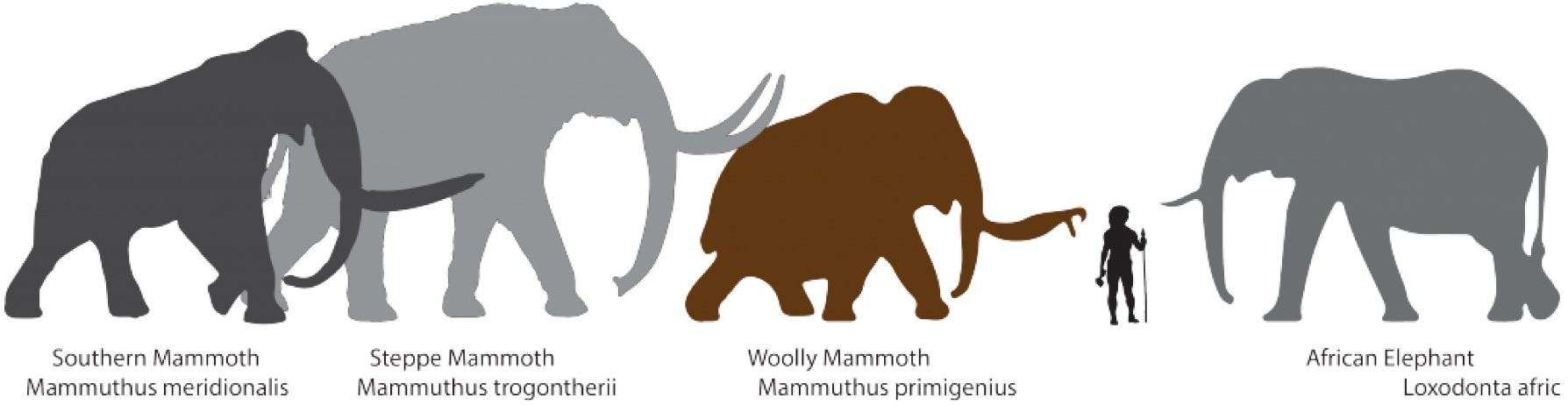
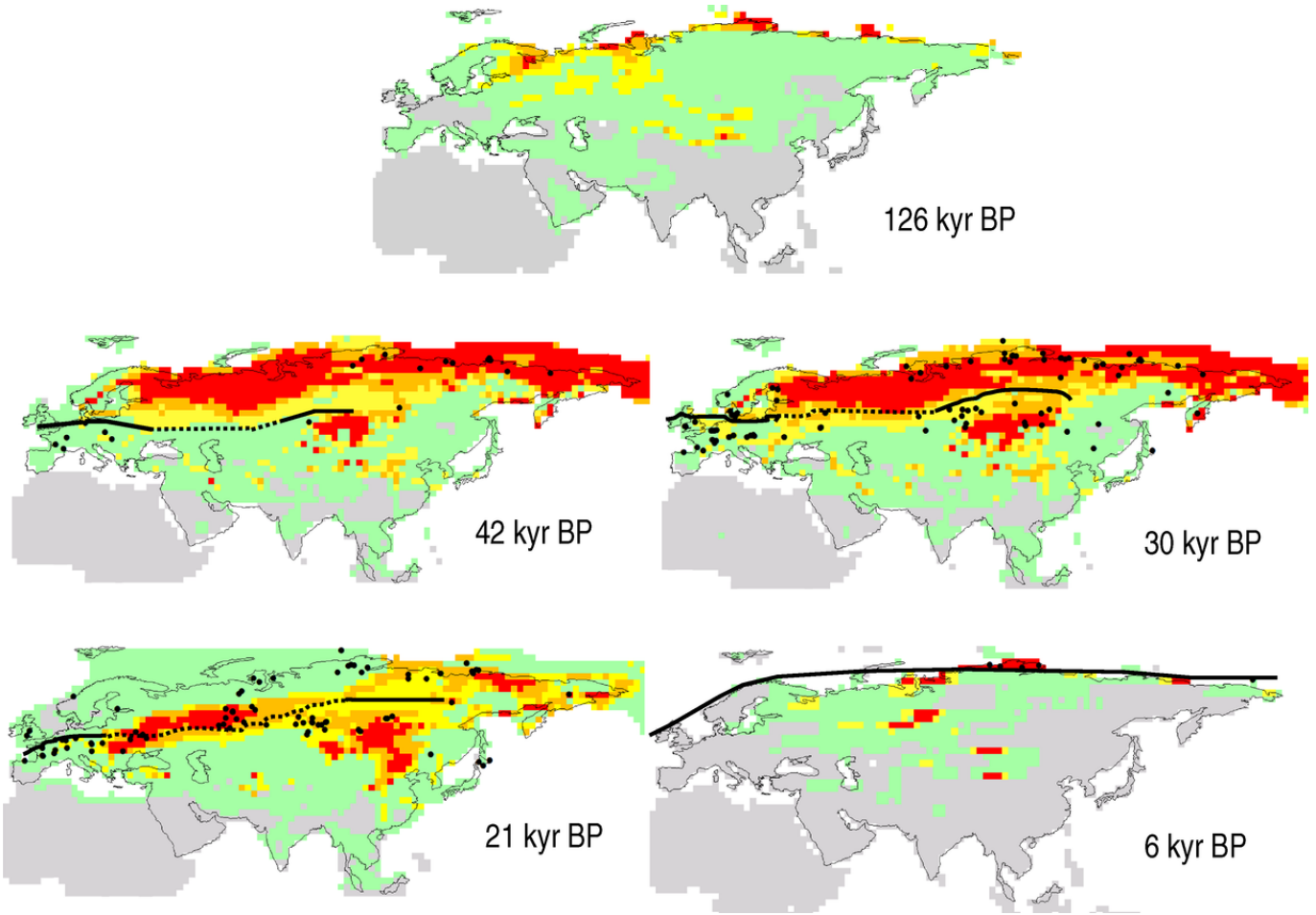
Esercitazione di laboratorio: allineamento di reads ad un genoma di riferimento

Specie

- Mammuth lanoso (*Mammuthus primigenius*), da 200 a 5 mila anni fa



Aerale di distribuzione



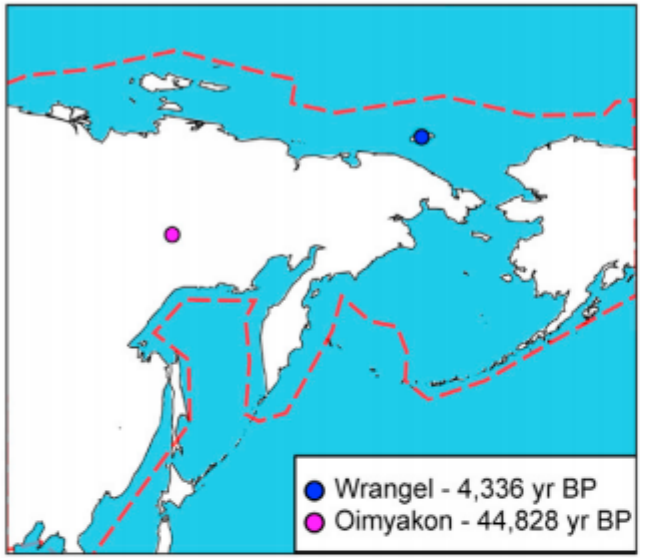
Current Biology
Report

Complete Genomes Reveal Signatures of Demographic and Genetic Declines in the Woolly Mammoth

Eleftheria Palkopoulou,[1,2,*] Swapan Mallick,[3,4,5] Pontus Skoglund,[3,4,6] Jacob Enk,[7,8] Nadin Rohland,[3,4] Heng Li,[3,4] Ayça Omrak,[6] Sergey Vartanyan,[9] Hendrik Poinar,[7] Anders Götherström,[6] David Reich,[3,4,5] and Love Dalén[1,*]

[1]Department of Bioinformatics and Genetics, Swedish Museum of Natural History, 10405 Stockholm, Sweden
[2]Department of Zoology, Stockholm University, 10691 Stockholm, Sweden
[3]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
[4]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[5]Howard Hughes Medical Institute, Harvard Medical School, Boston, MA 02115, USA
[6]Department of Archaeology and Classical Studies, Stockholm University, 10691 Stockholm, Sweden
[7]McMaster Ancient DNA Centre, Departments of Anthropology and Biology, and the Michael G. DeGroote Institute for Infectious Disease Research, McMaster University, 1280 Main Street West, Hamilton, ON L8S 4L9, Canada
[8]MYcroarray, 5692 Plymouth Road, Ann Arbor, MI 48105, USA
[9]N.A. Shilo North-East Interdisciplinary Scientific Research Institute, Far East Branch, Russian Academy of Sciences (NEISRI FEB RAS), Magadan 685000, Russia
*Correspondence: elle.palkopoulou@gmail.com (E.P.), love.dalen@nrm.se (L.D.)
http://dx.doi.org/10.1016/j.cub.2015.04.007

