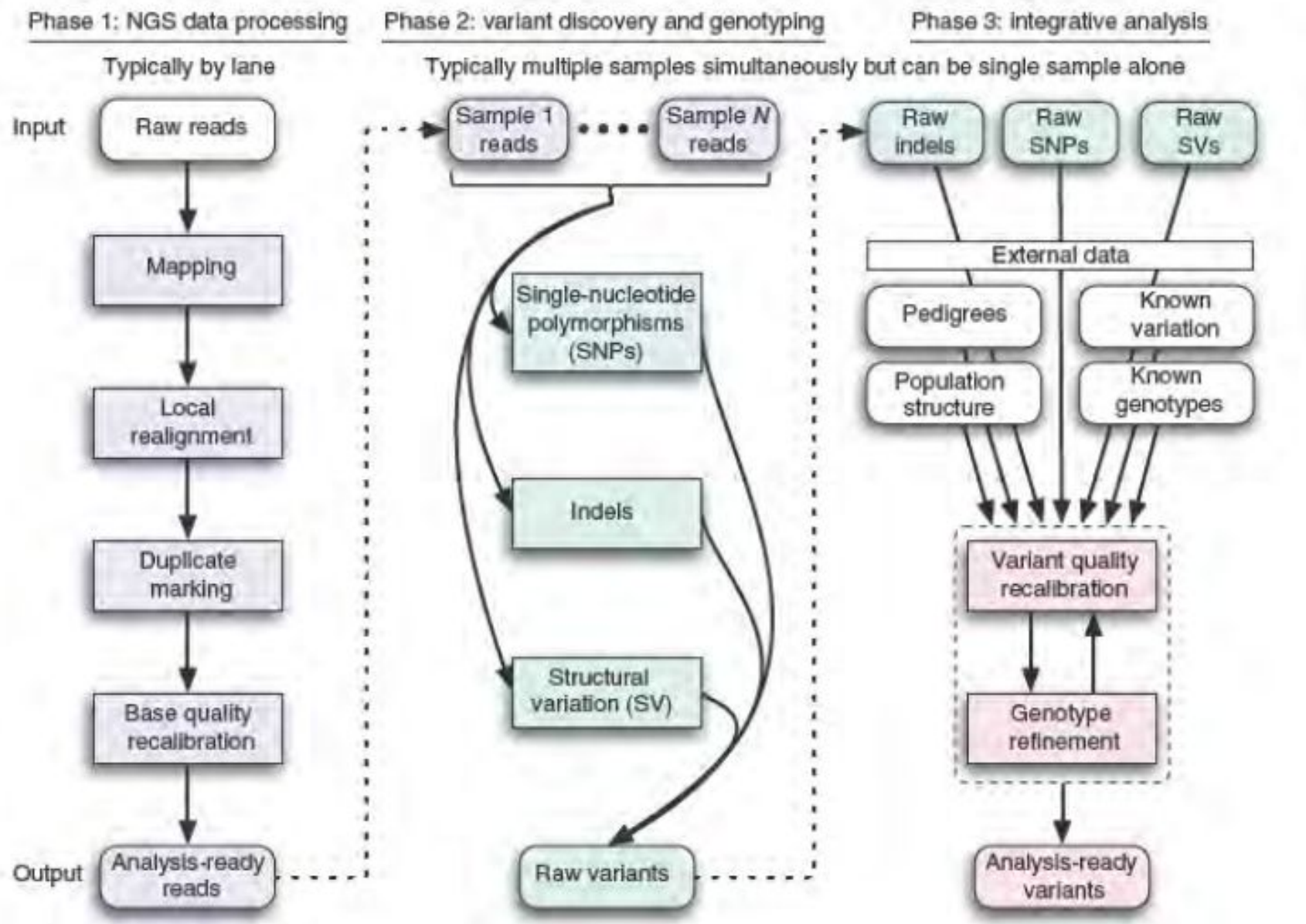
Yeni Nesil Dizileme Teknolojilerinde Varyant Çağırma

Emrah Kırdök


Phase 1: NGS data processing
Typically by lane
Input
Raw reads
Mapping
Local realignment
Duplicate marking
Base quality recalibration
Output
Analysis-ready reads
Phase 2: variant discovery and genotyping
Typically multiple samples simultaneously but can be single sample alone
Sample 1 reads
Sample N reads
Single-nucleotide polymorphisms (SNPs)
Indels
Structural variation (SV)
Raw variants
Phase 3: integrative analysis
Raw indels
Raw SNPs
Raw SVs
External data
Pedigrees
Known variation
Population structure
Known genotypes
Variant quality recalibration
Genotype refinement
Analysis-ready variants

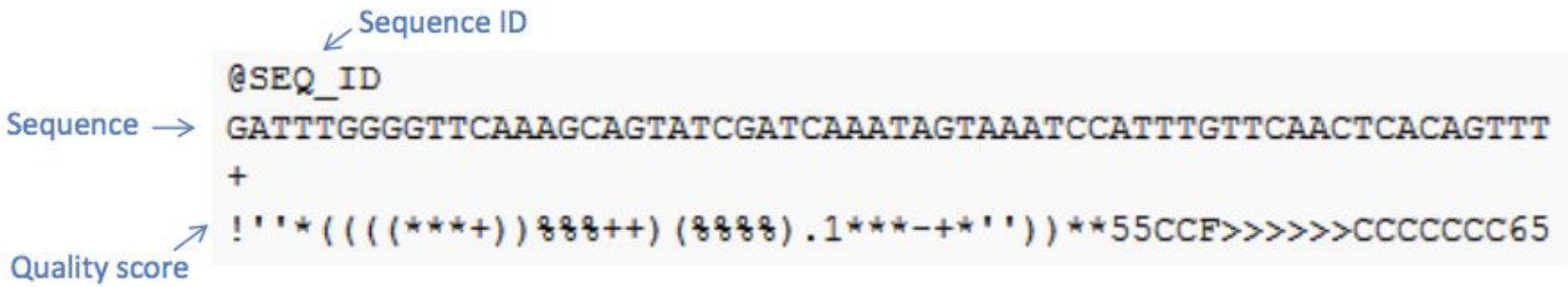
software	website
bwa	http://bio-bwa.sourceforge.net/
picard	http://picard.sourceforge.net/
samtools	http://samtools.sourceforge.net/
GATK	http://www.broadinstitute.org/gatk/
IGV	http://software.broadinstitute.org/software/igv/
tablet	http://bioinf.scri.ac.uk/tablet/
vcftools	http://vcftools.sourceforge.net/

Yeni Nesil Dizileme Teknolojileri?

Yeni Nesil Dizileme Teknolojileri?

FastQ dosyaları

- **FASTQ format**



- **Phred quality score**

!"#$%&'()*+,-./0123456789:;<=>?@

Phred score 0.....

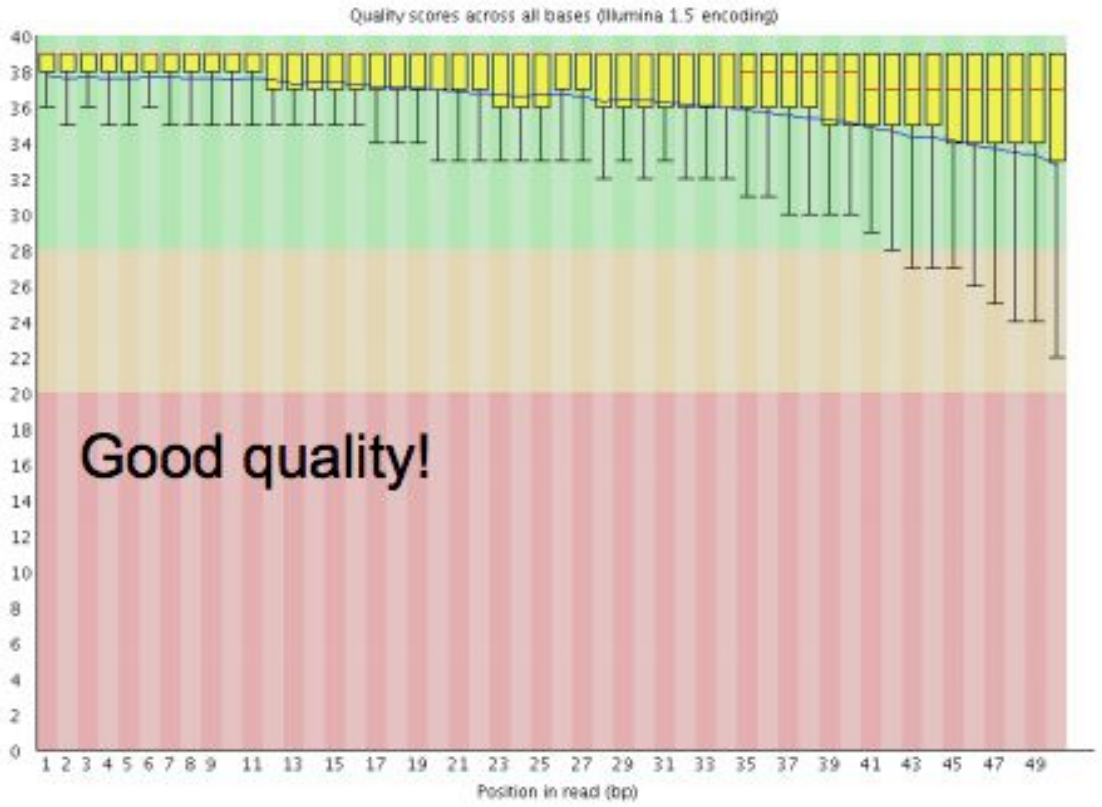
Error rate 1.....

