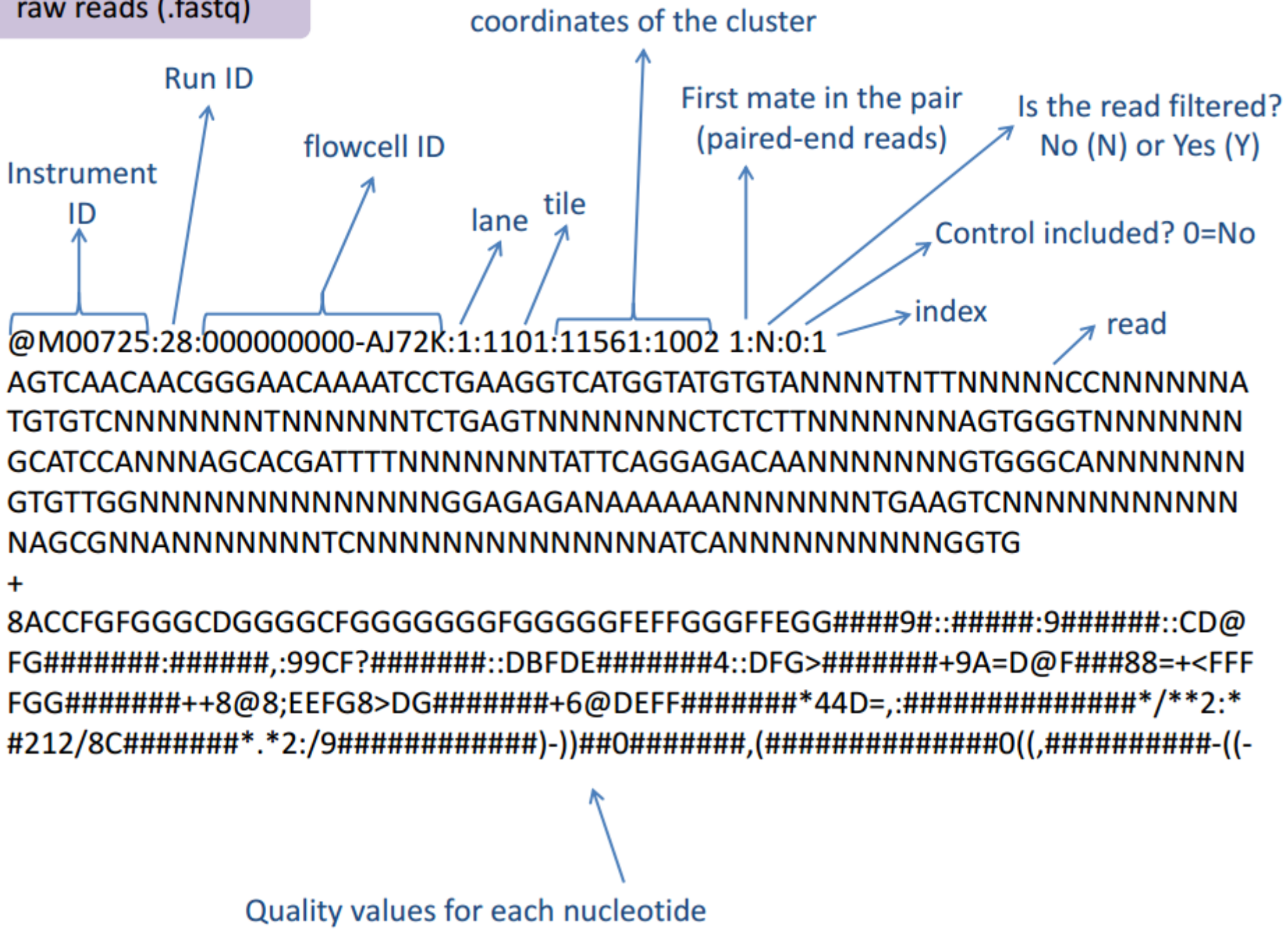
Lezione 7: Allineamento delle reads ad un genoma di riferimento

Il formato fastq

```
@M00725:28:000000000-AJ72K:1:1101:11561:1002 1:N:0:1
AGTCAACAACGGGAACAAAATCCTGAAGGTCATGGTATGTGTANNNNTNTTNNNNNCCNNNNNNA
TGTGTCNNNNNNNTNNNNNNTCTGAGTNNNNNNNCTCTCTTNNNNNNNAGTGGGTNNNNNNNN
GCATCCANNNAGCACGATTTTNNNNNNNTATTCAGGAGACAANNNNNNNGTGGGCANNNNNNN
GTGTTGGNNNNNNNNNNNNNNGGAGAGANAAAAAANNNNNNNTGAAGTCNNNNNNNNNNN
NAGCGNNANNNNNNNTCNNNNNNNNNNNNNNATCANNNNNNNNNGGTG
+
8ACCFGFGGGCDGGGGCFGGGGGGGFGGGGGFEFFGGGFFEGG####9#::#####:9######::CD@
FG#######:######,:99CF?#######::DBFDE#######4::DFG>#######+9A=D@F###88=+<FFF
FGG#######++8@8;EEFG8>DG#######+6@DEFF#######*44D=,:###############*/**2:*
#212/8C#######*.*2:/9############)-))##0#######,(###############0((,##########-((-
```