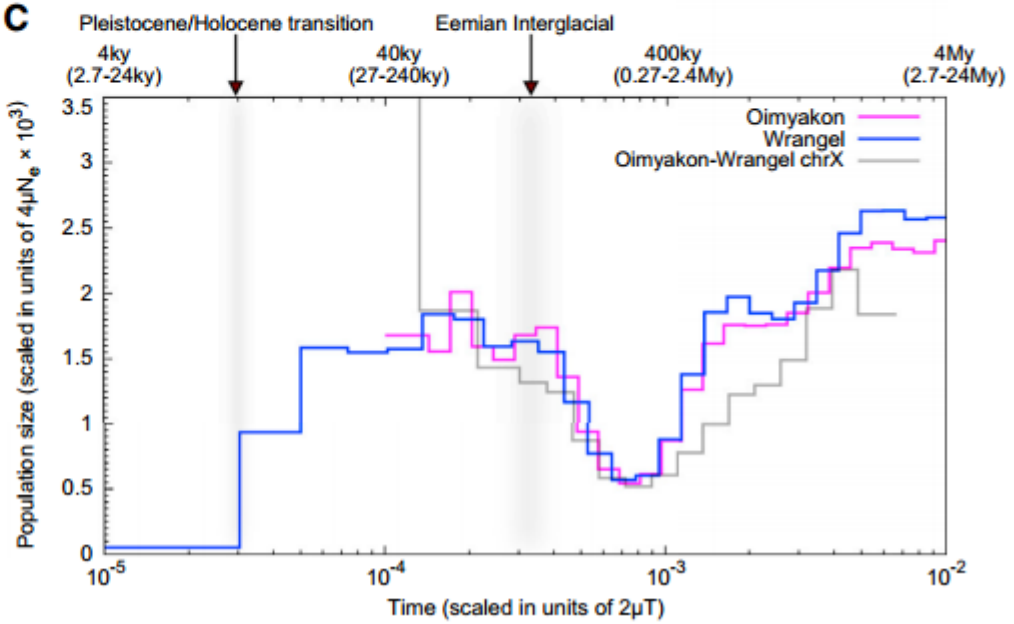
A



B

Sample	^{14}C date ± error (years)	Median calibrated date (years)	# raw reads ($\times 10^6$)	Average coverage	Average read length (bp)
Wrangel	3,905 ± 47	4,336	1,262	17.1	69
Oimyakon	41,300 ± 900	44,828	1,401	11.2	55

C



European Nucleotide Archive

- ID: PRJEB7929

Oimyakon

FASTQs

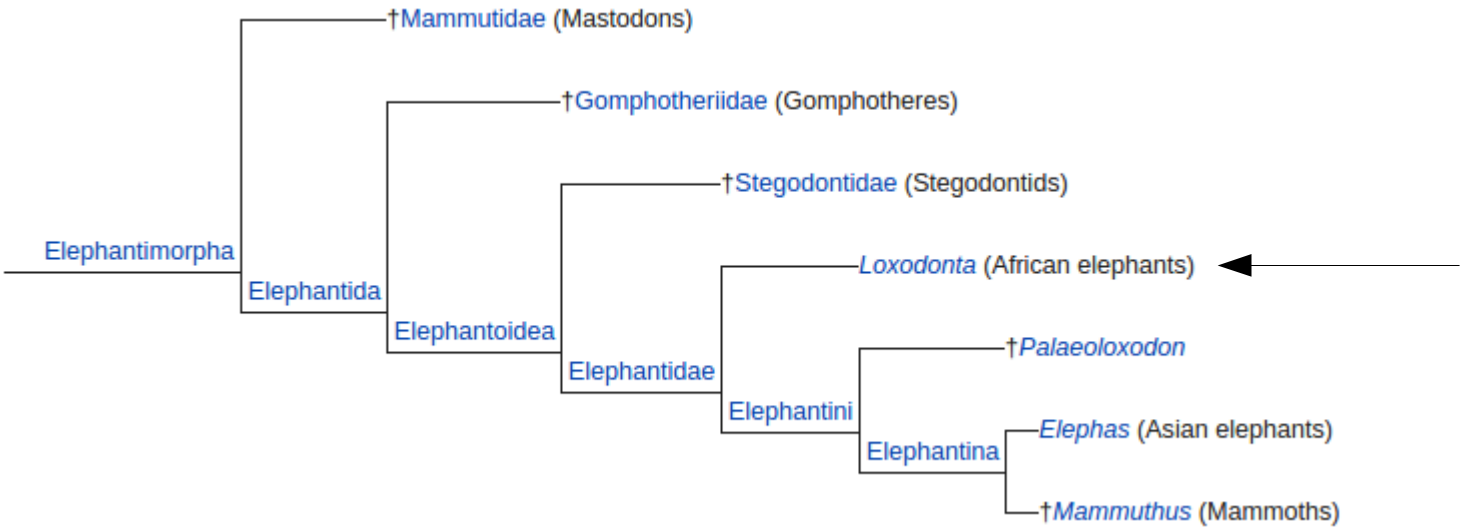
Showing results 1 - 2 of 2 results

Study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument model	Library layout	FASTQ files (FTP)	FASTQ files (Galaxy)	Submitted files (FTP)	Submitted files (Galaxy)
PRJEB7929	SAMEA3340290	ERS701783	ERX931666	ERR852028	37349	Mammuthus primigenius	Illumina HiSeq 2500	SINGLE	File 1	File 1	BAM File 1	BAM File 1
PRJEB7929	SAMEA3340289	ERS701782	ERX935618	ERR855944	37349	Mammuthus primigenius	Illumina HiSeq 2500	SINGLE	File 1	File 1	BAM File 1	BAM File 1

Wrangel Island

Reference genome (LoxAfr_v4)

- *Loxodonta africana (27 chr, 3.2 Gb)*



Submitter: Broad Institute
Assembly level: Scaffold
Assembly: GCA_000001905.1 Loxafr3.0 **scaffolds:** 2,352 **contigs:** 95,866 **N50:** 69,023 **L50:** 13,607
BioProjects: PRJNA70973, PRJNA12569
Whole Genome Shotgun (WGS): INSDC: AAGU00000000.3
Statistics: total length (Mb): 3196.74
protein count: 29784
GC%: 40.9
NCBI Annotation Release: 101

Risorse

- Genoma di riferimento:
  - /media/studenti/risorse/LoxAfr.chr3.reference.fa
- Reads (single-end trimmed):
  - /media/studenti/risorse/oimyakon.fastq
  - /media/studenti/risorse/wrangel.fastq

Reads

Oimyakon.fastq:

@HS2000-214:164:D204RACXX:1:2302:11944:55815
TAGCATTGCTGCTTTCCAGGTAGTTATCTACACTGTCAAAAGGCAACTGAAAATAGAATAAA
+
JJJJIIIHHHIJIJIJJIJJIJJJJIJJJIJJJIHGHIJJJJJJJJJJJJIHJJJJIIJJG

Wrangel.fastq:

@HWI-D00415:38:C3F6KACXX:4:1215:6202:19690
CAGGTGTTGGAGATGTCTAGATGAAAGCGTTTTACATAATTATATATCTTTTATTCTA
+
<<BF<BFFFFFFFFFBFFFFFFFFIIFFIFIIFFFIIIIIIIIIIIIIIIIIIFFII

Quante reads contengono i due files fastq?

Reads

Quante reads contiene oimyakon.fastq?

1) Possiamo contare le righe e dividere per 4:

```
wc -l /media/studenti/risorse/oimyakon.fastq
```

Otteniamo 369476 reads, diviso 4: 92369 reads

2) Possiamo cercare quanti “header” ci sono nel file:

```
grep “@HS2000” oimyakon.fastq | wc -l
```

Quante reads contiene wrangel.fastq?