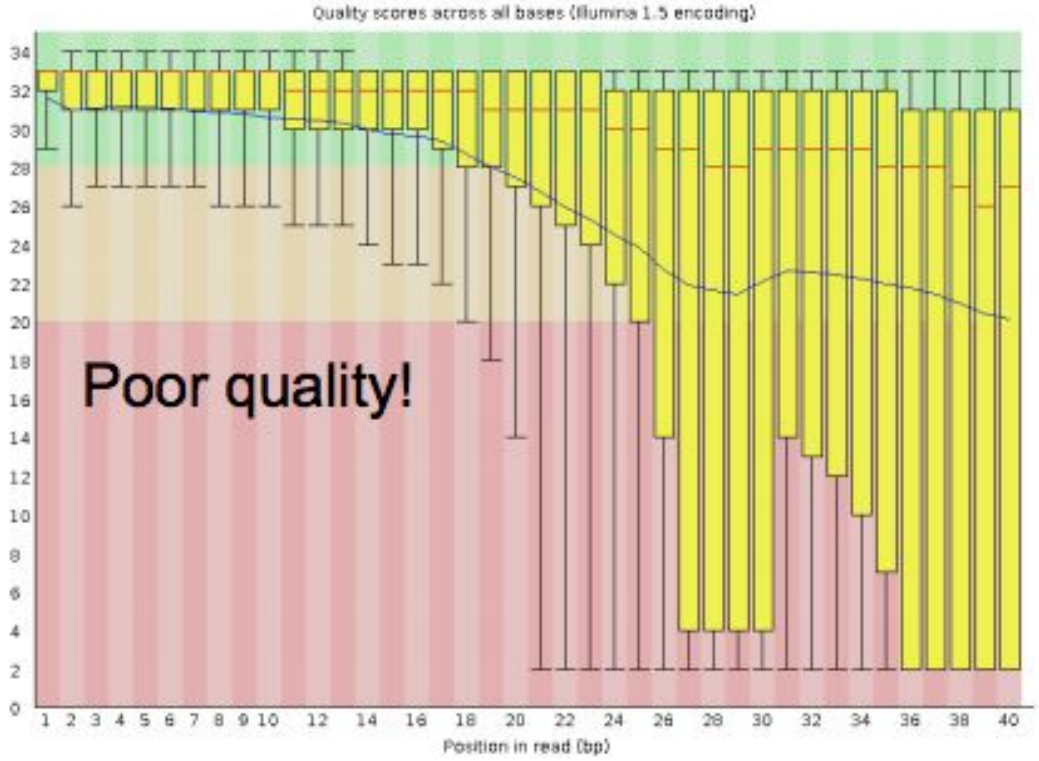
*Phred score = -10 * $log_{10}P$*

Table 1: Q-Scores and Error Probabilities

Quality Score	Error Probability
Q40	0.0001 (1 in 10,000)
Q30	0.001 (1 in 1,000)
Q20	0.01 (1 in 100)
Q10	0.1 (1 in 10)

DNA kütüphanesi kalite kontrolü

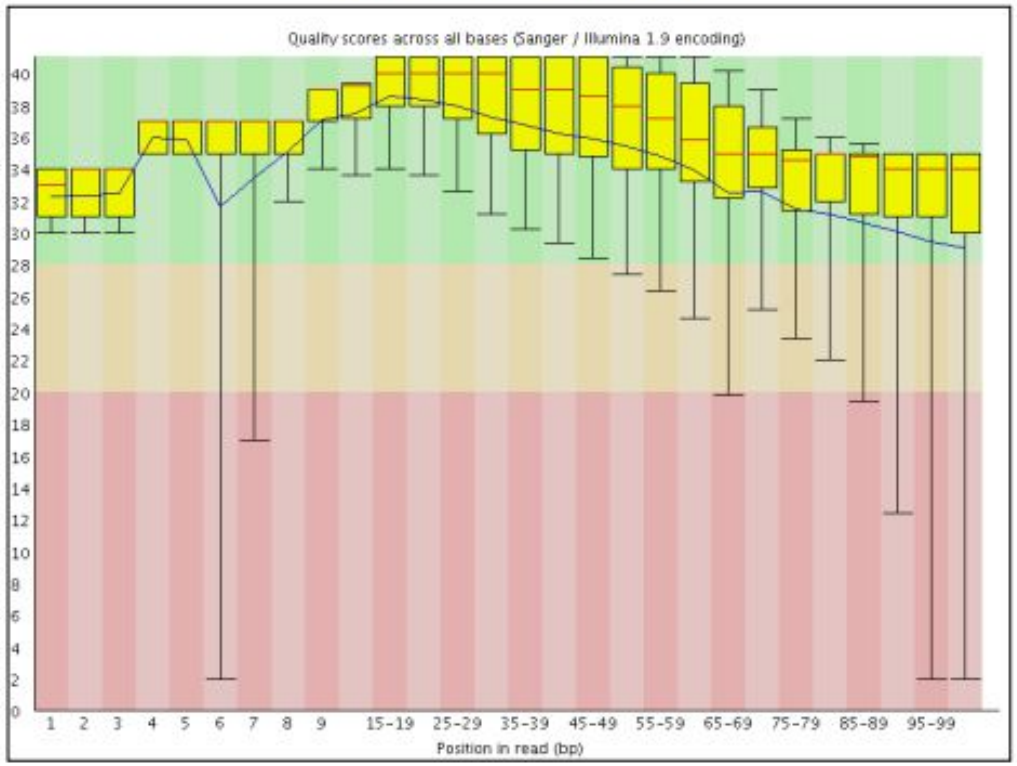




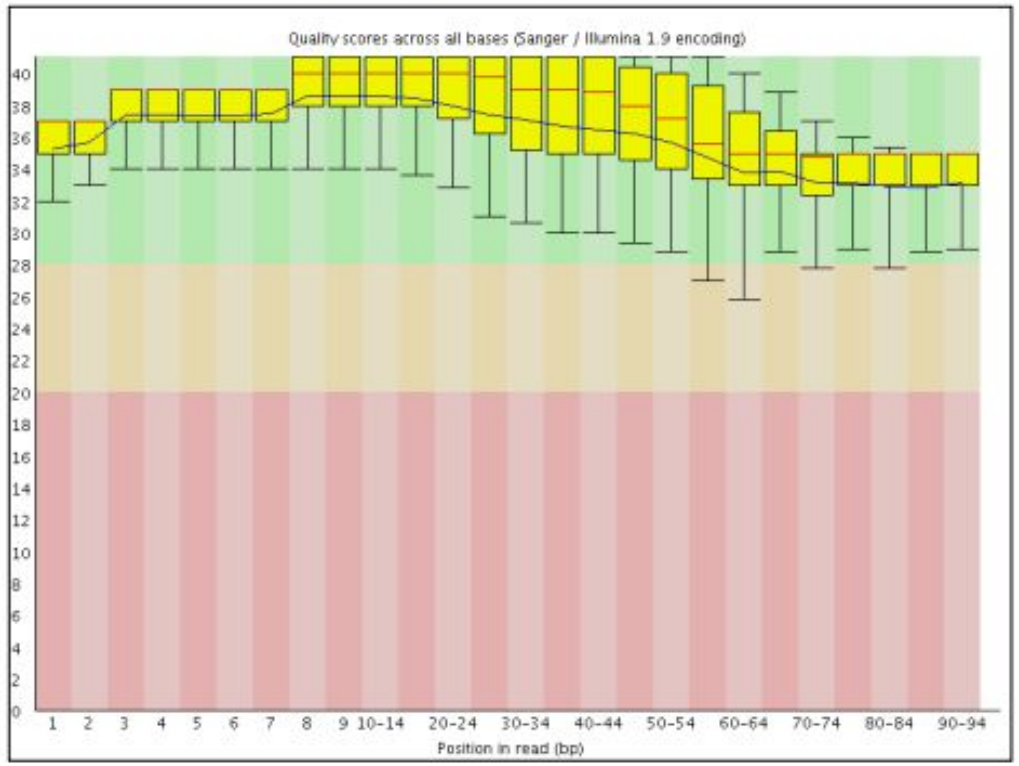
Adaptör içeriği

- Eğer DNA dizilerinde adaptörler varsa, bu analizlerimizi zorlaştırır
- Hata bile olabilir
  - AdaptorRemoval
  - trimmomatic

Before quality trimming



After quality trimming



Referans genom ile hizalama

- Referans genomu indeksleme (bwa index)
- DNA dizlerini referans genoma hizalama (bwa aln)
- Dizi hizlama dosyası (sequence alignment map)
- İkili dizi hizalama dosyası (bam)

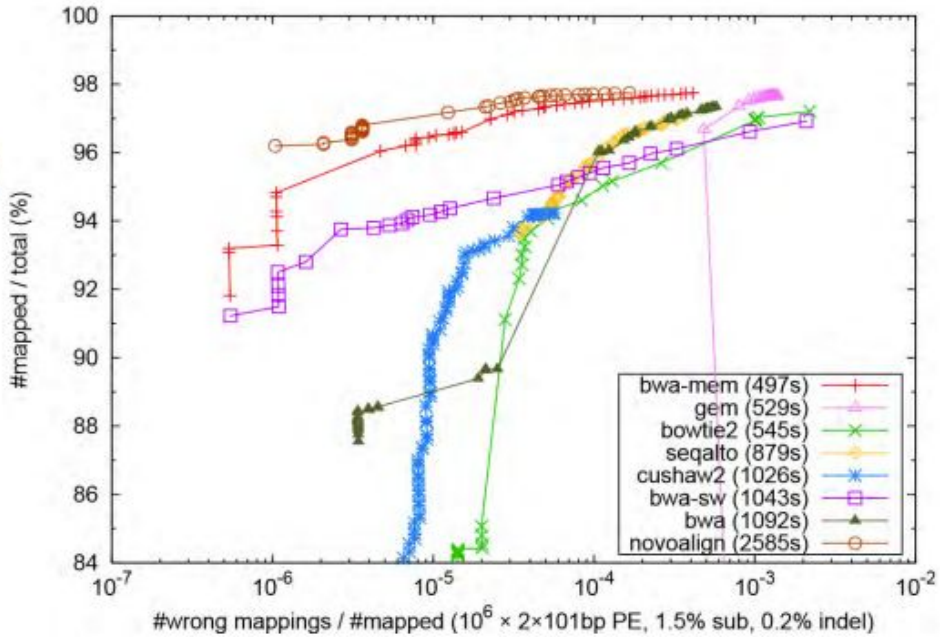
Referans genom hizalama

Hizalama

- Referans ile DNA okuması arasındaki en iyi eşleşmeyi bul
- Zorluklar
  - DNA okumalarındaki hatalar (okuma kalitesi)
  - DNA kütüphanelerindeki teknik sorunlar
  - Tekrar bölgeleri (homolog bölgeler, tekrarlar)
  - Homopolimerler