Un sequenziatore NGS produce milioni di “reads” cioè combinazioni di 4 righe di testo:
1° riga - Descrittore → include informazioni su dove e come è stata generata la read
2° riga - Basi nucleotidiche → I nucleotidi che compongono la sequenza di DNA
3° riga - Separatore → Divide la riga contenente i nucleotidi da quella con le qualità
4° riga - Qualità della lettura → Caratteri che codificano la confidenza nella chiamata di ogni nucleotide presente alla riga2

raw reads (.fastq)



Qualità della lettura di ogni base

- Phred-score Quality

- $Q = - 10 \log_{10} P$

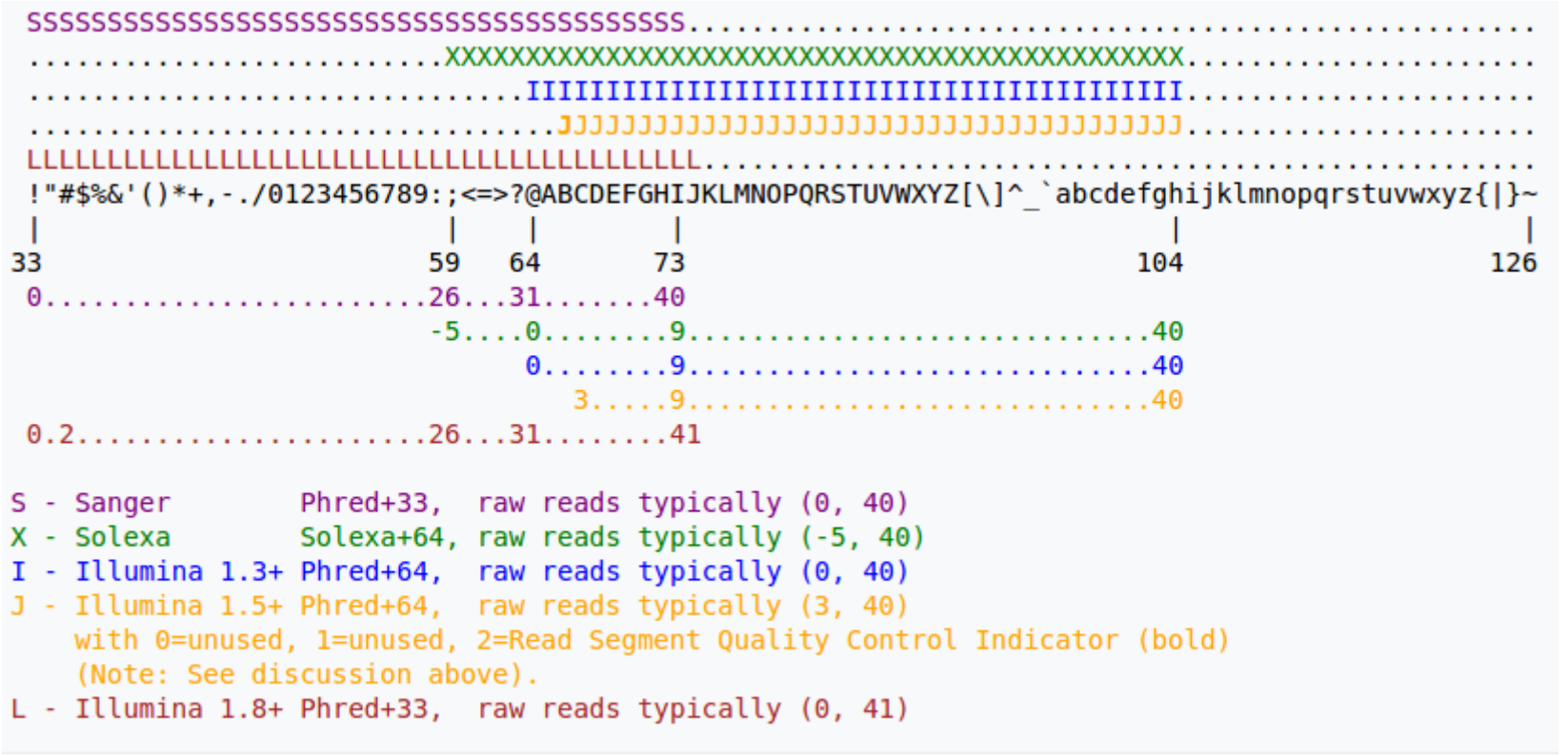
  Q is Phred-score quality

  P is base-calling error probability *P*

Phred quality scores are logarithmically linked to error probabilities Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %
50	1 in 100000	99.999 %

Codifica della qualità di lettura

- Sanger format can encode a Phred quality score from 0 to 93 using ASCII 33 to 126 (Phred+33)



Paired-end reads

- Nel caso di paired-end reads avremo due file separati contenenti coppie di reads (forward, reverse) posizionate nelle stesse righe

raw reads (.fastq)

Same lane, different read mate in the pair!