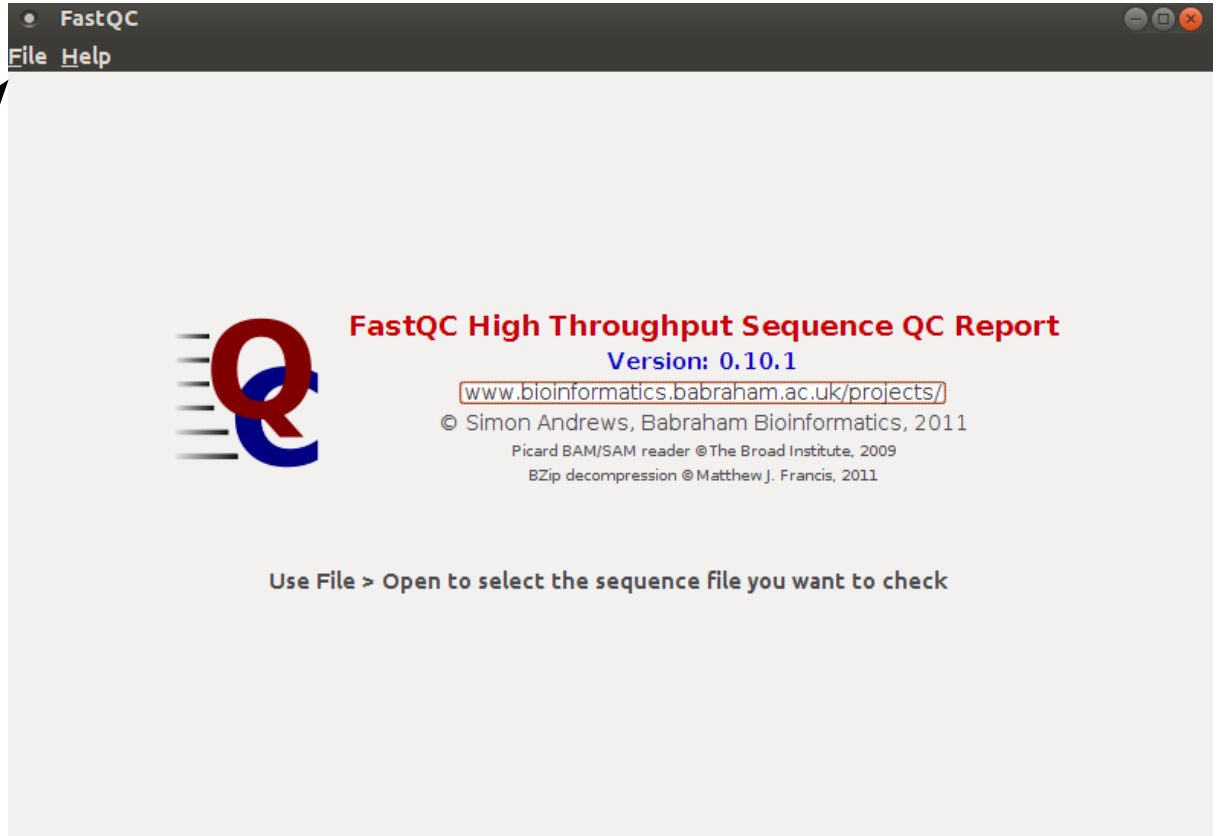
Controllo qualità: Fastqc

- Fastqc: controlla la qualità delle reads in uscita dal sequenziatore
  - Valuta la qualità delle reads generate
  - Produce indici di qualità (basi, adattatori, kmers,ecc)
  - Suggerisce il controllo di alcuni aspetti delle reads attraverso un codice colore (verde, arancione, giallo)

Controllo qualità: Fastqc

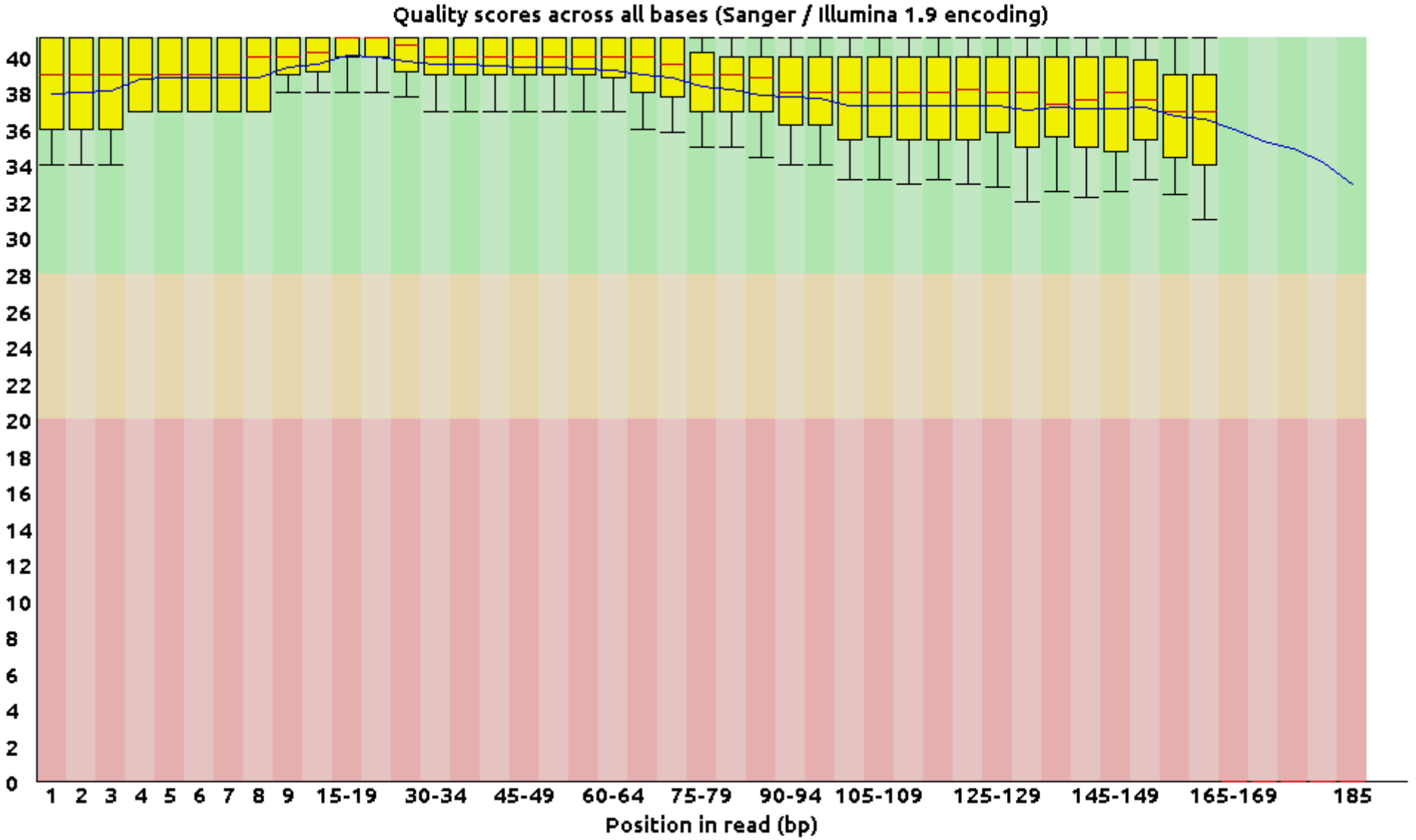
Per avviare il programma basta digitare “fastqc” sul terminale ed importare il file fastq di cui si vuole controllare la qualità

Click qui per importare un fastq



Controllo qualità: Fastqc

Oimyakon.fastq



Genoma di riferimento

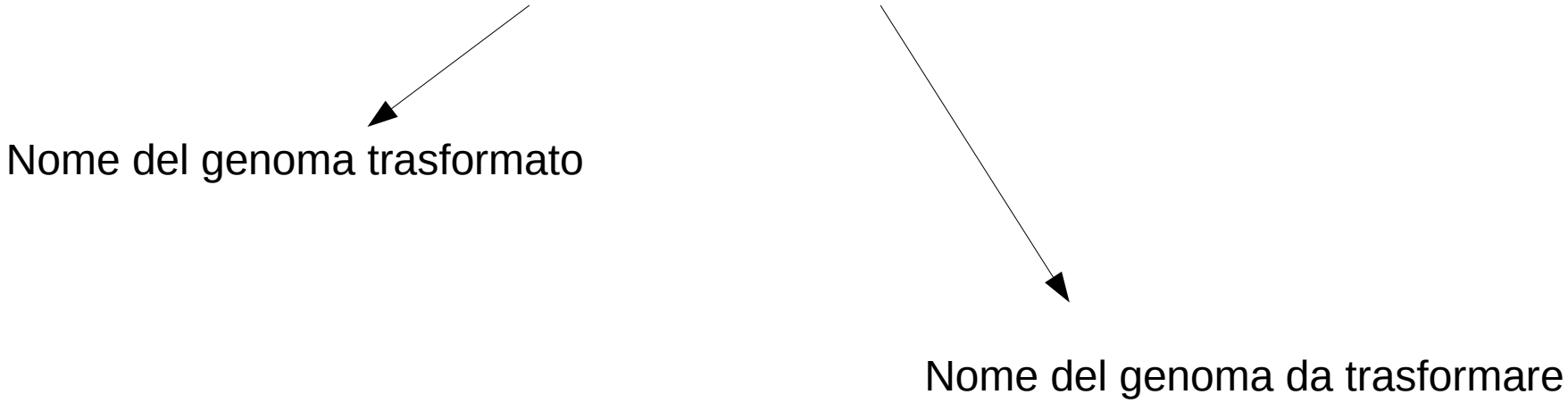
- L'allineamento di reads ad un genoma di riferimento utilizzando metodi di allineamento basati sulla trasformata di BW si compone di due fasi:
  - 1) Applicare la trasformata di BW al genoma di riferimento
  - 2) Allineare le reads utilizzando la trasformata