DNA hizalama algoritmaları

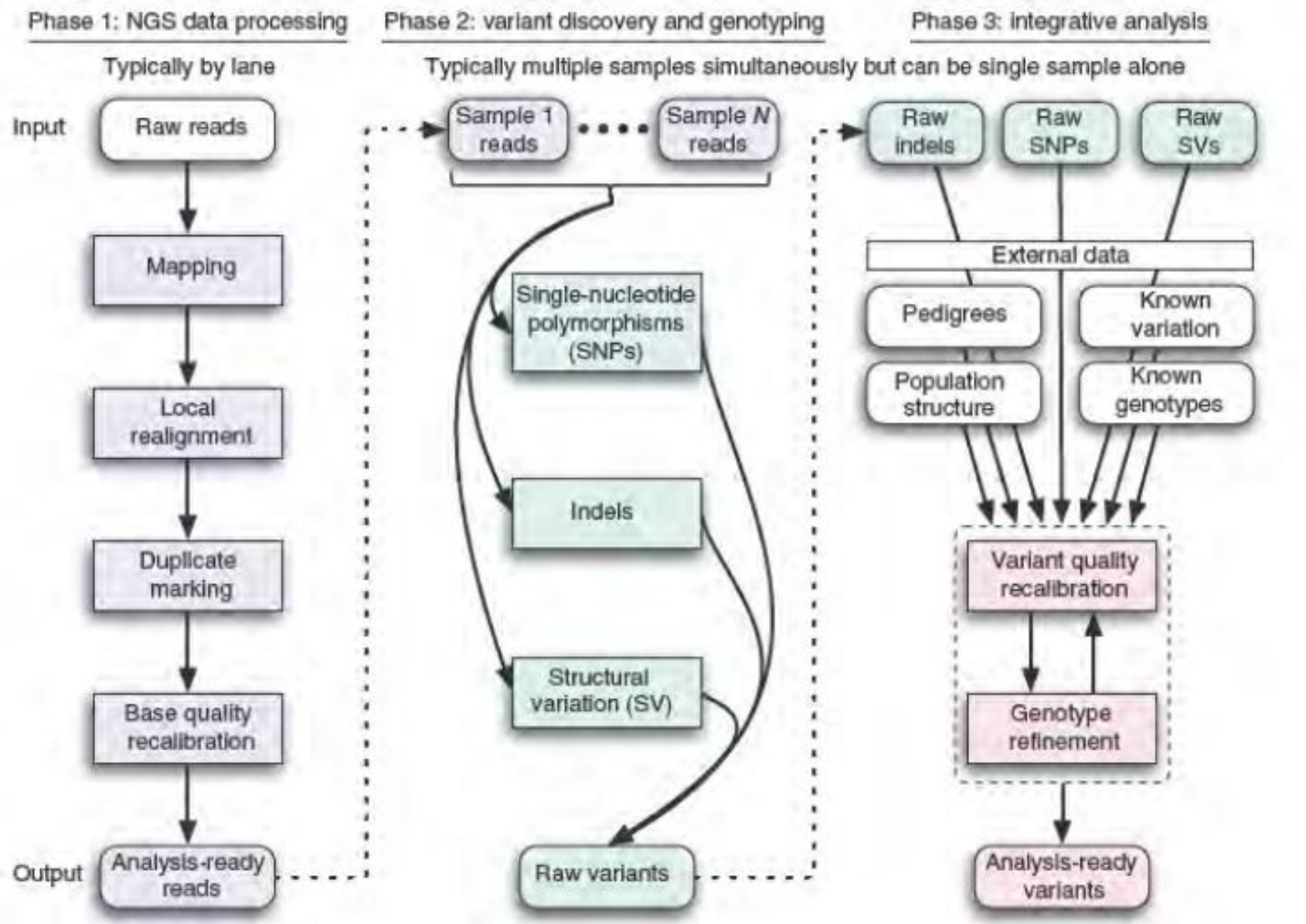
- BWA – 2009
- BWA-SW – 2010
- BWA-MEM – 2013
- Bowtie – 2009
- Bowtie2 – 2012
- Gem – 2012
- Cushaw2 – 2014
- Novoalign



Li, arXiv:1303.3997 (2013)