Prima posizione nel file R1

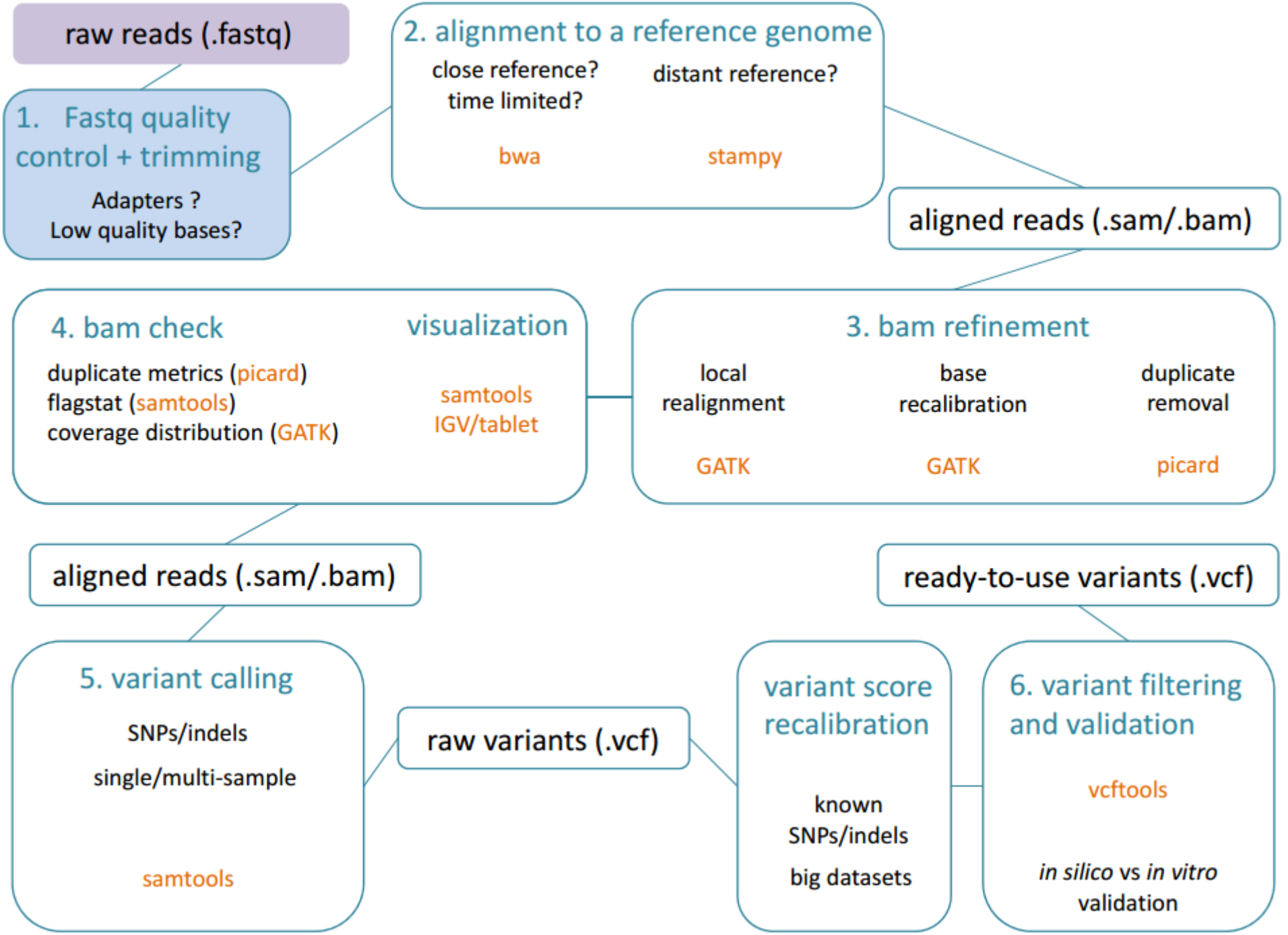
cc_gn2_R1_trimmed.fastq

@M00725:28:000000000-AJ72K:1:1101:11561:1002 1:N:0:1
AGTCAACAACGGGAACAAAATCCTGAAGGTCATGGTATGTGTANNNNTNTTNNNNNCCNNNNNNA
TGTGTCNNNNNNNTNNNNNNTCTGAGTNNNNNNNCTCTCTTNNNNNNNAGTGGGTNNNNNNN
GCATCCANNNAGCACGATTTTNNNNNNNTATTCAGGAGACAANNNNNNNGTGGGCANNNNNNN
GTGTTGGNNNNNNNNNNNNNGGAGAGANAAAAAANNNNNNNTGAAGTCNNNNNNNNNNN
NAGCGNNANNNNNNNTCNNNNNNNNNNNNNNATCANNNNNNNNNGGTG