Preparazione del genoma

- La trasformata di BW viene applicata tramite il comando “index” del programma “bwa”:

  $ cd /media/studenti/bioinfo/risorse

  $ bwa index -p LoxAfr.chr3 LoxAfr.chr3.reference.fa

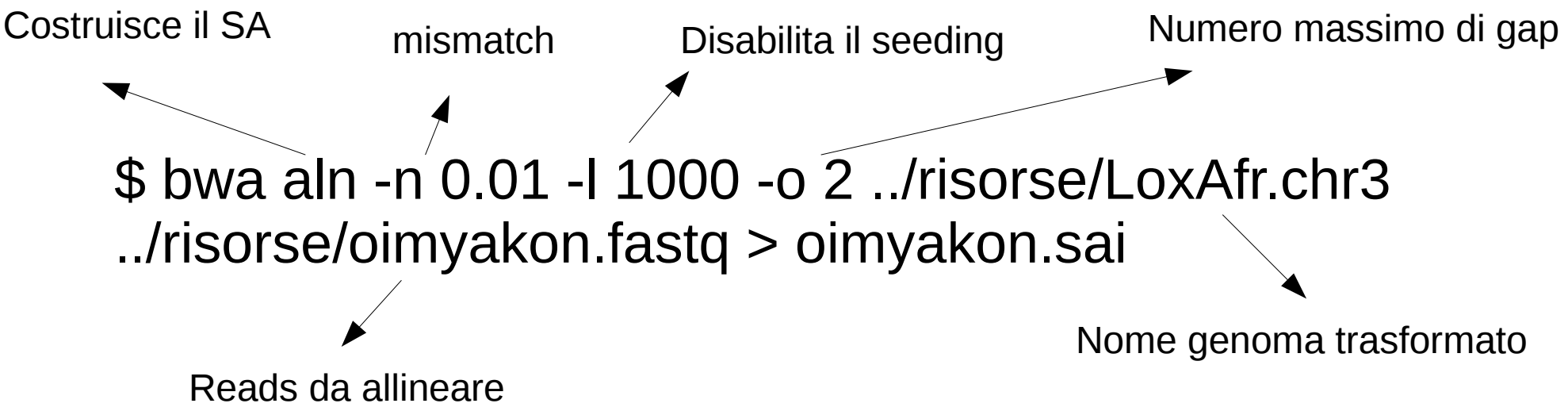


Allineamento delle reads

- Le reads vengono allineate al genoma utilizzando il programma “bwa”.

1) viene calcolato il “suffix array interval” per ogni read:

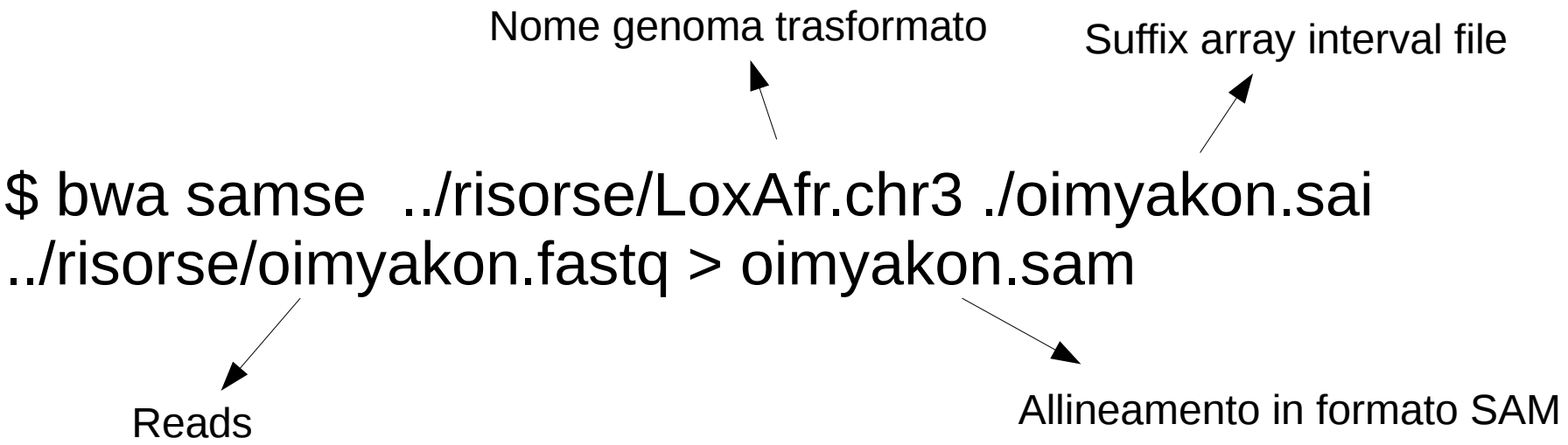
$ cd /media/studenti/bioinfo/Aulabio??



Parametri ottimizzati per reads da DNA antico

Allineamento delle reads

2) il suffix array interval di ogni reads viene usato per ottenere l'allineamento in formato sam:



$ samtools view -S -b oimyakon.sam > oimyakon.bam

Allineamento delle reads

- Quante reads abbiamo allineato?

  $ samtools view oimyakon.bam | wc -l

  $ samtools view -c oimyakon.bam

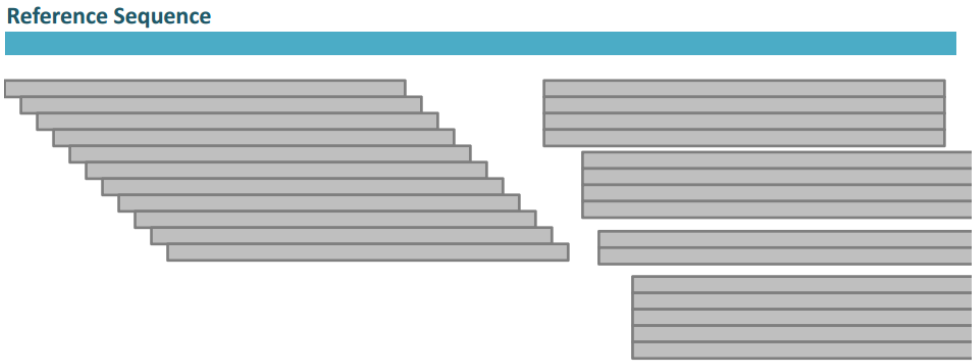
- Quante reads abbiamo allineato con buona confidenza (> MapQ20)?

  $ samtools view -q 20 -c oimyakon.bam

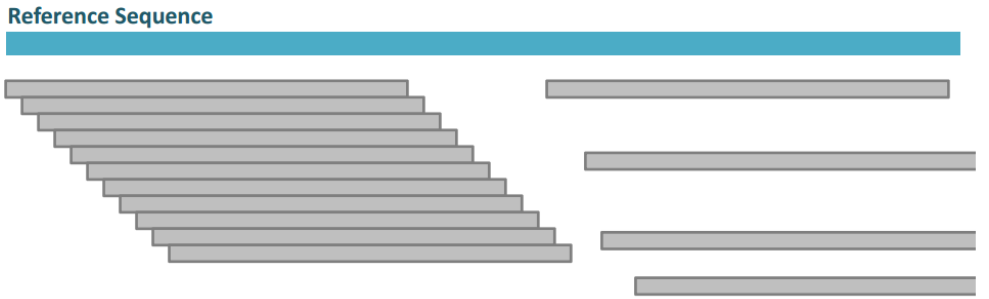
- Ripetere per le reads dell'individuo wrangel.