Dizi hizalama dosyaları

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714       16      chr20   190930  3       100M    *       0       0
       CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C    BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961       16      chr20   193953  50      100M    *       0       0
       TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G    DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030       16      chr20   270877  50      100M    *       0       0
       GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C    DDDDDDDDDCCDDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699       0       chr20   271218  50      50M4700N50M     *       0
       0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
accepted_hits.sam
```

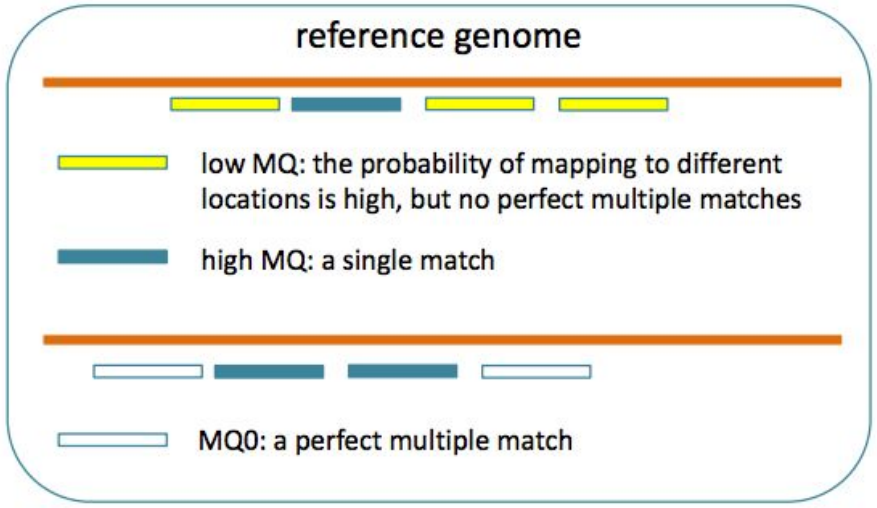

Phase 1: NGS data processing
Typically by lane
Input
Raw reads
Mapping
Local realignment
Duplicate marking
Base quality recalibration
Output
Analysis-ready reads
Phase 2: variant discovery and genotyping
Typically multiple samples simultaneously but can be single sample alone
Sample 1 reads
Sample N reads
Single-nucleotide polymorphisms (SNPs)
Indels
Structural variation (SV)