Prima posizione nel file R2

cc_gn2_R2_trimmed.fastq

@M00725:28:000000000-AJ72K:1:1101:11561:1002 2:N:0:1
CCATTTCTNNNNNNNAGGACCTNNNNNNNAGCCCTNNNNNNNNNNNNNAGNATATGANNNNN
NNTCTTATTNANCCANNNTCTAGNNNNNNNCTTTCCTNNNNNNNTCTCTGANNNNNNNNNNN
NNCCCTTCCNNTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTTNTCTCNTNNNNNNNN
NNNNAAAATCCNNNNNNNNNNNNNCCACTAANNNNNNNNNNNNNNAAGAAATAACACACNN
NNNNNACAAAAANNNNNNNACAACACNNNNNNNGCATAAANNNA


raw reads (.fastq)
1. Fastq quality control + trimming
Adapters ?
Low quality bases?
2. alignment to a reference genome
close reference?
time limited?
bwa
distant reference?
stampy
aligned reads (.sam/.bam)
3. bam refinement
local realignment
GATK
base recalibration
GATK
duplicate removal
picard
4. bam check
duplicate metrics (picard)
flagstat (samtools)
coverage distribution (GATK)
visualization
samtools
IGV/tablet
aligned reads (.sam/.bam)
5. variant calling
SNPs/indels
single/multi-sample
samtools
raw variants (.vcf)
variant score recalibration
known SNPs/indels
big datasets
ready-to-use variants (.vcf)
6. variant filtering and validation
vcftools
in silico vs in vitro validation

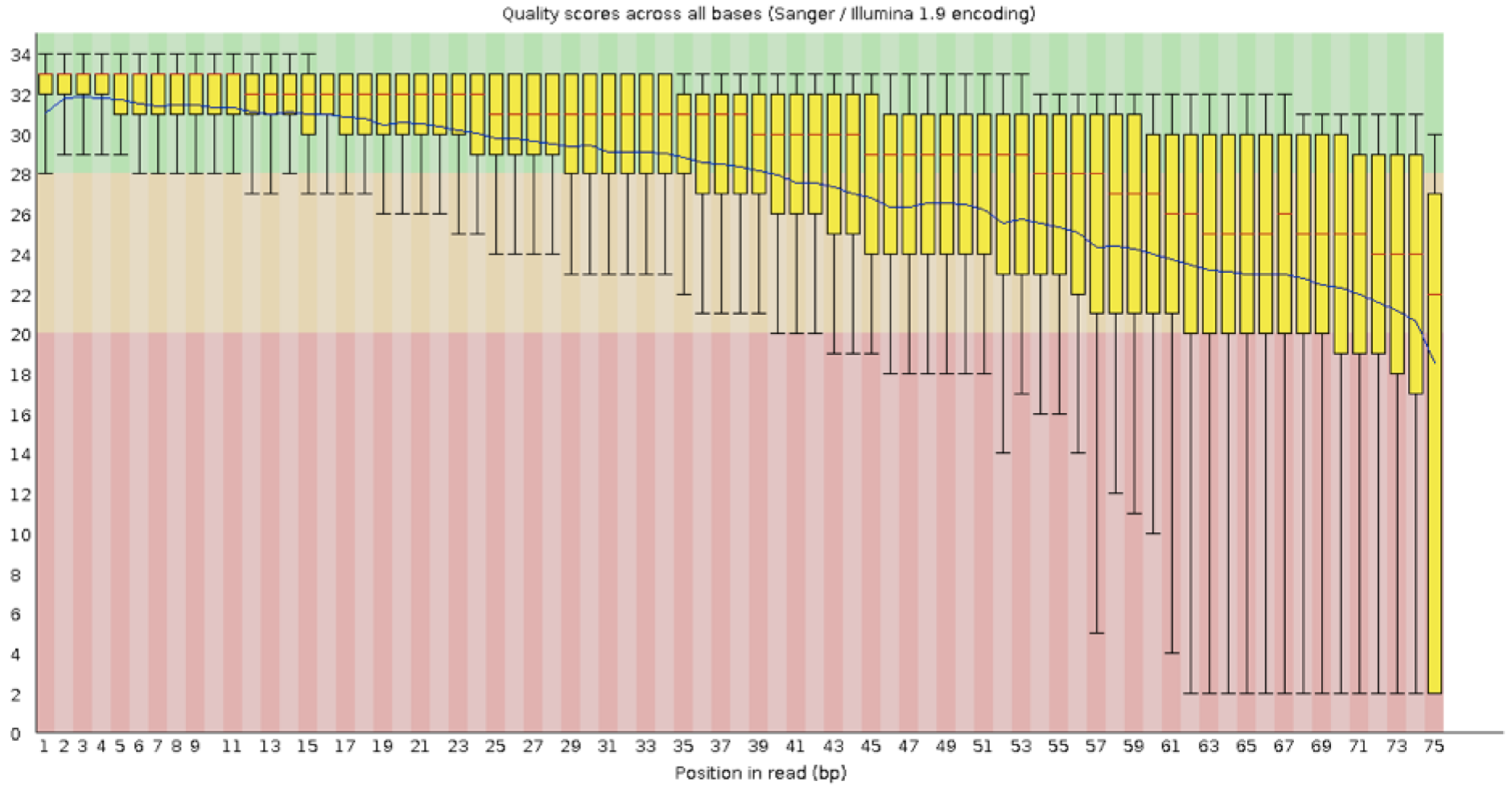
Controllo qualità delle reads

- Non tutte le basi che compongono una reads hanno lo stesso livello di qualità
  - La qualità generalmente tende a diminuire più ci avviciniamo al 3'
  - E' necessario verificare in ogni esperimento di sequenziamento come varia la qualità al variare della posizione sulla read
  - Basi con qualità < 20 vengono generalmente rimosse (trimmed)