Rimozione dei duplicati di PCR

Prima della rimozione



Dopo la rimozione



Rimozione dei duplicati di PCR

- Prima di identificare i duplicati, l'allineamento (bam) deve essere ordinato per coordinata (chr, posizione):

  $ samtools sort oimyakon.bam oimyakon_sorted

- Identificazione e rimozione dei duplicati:

  $ samtools rmdup -s oimyakon_sorted.bam oimyakon_sorted_rmdup.bam

- Rimuovere i duplicati di PCR anche sul file wrangel.bam

Riallineamento degli indels

- Identificare le regioni che contengono INDELS e ri-allineare le reads al genoma di riferimento solo in queste regioni

- 1) indicizzare il genoma di riferimento

  $ samtools faidx ../risorse/LoxAfr.chr3.reference.fa

  $ picard-tools CreateSequenceDictionary.jar R=../risorse/LoxAfr.chr3.reference.fa O=../risorse/LoxAfr.chr3.reference.dict

- 2) aggiungo l'etichetta RG al file bam

  $ picard-tools AddOrReplaceReadGroups.jar I=oimyakon_sorted_rmdup.bam O=oimyakon_sorted_rmdup_rg.bam RGID=oim RGLB=oim RGPU=oim RGPL=illumina RGSM=oim VALIDATION_STRINGENCY=SILENT

- 3) indicizzo il file bam

  $ samtools index oimyakon_sorted_rmdup_rg.bam

Riallineamento degli indels

- 4) Identificare le regioni genomiche che contengono indels

  $ java -jar GATK -I oimyakon_sorted_rmdup_rg.bam -R ../risorse/LoxAfr.chr3.reference.fa -T RealignerTargetCreator -o oimyakon.intervals

- 5) Effettuo un ri-allineamento solo nelle regioni contenute nel file oimyakon.intervals

  $ java -jar GATK -I oimyakon_sorted_rmdup_rg.bam -R ../risorse/LoxAfr.chr3.reference.fa -T IndelRealigner -targetIntervals oimyakon.intervals -o oimyakon_sorted_rmdup_rg_real.bam

Visualizzare un BAM

- Le reads allineate possono essere visualizzate tramite diversi tool (IGV, tablet, samtools):
  - Samtools tview
  - Tool pratico per la visualizzazione
  - Visualizza le reads direttamente nel terminale
  - Non necessita di grandi risorse computazionali
  - Fornisce solo le informazioni principali associate alle reads

tview

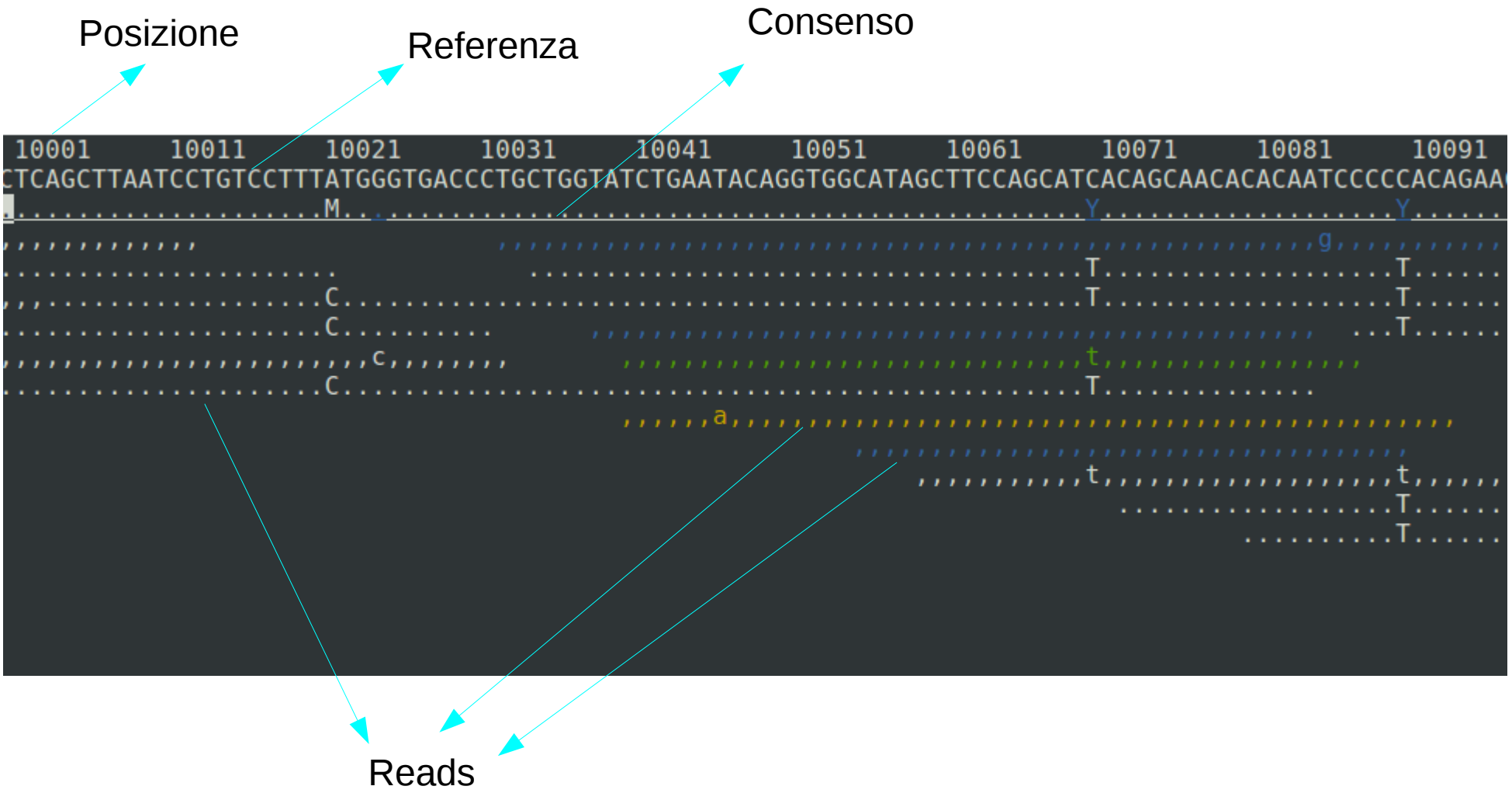
- $ samtools tview oimyakon_sorted_rmdup_rg_real.bam ../risorse/LoxAfr.chr3.reference.fa -p chr03:10000

Allineamento bam

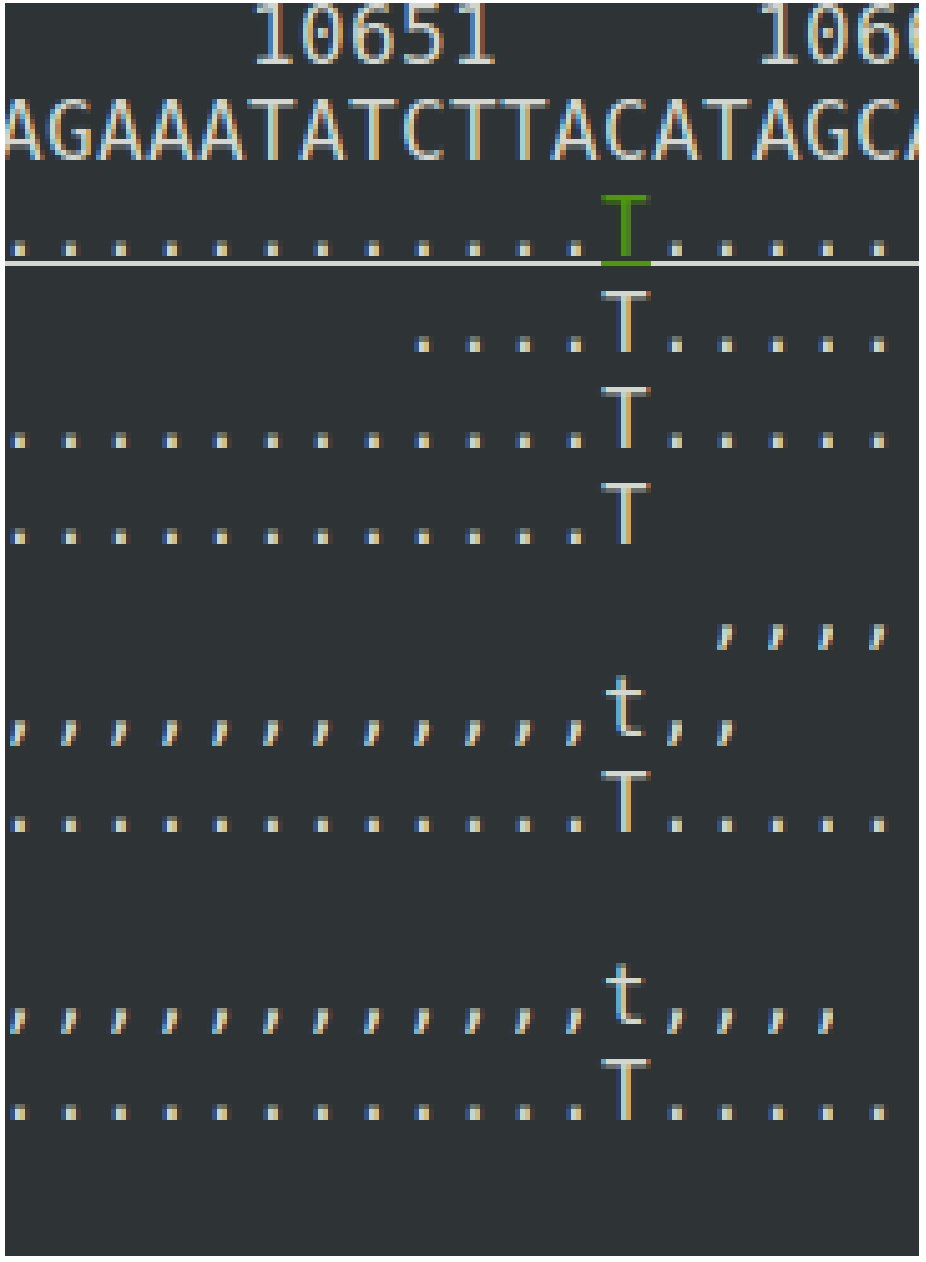
Referenza

Posizione cromosomica

tview


Posizione
Referenza
Consenso
10001 10011 10021 10031 10041 10051 10061 10071 10081 10091
Reads