Raw variants
Phase 3: integrative analysis
Raw indels
Raw SNPs
Raw SVs
External data
Pedigrees
Known variation
Population structure
Known genotypes
Variant quality recalibration
Genotype refinement
Analysis-ready variants

Hizalamadan sonra

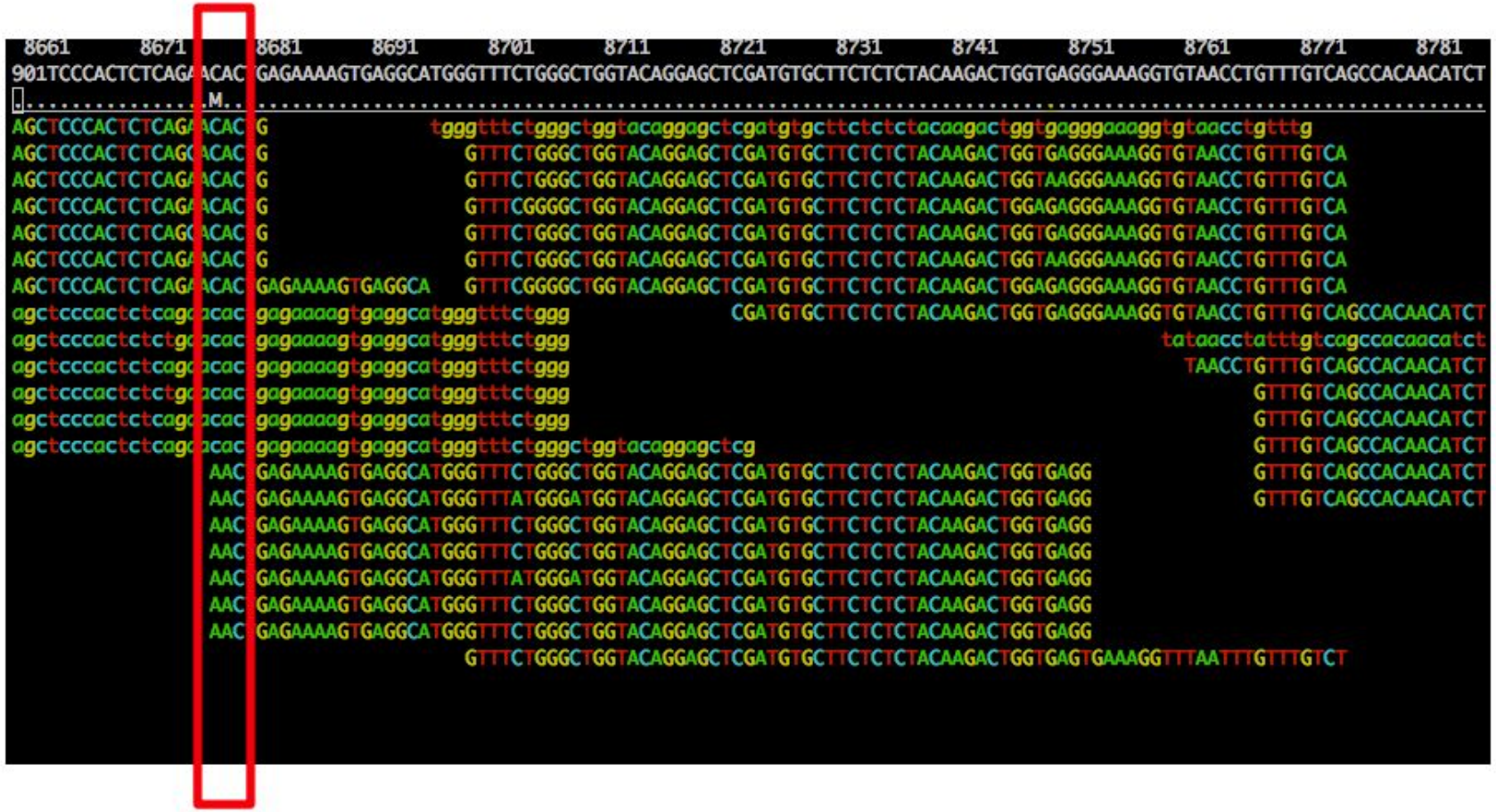
- PCR duplikasyonları sorun teşkil edebilir
- Bazı bölgeleri yeniden hizalayabiliriz (local realignment)
- Kalite değerleri kalibrasyonu

Hizalama kalitesi

- Eğer, bir DNA dizisi birden fazla bölgeye hizalanmışsa?
  - Tekrarlı dizileri yüzünden
  - Homopolimer bölgeler (düşük karmaşıklık)
  - Hatalar ve boşluklar
  - MQ (hizalama kalitesi skoru)



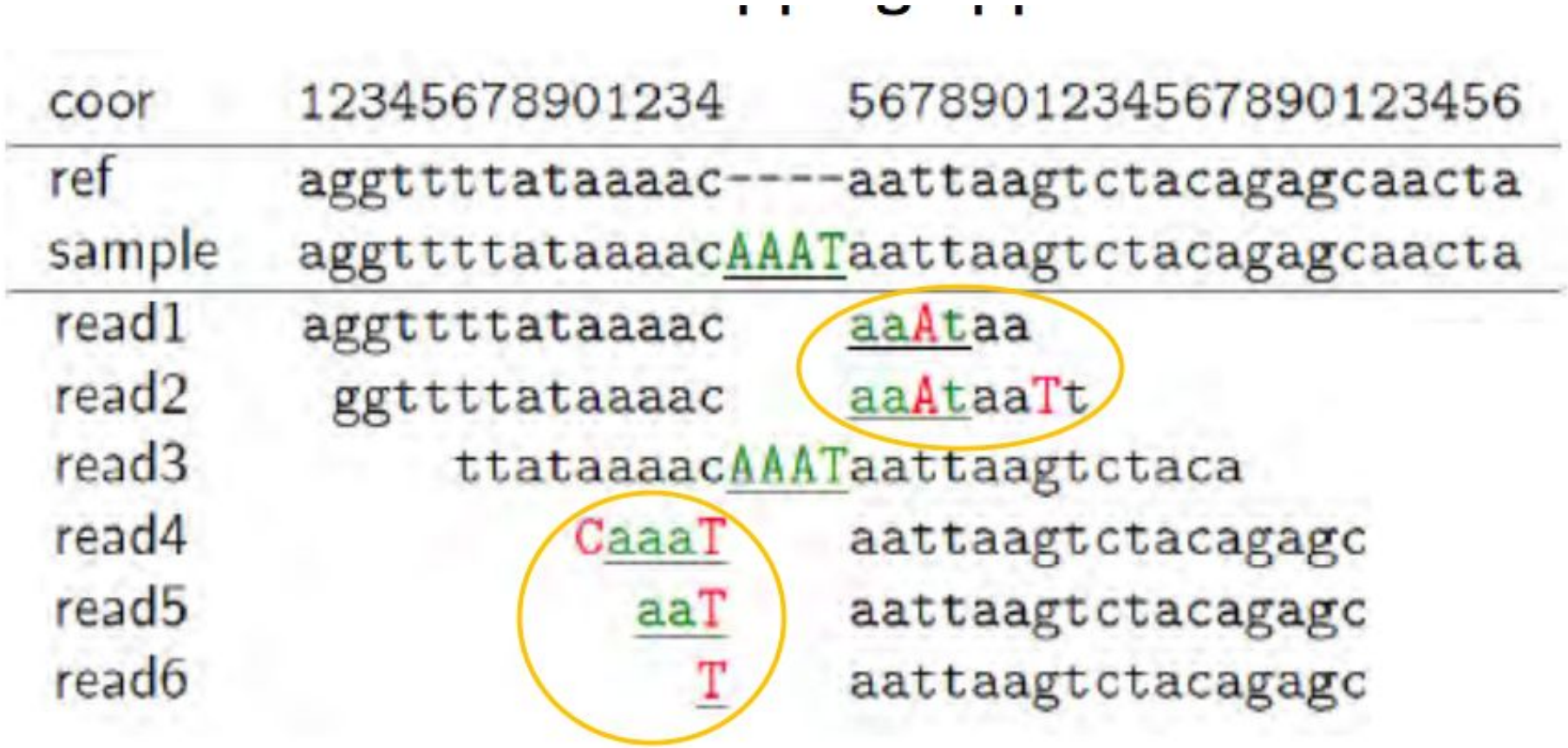
Optik PCR duplikasyonları

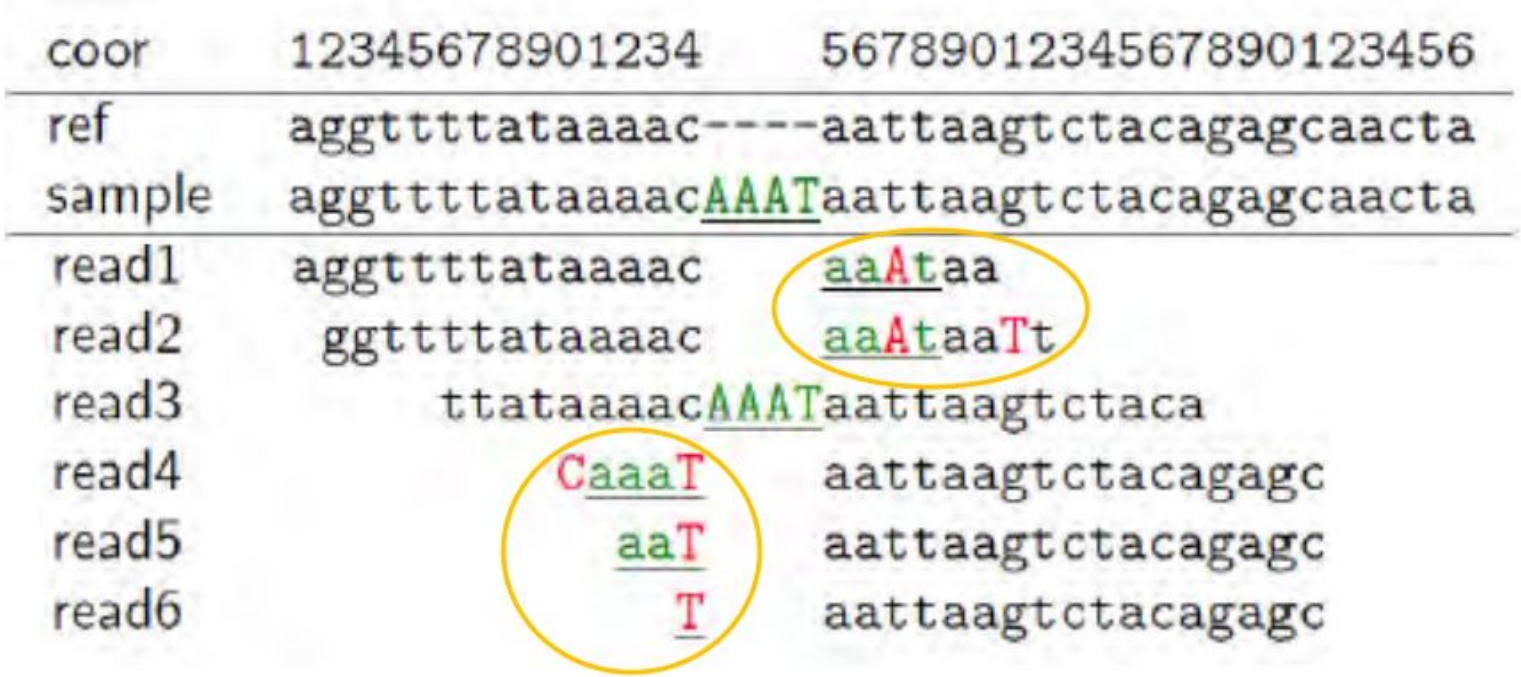


PCR duplikasyonları

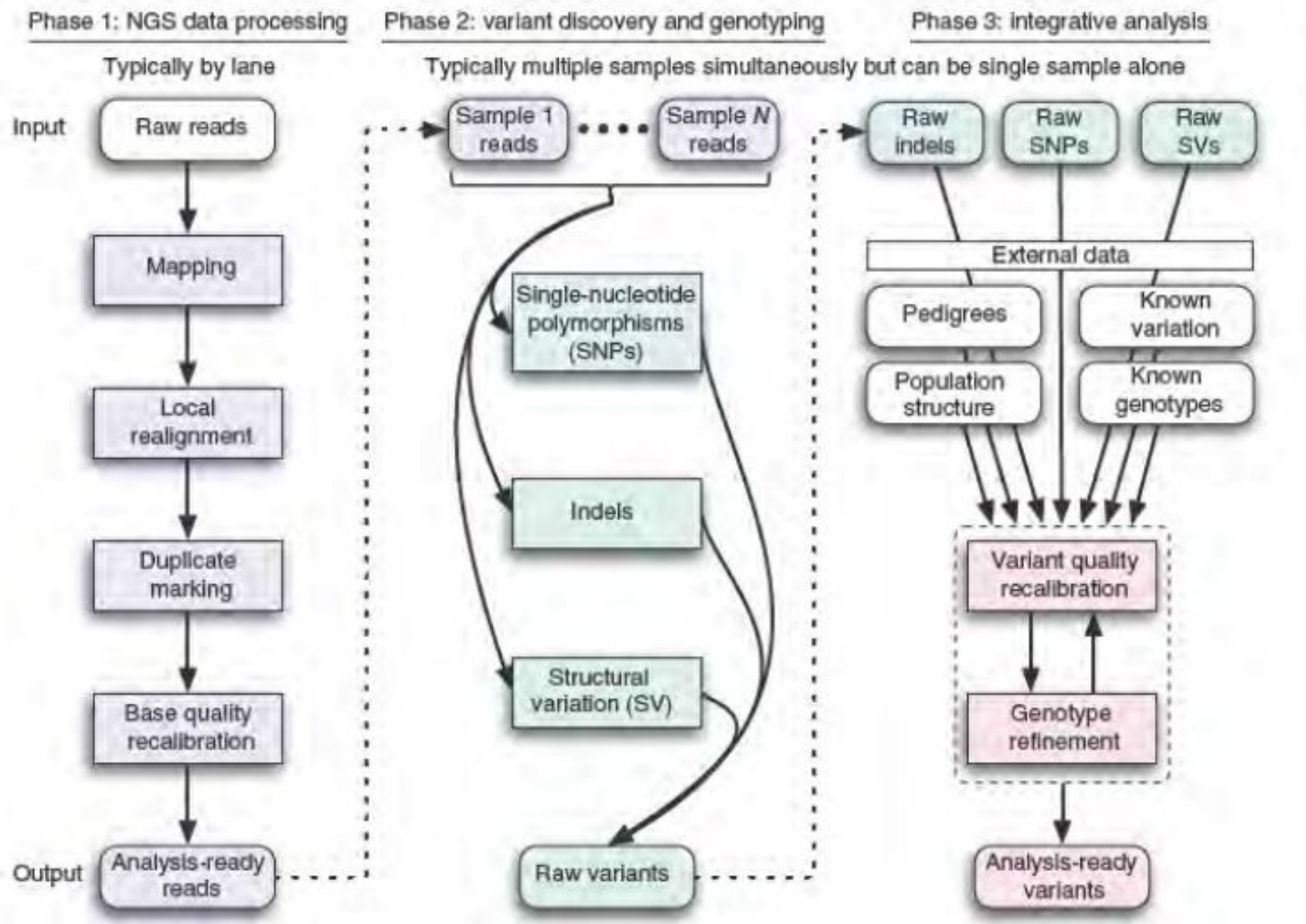
- Birbirinin eşdeğeri olan okumaları çıkartalım
- Aynı bölgeye hizalanan diziler

Bölgesel yeniden hizalama




coor 12345678901234 567890123456789012345678
ref aggttttataaaac----aattaagtctacagagcaacta
sample aggttttataaaacAAAATaattaagtctacagagcaacta
read1 aggttttataaaac aaAtaa
read2 ggttttataaaac aaAtaaTt
read3 ttataaaacAAAATaattaagtctaca
read4 CaaaT aattaagtctacagagc
read5 aaT aattaagtctacagagc
read6 T aattaagtctacagagc