Controllo qualità delle reads

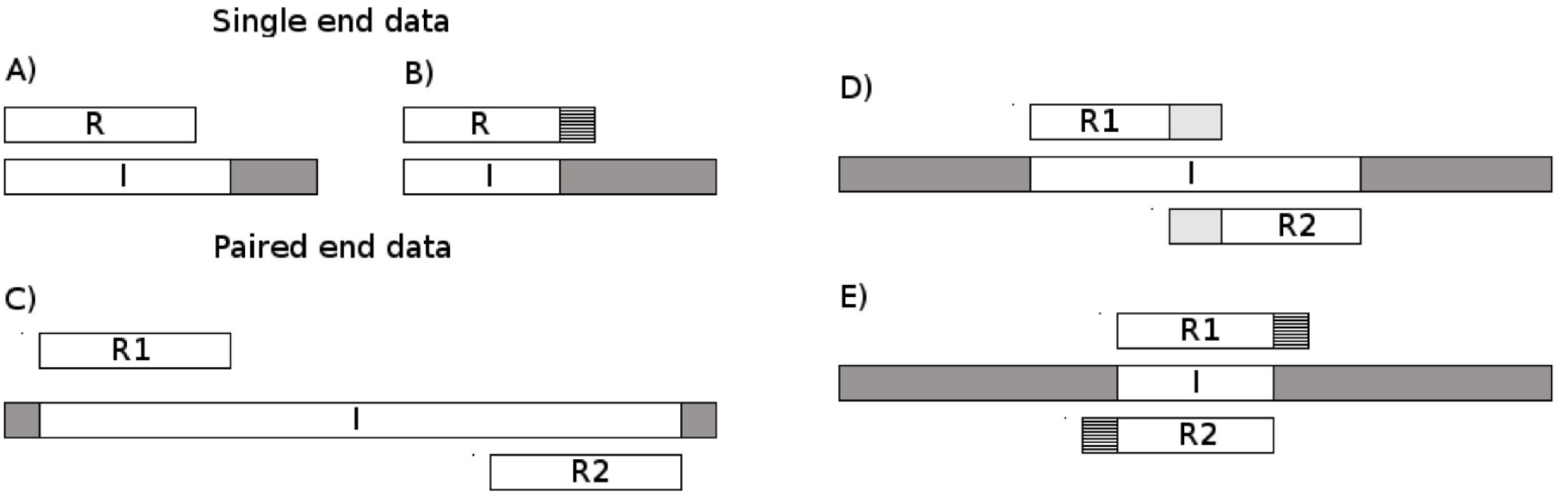
- Procedendo dal 3' verso il 5' si rimuovono nucleotidi da ogni reads fino a raggiungere una qualità minima (Phred quality score >=20)