tview

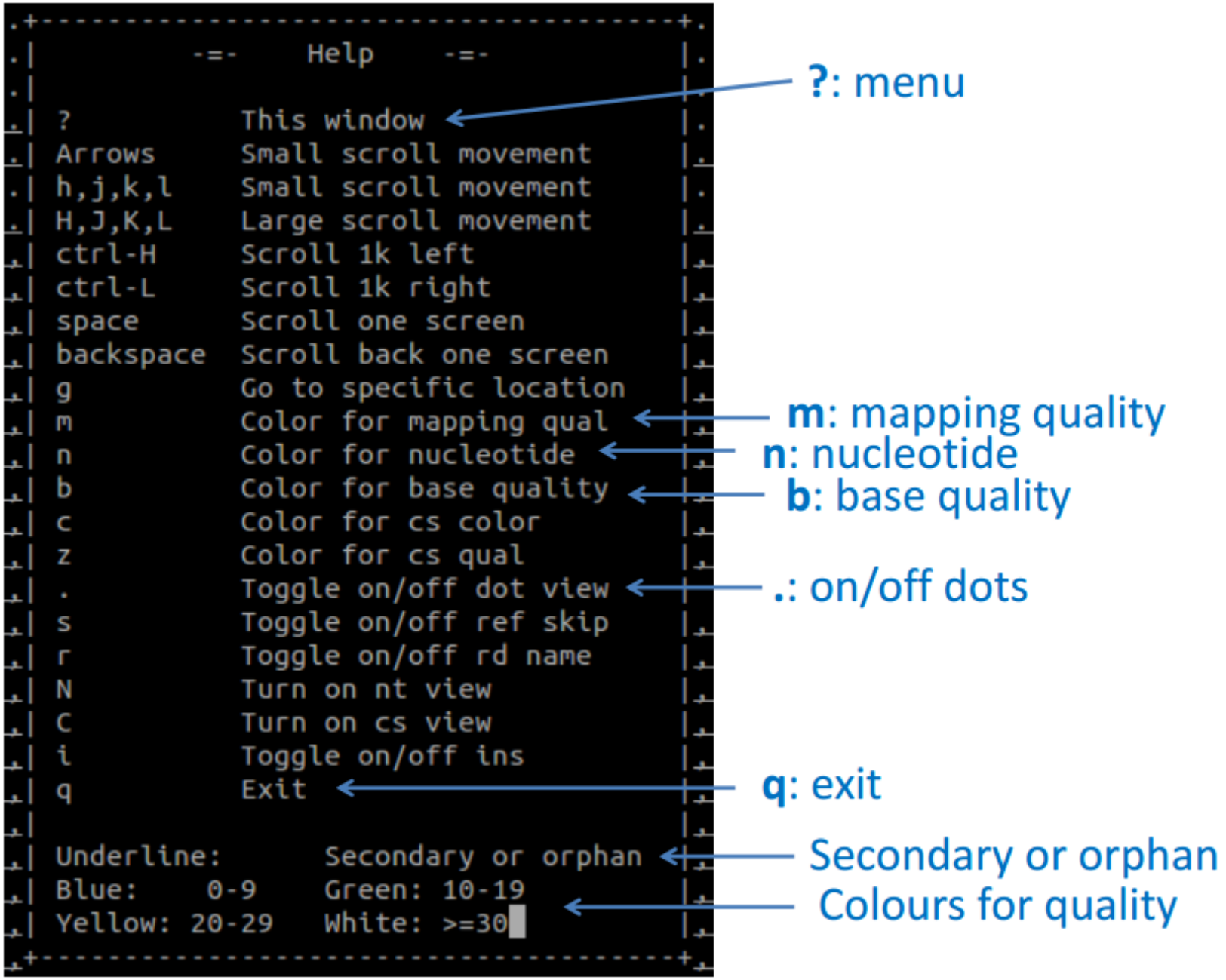


. : base reference su strand positiva

, : base reference su strand negativa

Base maiuscola: base alternativa su strand positiva

Base minuscola: base alternativa su strand negativa

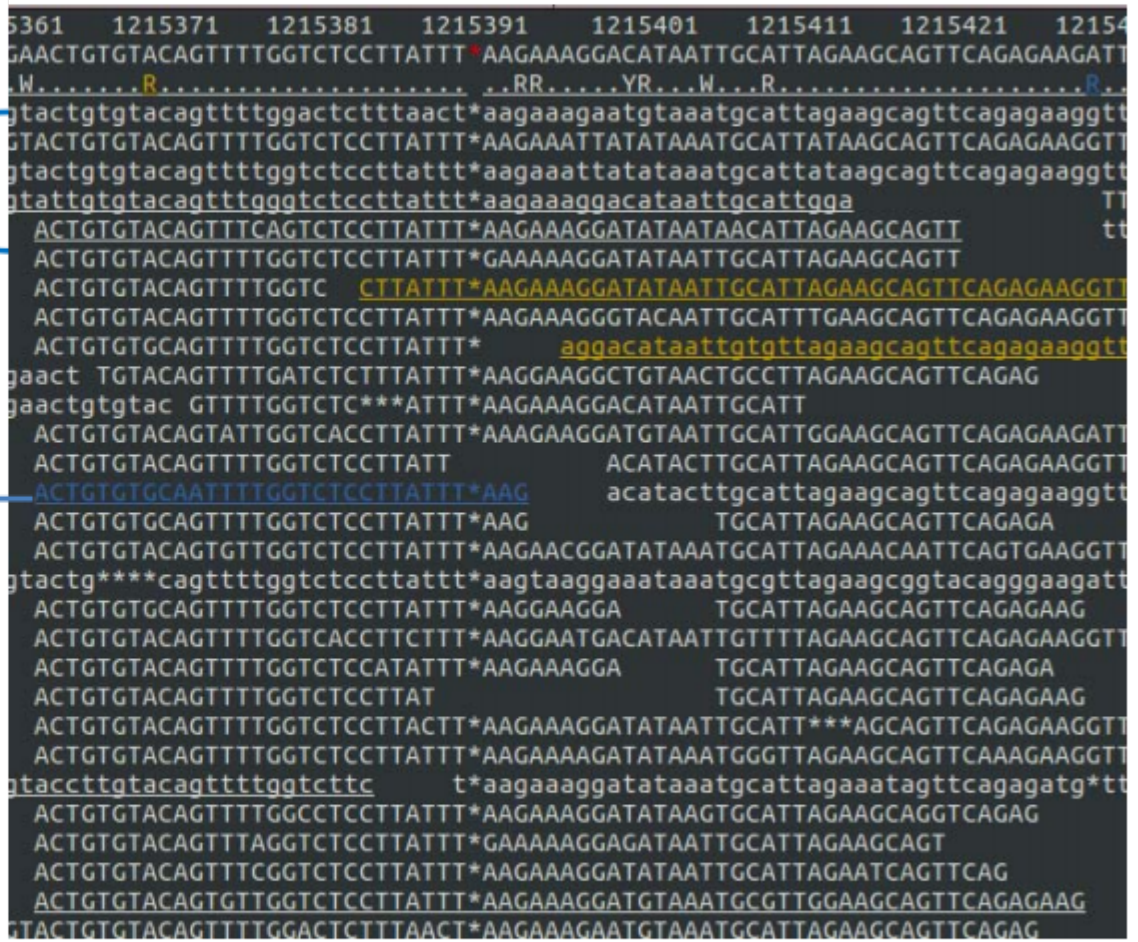

-=- Help -=-
? This window
Arrows Small scroll movement
h,j,k,l Small scroll movement
H,J,K,L Large scroll movement
ctrl-H Scroll 1k left
ctrl-L Scroll 1k right
space Scroll one screen
backspace Scroll back one screen
g Go to specific location
m Color for mapping qual
n Color for nucleotide
b Color for base quality
c Color for cs color
z Color for cs qual
. Toggle on/off dot view
s Toggle on/off ref skip
r Toggle on/off rd name
N Turn on nt view
C Turn on cs view
i Toggle on/off ins
q Exit
Underline: Secondary or orphan
Blue: 0-9 Green: 10-19
Yellow: 20-29 White: >=30
?: menu
m: mapping quality
n: nucleotide
b: base quality
.: on/off dots
q: exit
Secondary or orphan
Colours for quality