```
coor      12345678901234     5678901234567890123456
ref       aggttttataaaac----aattaagtctacagagcaacta
sample    aggttttataaaacAAATaattaagtctacagagcaacta
read1     aggttttataaaacaaataa
read2      ggttttataaaacaaataatt
read3          ttataaaacaaataattaagtctaca
read4                caaataattaagtctaca
read5                  aataattaagtctaca
read6                    taattaagtctaca
```


Phase 1: NGS data processing
Typically by lane
Input
Raw reads
Mapping
Local realignment
Duplicate marking
Base quality recalibration
Output
Analysis-ready reads
Phase 2: variant discovery and genotyping
Typically multiple samples simultaneously but can be single sample alone
Sample 1 reads
Sample N reads
Single-nucleotide polymorphisms (SNPs)
Indels
Structural variation (SV)
Raw variants
Phase 3: integrative analysis
Raw indels
Raw SNPs
Raw SVs
External data
Pedigrees
Known variation
Population structure
Known genotypes
Variant quality recalibration
Genotype refinement
Analysis-ready variants

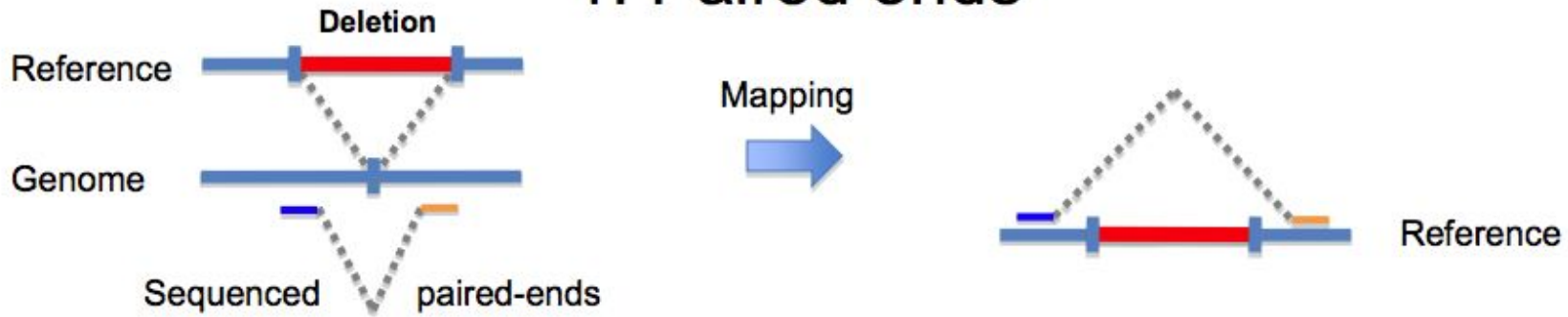
Varyant Çağırma

- Varyant : Bir sonuç
  - Bir pozisyonda, referans genomdan farklı olan bölgeler
  - TNP
  - Kısa INDEL
  - Yapısal varyant

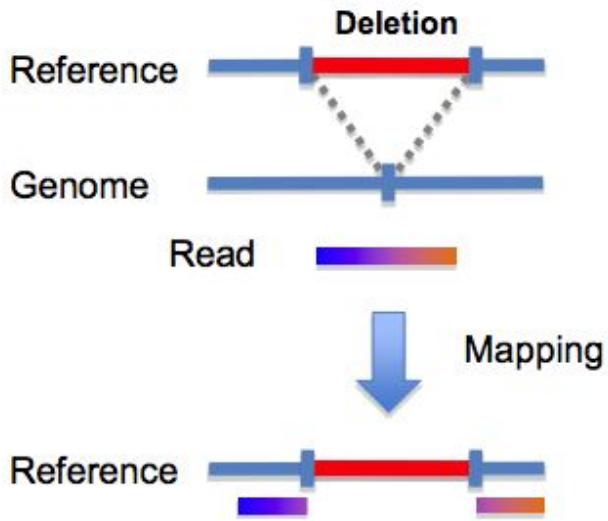
Yapısal varyant

- Farklı tipte kromozomal değişimleri
  - INDEL
  - Translokasyonlar
  - Inversiyon

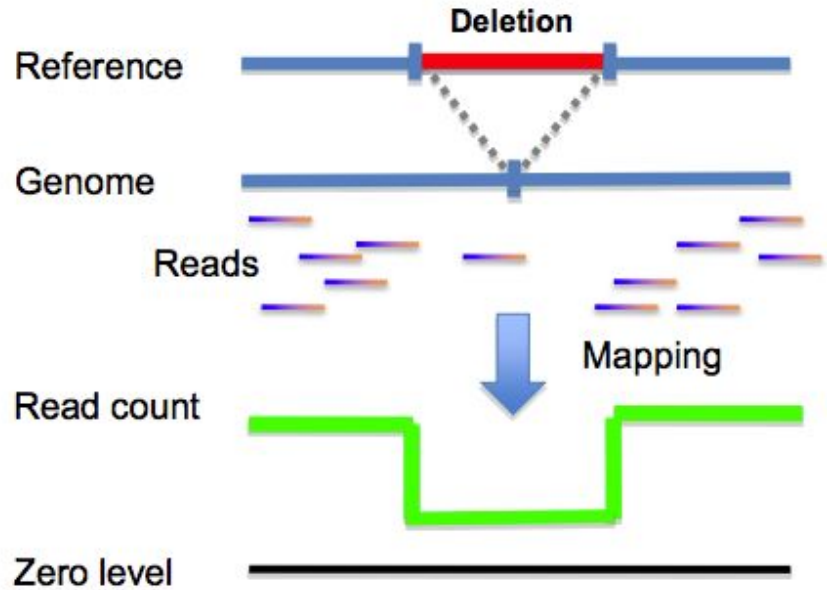
1. Paired ends

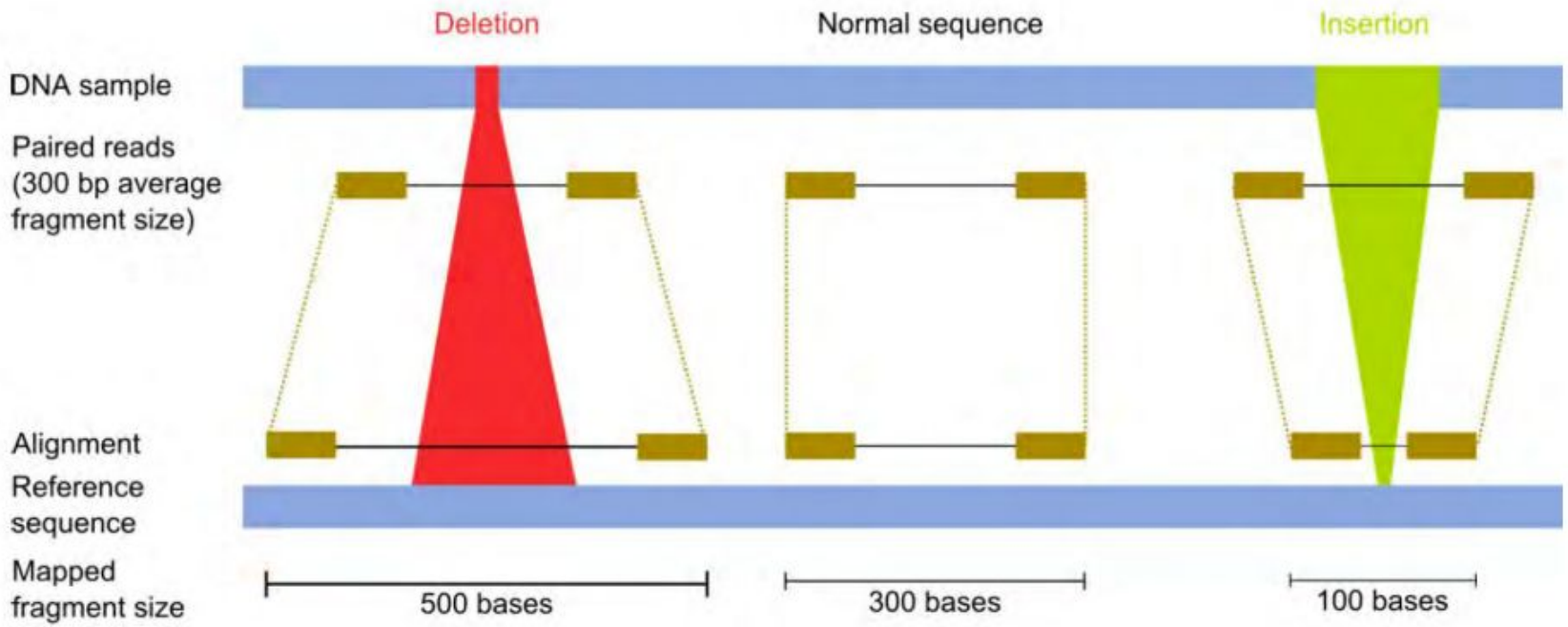


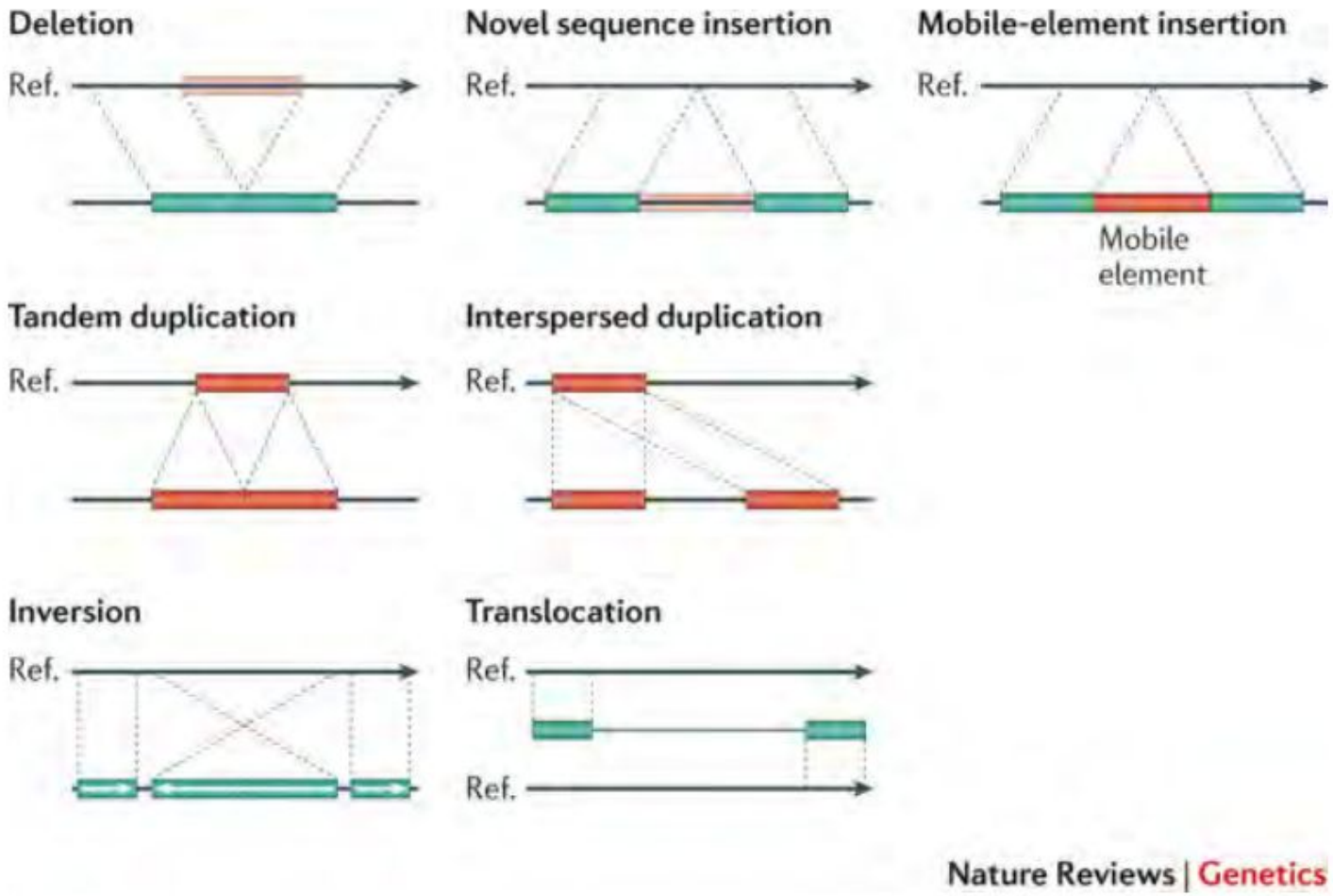
3. Split read



2. Read depth



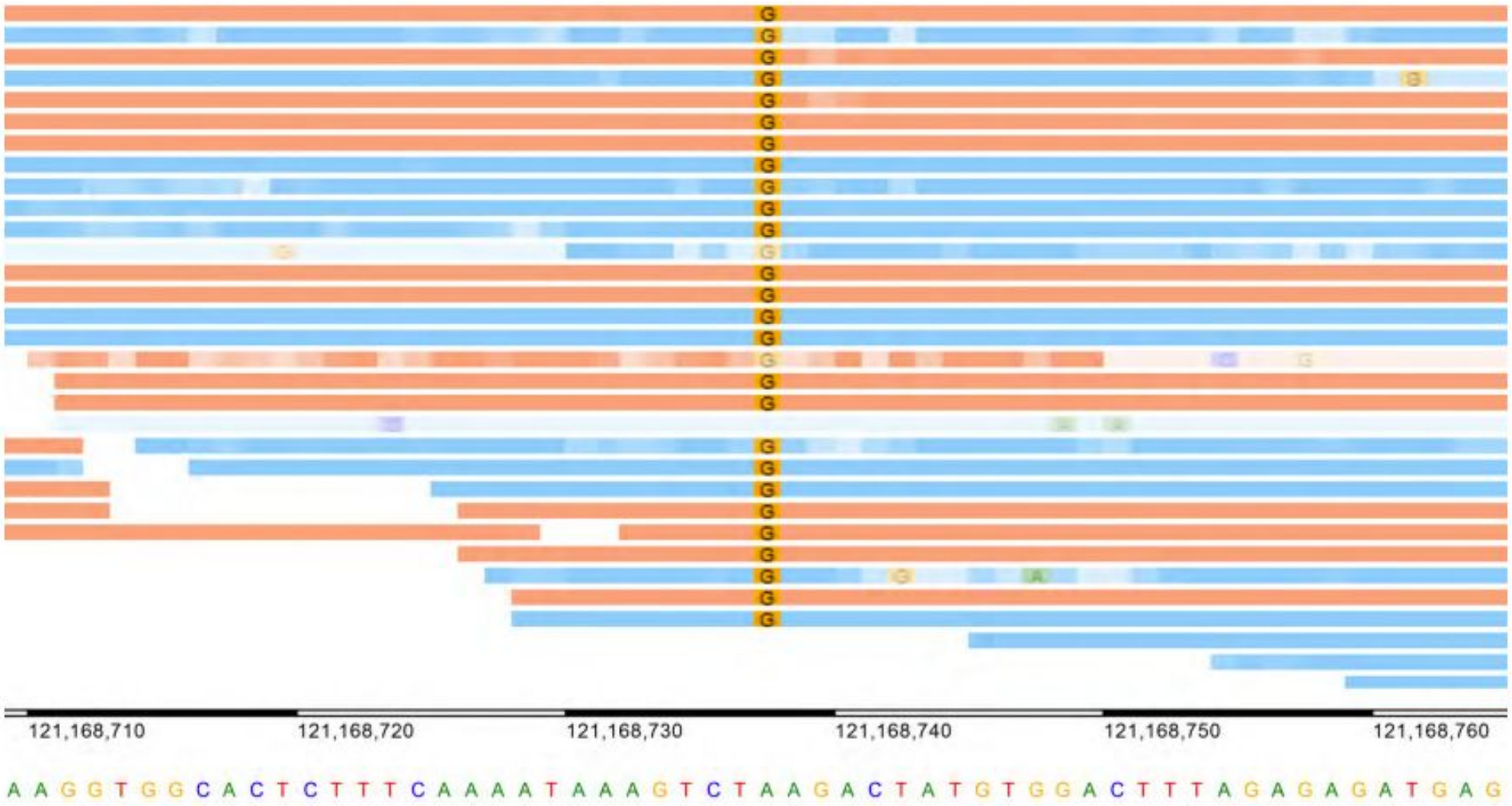

Deletion
Normal sequence
Insertion
DNA sample
Paired reads
(300 bp average
fragment size)
Alignment
Reference