Per base sequence quality



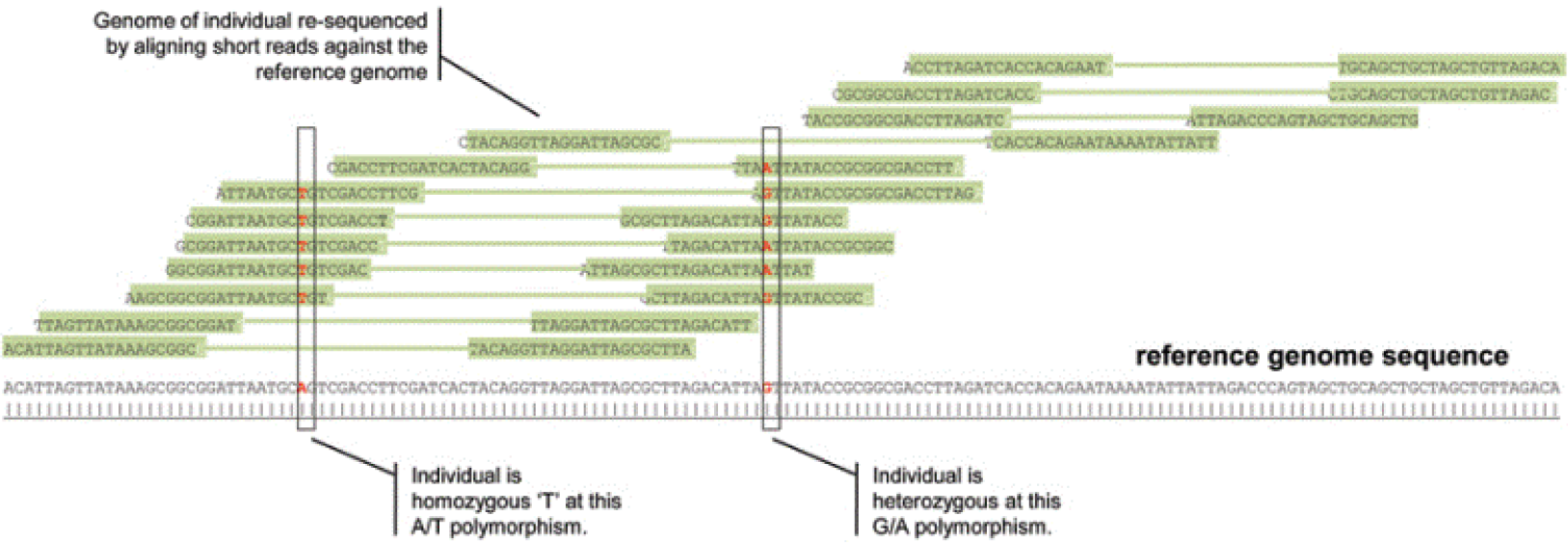
Controllo qualità delle reads

- In alcuni casi è possibile che alcune reads contengano porzioni di adattatori (Illumina:p5/p7)
  - Se non rimossi possono originare problemi di allineamento o falsi polimorfismi
  - Le sequenze nucleotidiche degli adattatori sono conosciute quindi è possibile identificarle e rimuoverle



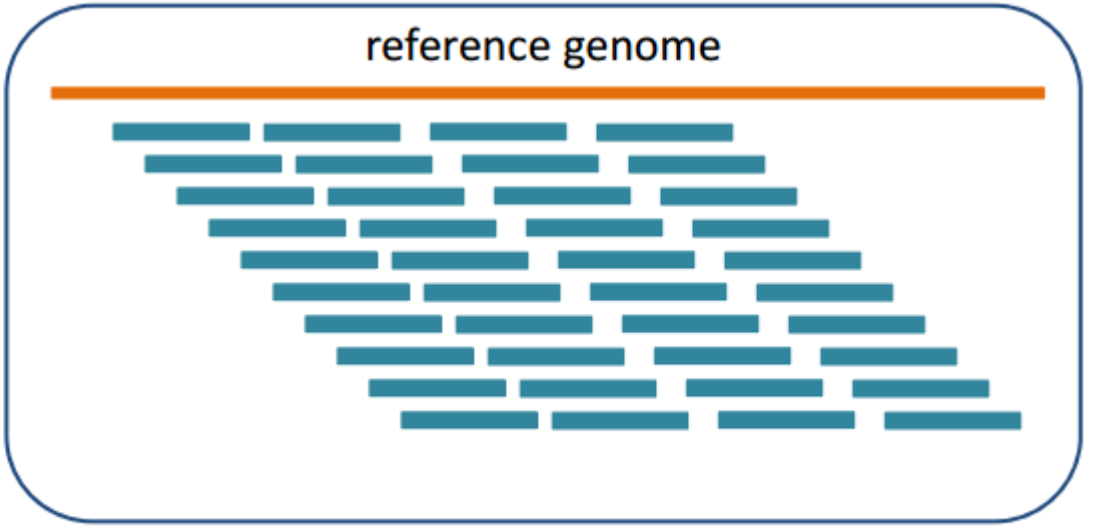
Allineamento a un genoma di riferimento

- Genoma di riferimento: una o più sequenze di DNA che rappresentano il genoma di un organismo



Allineamento

- Allineamento: identificare la posizione delle reads rispetto al genoma di riferimento
- Processo nel quale si determina **la posizione di provenienza più probabile** di una read all'interno del genoma