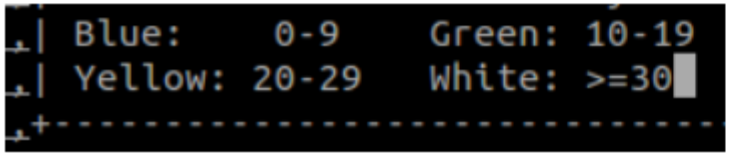
Press “ q ”, “ m ”

Press “ ? ”

MQ >=30

0 ≤ MQ ≤ 9





Press “ b ”

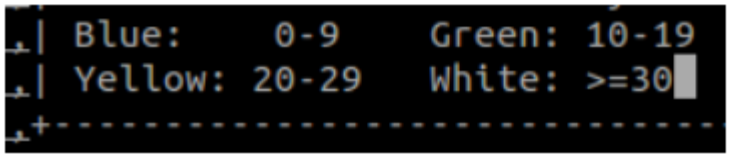
Press “ ? ”

20 ≤ BQ ≤ 29

10 ≤ BQ ≤ 19

BQ ≥ 30

0 ≤ BQ ≤ 9



Domande (oimyakon):

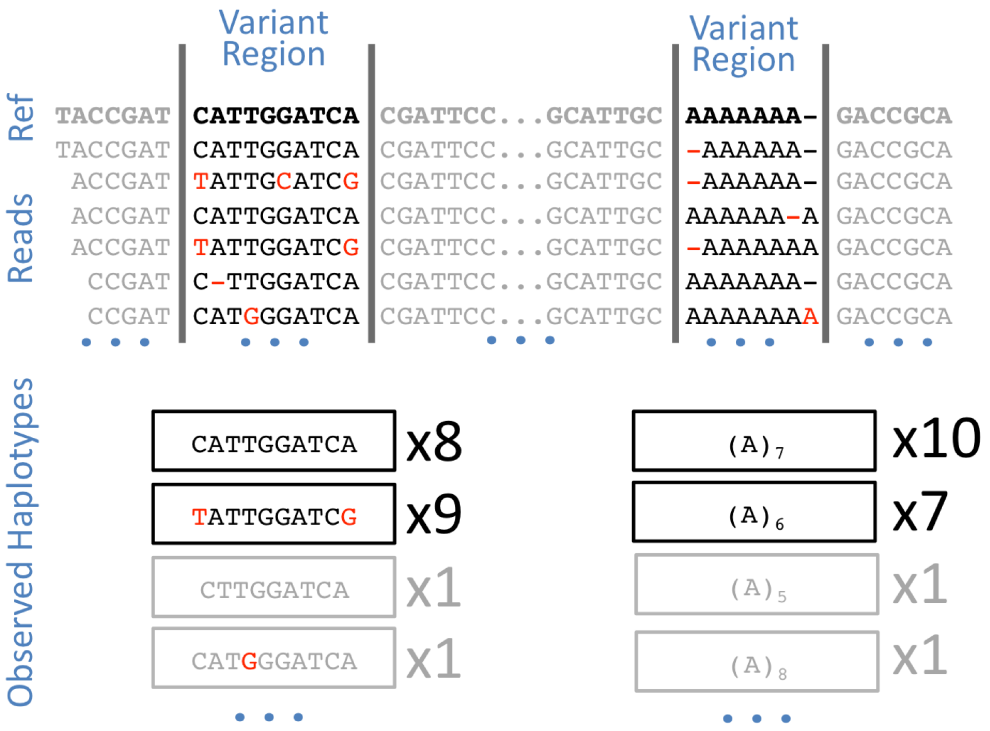
- Esiste un sito polimorfico alla base 102 ?
- Esiste un sito polimorfico alla base 611 ?
- Esiste un sito polimorfico alla base 7694 ?

Se si, elencare gli alleli ref & alt, la qualità media delle basi, la qualità media di allineamento delle reads e il possibile genotipo (AA/AR/RR)

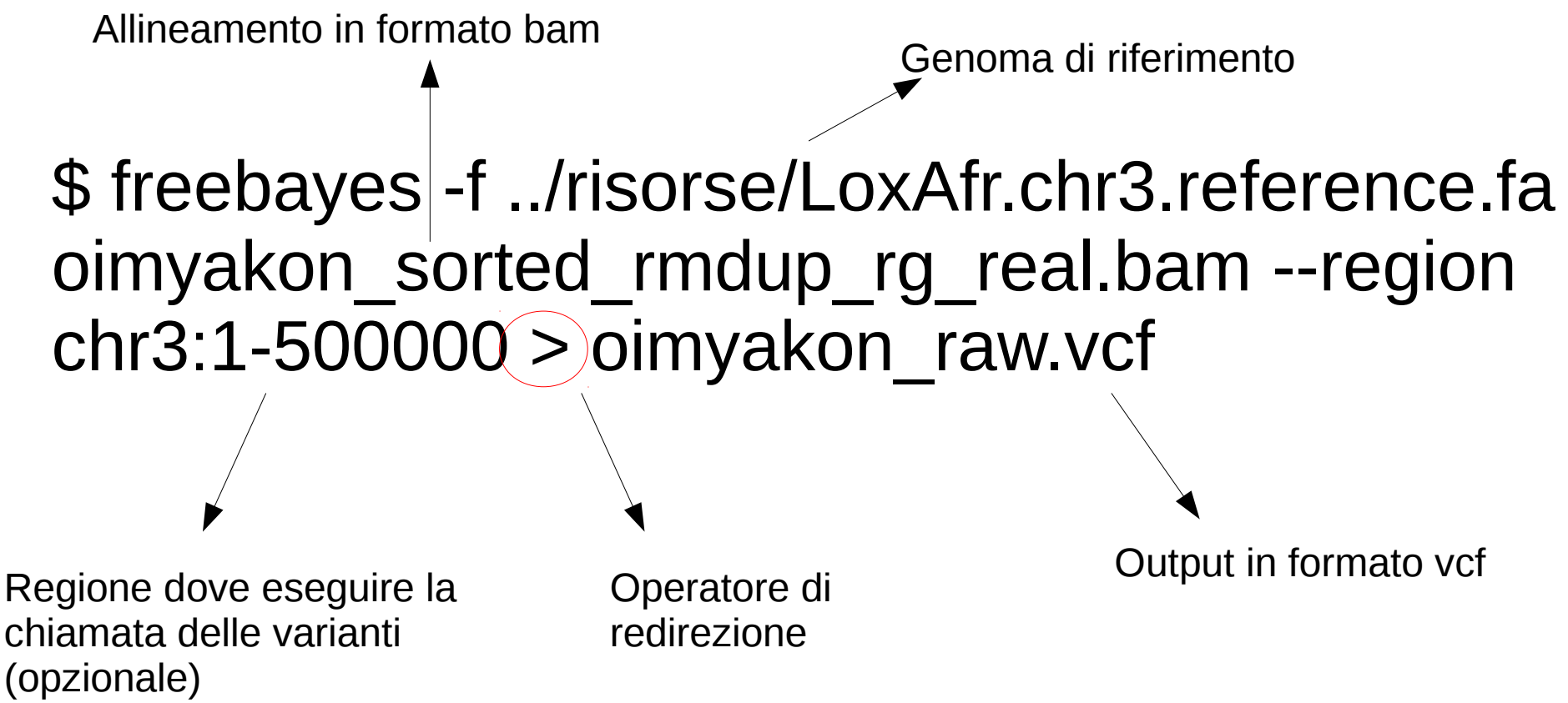
Chiamata delle varianti

- Abbiamo bisogno di metodi che identifichino i siti polimorfici e determinino i genotipi

***freebayes*, a haplotype-based variant detector**



Chiamata delle varianti



Freebayes identificherà le posizioni polimorfiche e attribuirà il genotipo più probabile per ognuna, e memorizzerà le informazioni in formato vcf