sequence
Mapped
fragment size
500 bases
300 bases
100 bases



Nature Reviews Genetics 12, 363-376 (May 2011) | doi:10.1038/nrg2958

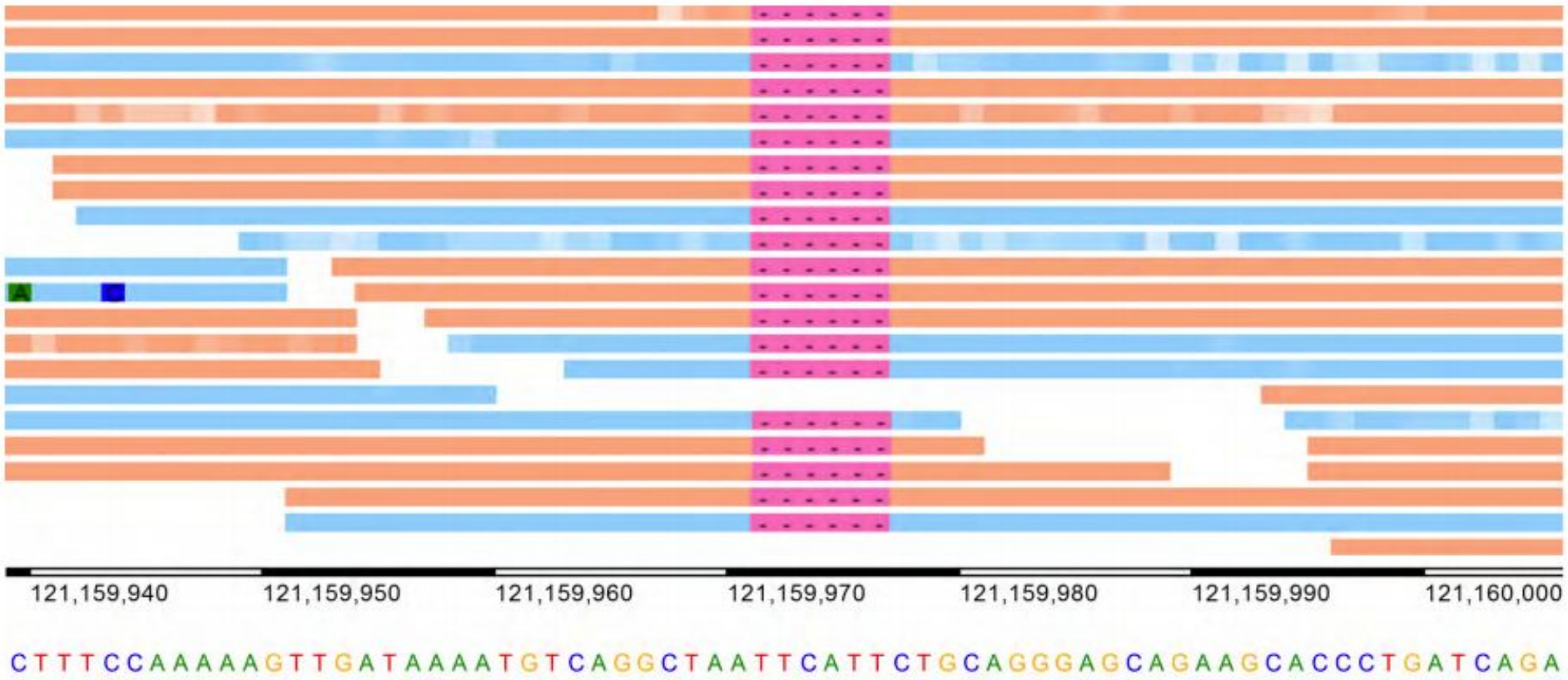
TNP çağırma

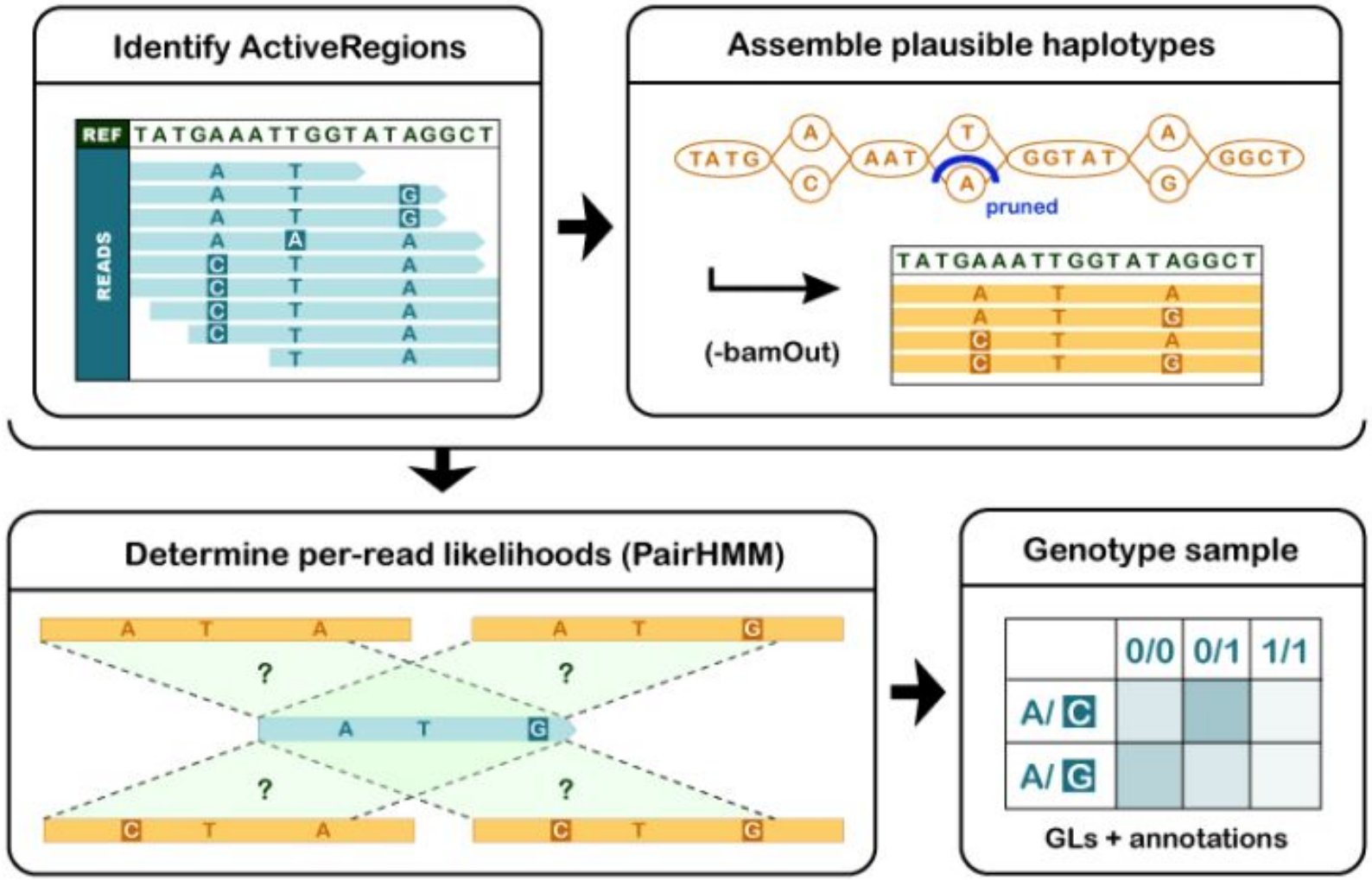
- Tek nükleotid polimorfizmi
  - Bir pozisyona hizalanmış bazlara bak, ve farklılığı tespit et
- Düşünülmesi gereken faktörler:
  - Baz okuma kaliteleri
  - INDEL bölgelere olan yakınlık
  - O bölgeye hizalanmış bazların hizalama kaliteleri
  - Dizi uzunluğu

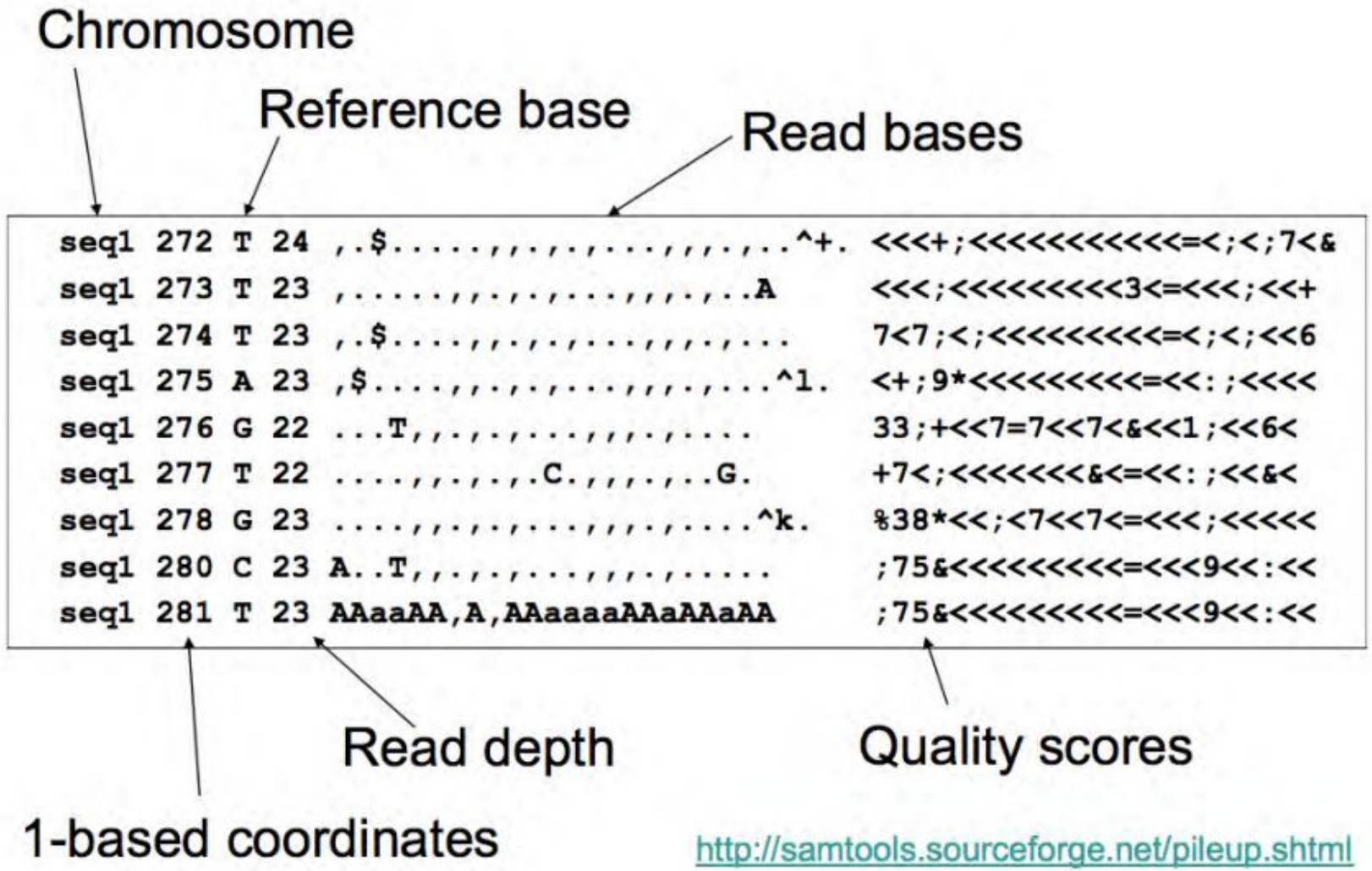


http://www.sanger.ac.uk/mousegenomes

INDEL çağırma



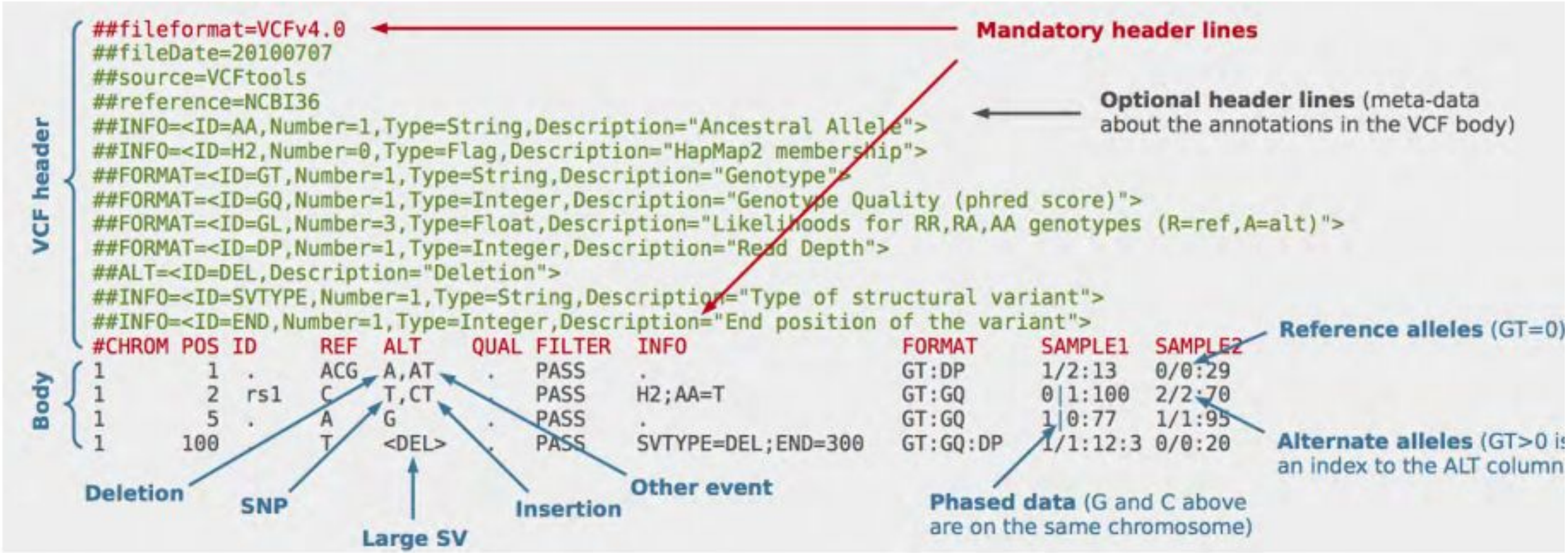

Identify ActiveRegions
REF
TATGAAATTGGTATAGGCT
READS
Assemble plausible haplotypes
TATG
AAT
GGTAT
GGCT
pruned
(-bamOut)
TATGAAATTGGTATAGGCT
Determine per-read likelihoods (PairHMM)
Genotype sample
0/0
0/1
1/1
A/C
A/G
GLs + annotations


Chromosome
Reference base
Read bases
seq1 272 T 24 ,.$.....,.,.,.....,,,,,..^+. <<<+;<<<<<<<<<<=<;<;7<&
seq1 273 T 23 ,.....,.,.,.....,,,,,..A <<<;<<<<<<<<3<=<<<;<<+
seq1 274 T 23 ,.$.....,.,.,.....,,,,,... 7<7;<;<<<<<<<<<=<;<;<<6
seq1 275 A 23 ,$.....,.,.,.....,,,,,...^1. <+;9*<<<<<<<<<=<<:;<<<<
seq1 276 G 22 ...T,,.,.,.,.....,,,,,.... 33;+<<7=7<<7<&<<1;<<6<
seq1 277 T 22 .....,.,.,.C.,,,,,,..G. +7<;<<<<<<<&<=<<:;<<&<
seq1 278 G 23 .....,.,.,.....,,,,,....^k. %38*<<;<7<<7<=<<<;<<<<<
seq1 280 C 23 A..T,,.,.,.....,,,,,..... ;75&<<<<<<<<<=<<<9<<:<<
seq1 281 T 23 AAaaAA,A,AAaaaaAAaAAaAA ;75&<<<<<<<<<=<<<9<<:<<
Read depth
Quality scores
1-based coordinates
http://samtools.sourceforge.net/pileup.shtml

Variant Call Format (VCF)

- DNA polimorfizm verisini saklamak için standart bir format