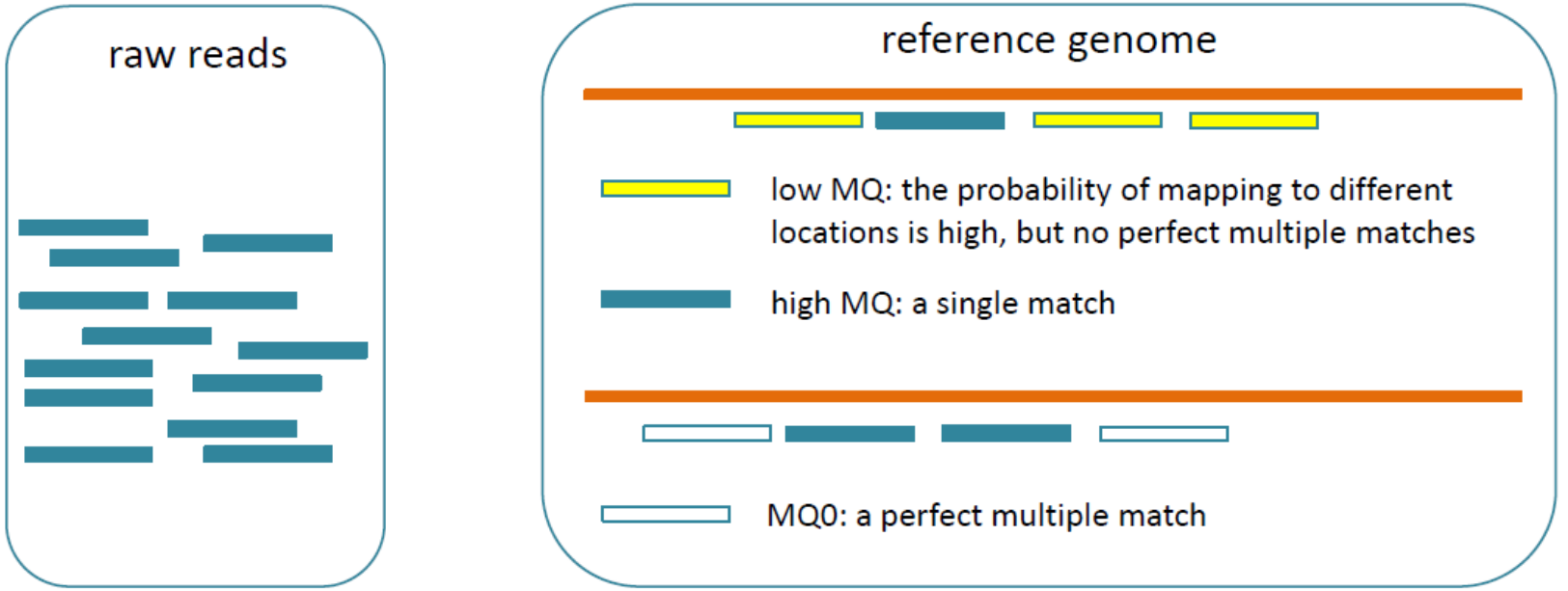
Qualità dell'allineamento

alignment to a reference genome – mapping qualities (MQ)

What if there are several possible places to align your sequencing read?
This may be due to:

- Repeated elements in the genome
- Low complexity sequences
- Reference errors and gaps

MQ is a phred-score of the quality of the alignment
With paired-end reads: mapping quality is determined on the pair, thus even if one read can be mapped in several places, the mapping of its pair can help to locate it properly.