Chiamata delle varianti

- VCF (intestazione)

```
##fileformat=VCFv4.1
##fileDate=20170515
##source=freeBayes v1.0.2-29-g41c1313
##reference=LoxAfr.chr3.reference.fa
##contig=<ID=chr3,length=205080690>
##phasing=none
##commandline="/opt/software/ngs/freebayes/bin/freebayes -f LoxAfr.chr3.reference.fa test_sorted_rmdup_rg_real.bam"
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of samples with data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total read depth at the locus">
##INFO=<ID=DPB,Number=1,Type=Float,Description="Total read depth per bp at the locus; bases in reads overlapping / bases in haplotype">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Total number of alternate alleles in called genotypes">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=AF,Number=A,Type=Float,Description="Estimated allele frequency in the range (0,1]">
##INFO=<ID=RO,Number=1,Type=Integer,Description="Reference allele observation count, with partial observations recorded fractionally">
##INFO=<ID=AO,Number=A,Type=Integer,Description="Alternate allele observations, with partial observations recorded fractionally">
##INFO=<ID=PRO,Number=1,Type=Float,Description="Reference allele observation count, with partial observations recorded fractionally">
##INFO=<ID=PAO,Number=A,Type=Float,Description="Alternate allele observations, with partial observations recorded fractionally">
##INFO=<ID=QR,Number=1,Type=Integer,Description="Reference allele quality sum in phred">
##INFO=<ID=QA,Number=A,Type=Integer,Description="Alternate allele quality sum in phred">
##INFO=<ID=PQR,Number=1,Type=Float,Description="Reference allele quality sum in phred for partial observations">
##INFO=<ID=PQA,Number=A,Type=Float,Description="Alternate allele quality sum in phred for partial observations">
##INFO=<ID=SRF,Number=1,Type=Integer,Description="Number of reference observations on the forward strand">
##INFO=<ID=SRR,Number=1,Type=Integer,Description="Number of reference observations on the reverse strand">
##INFO=<ID=SAF,Number=A,Type=Integer,Description="Number of alternate observations on the forward strand">
##INFO=<ID=SAR,Number=A,Type=Integer,Description="Number of alternate observations on the reverse strand">
```

...

Chiamata delle varianti

- VCF (corpo)

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  oim
chr3    63      .       C       G       33.7366 .       AB=0.428571;ABP=3.32051;AC=1;AF=0.5;AN=2;AO=3;CIGAR=1X;DP=7;DPB=7;DPRA=0;EPP=3.73412;EPPR=3.0103;GTI=0;LEN=1;MEANALT=1;MQM=37;MQMR=37;N
S=1;NUMALT=1;ODDS=7.76772;PAIRED=0;PAIREDR=0;PAO=0;PQA=0;PQR=0;PRO=0;QA=118;QR=160;RO=4;RPL=0;RPP=9.52472;RPPR=5.18177;RPR=3;RUN=1;SAF=1;SAP=3.73412;SAR=2;SRF=3;SRP=5.18177;SRR=1;TYPE=snp;tec
hnology.illumina=1      GT:DP:DPR:RO:QR:AO:QA:GL        0/1:7:7,3:4:160:3:118:-7.69165,0,-10.9156
chr3    85      .       A       G       13.083  .       AB=0.25;ABP=9.52472;AC=1;AF=0.5;AN=2;AO=3;CIGAR=1X;DP=12;DPB=12;DPRA=0;EPP=3.73412;EPPR=3.25157;GTI=0;LEN=1;MEANALT=1;MQM=37;MQMR=37;NS
=1;NUMALT=1;ODDS=2.96205;PAIRED=0;PAIREDR=0;PAO=0;PQA=0;PQR=0;PRO=0;QA=122;QR=358;RO=9;RPL=2;RPP=3.73412;RPPR=3.25157;RPR=1;RUN=1;SAF=0;SAP=9.52472;SAR=3;SRF=7;SRP=9.04217;SRR=2;TYPE=snp;tech
nology.illumina=1       GT:DP:DPR:RO:QR:AO:QA:GL        0/1:12:12,3:9:358:3:122:-6.31137,0,-25.0856
chr3    102     .       C       T       88.2584 .       AB=0.333333;ABP=7.35324;AC=1;AF=0.5;AN=2;AO=6;CIGAR=1X;DP=18;DPB=18;DPRA=0;EPP=4.45795;EPPR=3.73412;GTI=0;LEN=1;MEANALT=1;MQM=37;MQMR=3
7;NS=1;NUMALT=1;ODDS=20.3222;PAIRED=0;PAIREDR=0;PAO=0;PQA=0;PQR=0;PRO=0;QA=238;QR=464;RO=12;RPL=4;RPP=4.45795;RPPR=3.73412;RPR=2;RUN=1;SAF=4;SAP=4.45795;SAR=2;SRF=4;SRP=5.9056;SRR=8;TYPE=snp;
technology.illumina=1   GT:DP:DPR:RO:QR:AO:QA:GL        0/1:18:18,6:12:464:6:238:-13.866,0,-32.2601
```

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  FORMAT  oim
chr3    63      .       C       G       33.7366 .       GT:DP:DPR:RO:QR:AO:QA:GL        0/1:7:7,3:4:160:3:118:-7.69165,0,-10.9156
chr3    85      .       A       G       13.083  .       GT:DP:DPR:RO:QR:AO:QA:GL        0/1:12:12,3:9:358:3:122:-6.31137,0,-25.0856
chr3    102     .       C       T       88.2584 .       GT:DP:DPR:RO:QR:AO:QA:GL        0/1:18:18,6:12:464:6:238:-13.866,0,-32.2601
chr3    157     .       G       A       53.5175 .       GT:DP:DPR:RO:QR:AO:QA:GL        1/1:2:2,2:0:0:2:77:-6.5556,-0.60206,0
chr3    278     .       A       G       88.3519 .       GT:DP:DPR:RO:QR:AO:QA:GL        0/1:5:5,4:1:41:4:154:-11.2499,0,-2.04933
chr3    334     .       G       A       109.254 .       GT:DP:DPR:RO:QR:AO:QA:GL        1/1:4:4,4:0:0:4:151:-12.5555,-1.20412,0
chr3    345     .       T       C       172.722 .       GT:DP:DPR:RO:QR:AO:QA:GL        1/1:6:6,6:0:0:6:236:-19.1625,-1.80618,0
```

Non tutti i polimorfismi hanno la stessa qualità (Phread Scaled)

Non tutti i polimorfismi hanno le stesse caratteristiche

Necessario definire dei filtri di qualià e trattenere solo i polimorfismi che li soddisfino

Filtro delle varianti

- I filtri dipendono dalle caratteristiche dell'esperimento. Non ci sono filtri universali. Due tipi di filtri:

- Filtri sull'input, che modulano le caratteristiche delle reads che serviranno per scoprire i siti polimorfici:
  - Usare solo le reads con MQ > 20
  - Usare solo le basi con BQ > 20
  - Una copertura di alemeno 5 reads

- Filtri sulle varianti, che eliminano i siti con determinate caratteristiche. Ad esempio potremmo voler tenere solo le posizioni polimorfiche che:
  - Hanno una probabilità di essere polimorfiche maggiore del 99% (Q > 20)
  - Strand bias non significativo

Guardiamo le opzioni di freebayes

$ freebayes -h

Filtro delle varianti

I filtri sull'input vanno specificati su freebayes:

MQ > 20 : -m 20

BQ > 20 : -q 20

Copertura di almento 5 reads: --min-coverage 5

$ freebayes -m 20 -q 20 –min-coverage 5 -f ../risorse/LoxAfr.chr3.reference.fa oimyakon_sorted_rmdup_rg_real.bam > oimyakon_raw.vcf

Filtro delle varianti

I filtri sulle varianti vanno applicati al file vcf tramite tool specifici, ad esempio “vcffilter”:

Probabilità del polimorfismo > 99% (QUAL > 20)

Strand bias assente (P=0.05;Phred=13; SRP < 13 & SAP < 13)

La descrizione di questi due parametri è presente nell'intestazione del vcf:

##INFO=<ID=SRP,Number=1,Type=Float,Description="Strand balance probability for the reference allele: Phred-scaled upper-bounds estimate of the probability of observing the deviation between SRF and SRR given E(SRF/SRR) ~ 0.5, derived using Hoeffding's inequality">

##INFO=<ID=SAP,Number=A,Type=Float,Description="Strand balance probability for the alternate allele: Phred-scaled upper-bounds estimate of the probability of observing the deviation between SAF and SAR given E(SAF/SAR) ~ 0.5, derived using Hoeffding's inequality">

$ vcffilter -f “QUAL > 20” -f “SRP < 13” -f “SAP < 13” oimyakon_raw.vcf > oimyakon_filt.vcf

Quante varianti?

- Per contare il numero di varianti:

  $ grep -v “^#” oimyakon_filt.vcf | wc -l

- Quanti SNP abbiamo trovato?

- Quanti INDELs abbiamo trovato?

Esercizio

- Replicare la procedura di chiamata delle varianti per l'individuo “wrangel”.
- Quanti SNP identifichiamo nel mammuth più recente?
- Quale individuo mostra più siti eterozigoti?