  - TNP
  - INDEL
  - Yapısal varyatlar
- Indekslenerek, daha hızlı erişim sağlanabilir

VCF formatı


VCF header
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
Body
1 1 . ACG A,AT . PASS . GT:DP 1/2:13 0/0:29
1 2 rs1 C T,CT . PASS H2;AA=T GT:GQ 0|1:100 2/2:70
1 5 . A G . PASS . GT:GQ 1|0:77 1/1:95
1 100 T <DEL> . PASS SVTYPE=DEL;END=300 GT:GQ:DP 1/1:12:3 0/0:20
Mandatory header lines
Optional header lines (meta-data about the annotations in the VCF body)
Reference alleles (GT=0)
Alternate alleles (GT>0 is an index to the ALT column
Deletion
SNP
Large SV
Insertion
Other event
Phased data (G and C above are on the same chromosome)

Header
lines starting with ##: arbitrary number of meta-information lines
line starting with #: column definition – mandatory columns include:

CHROM chromosome
POS position of the start of the variant
ID unique identifier of the variant (e.g. rs number for SNPs)
REF reference allele
ALT comma separated list of alternate non-reference alleles
QUAL phred-scaled quality score
FILTER site filtering information
INFO user extensible annotation (e.g. samtools and GATK may differ in this)

Peki hangi SNP’ler gerçek pozitif?

- Filtreleme kriterleri
  - Baz kalitesi (20)
  - Okuma derinliği
  - Hizalama derinliği (50 – 60)
  - İleri veya geri iplik
  - TNP yoğunluğu
  - INDEL yakınlığı

Varyant anotasyonu

- Bu varyantların nerede olduklarını biliyoruz
- Peki bu varyantları anote edebilir miyiz?
  - SnpEFF: Varyant anotasyonu
  - Bir genom anotasyon dosyası varlığında, bu varyantları daha iyi anlayabiliriz
  - Mesela hangi genlerde?
  - Amino asit değişimleri oluşturuyor mu?