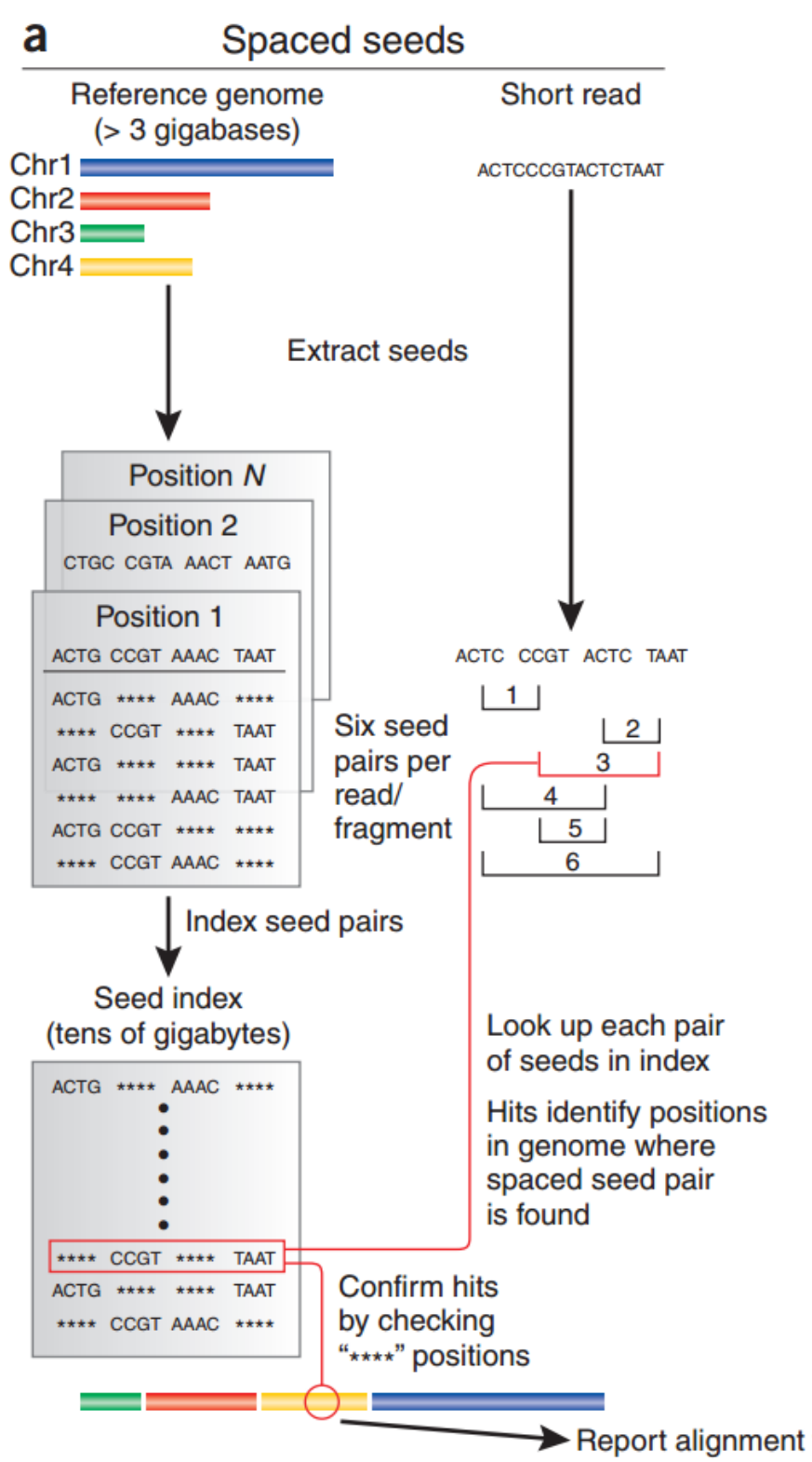
Metodi di allineamento

- Un buon metodo di allineamento dovrebbe:
  - Gestire grandi **quantità** di reads: allineare centinaia di milioni di reads velocemente e senza richiedere enormi risorse computazionali
  - Gestire sequenze di DNA **corte**: identificare la posizione genomica di piccole sequenze di DNA (35-300bp) con sicurezza
  - Buon rapporto tra **sensibilità e velocità** di ricerca: identificare velocemente la giusta posizione genomica
  - Gestire **genomi** di riferimento di **grandi dimensioni**: ricercare in maniera veloce le posizioni delle reads in genomi lunghi alcune GB senza richiedere enormi risorse computazionali

Metodi di allineamento

- Spaced-seed indexing: divisione del genoma e delle reads in sottoinsiemi (seed) e ricerca delle corrispondeze tramite l'uso di indici

- Trasformata di Burrows-Wheeler: compressione del genoma di riferimento per una ricerca più veloce



Spaced seeds

- Spezzare il genoma in una serie di 4 sottosequenze di lunghezza uguale, chiamati “seed”
- Creare le possibili sei combinazioni di due coppie di seed lungo il genoma
- Creare un indice delle possibili combinazioni (molto voluminoso)
- Ripetere il processo di creazione dei seed per ogni read
- Identificare il match tra i seed sulla read e quelli nell'indice
- SNPs e INDELs saranno tollerati grazie agli “spazi” (****) tra i seed

Svantaggi

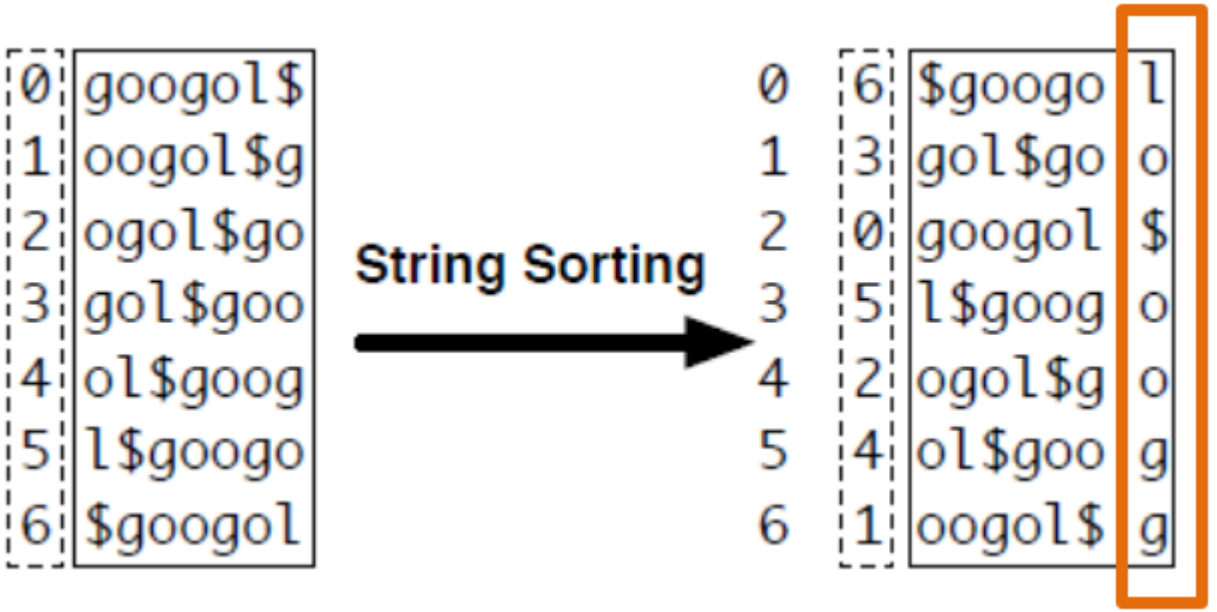
- Tollerante a polimorfismi che colpiscono al massimo due seed
- La creazione dell'indice del genoma di riferimento è un passaggio dispendioso in termini di tempo e risorse (circa 50GB di memoria per indicizzare il genoma umano)
- Velocità dipendente dalle risorse computazionali a disposizione

Trasformata di Burrows-Wheeler

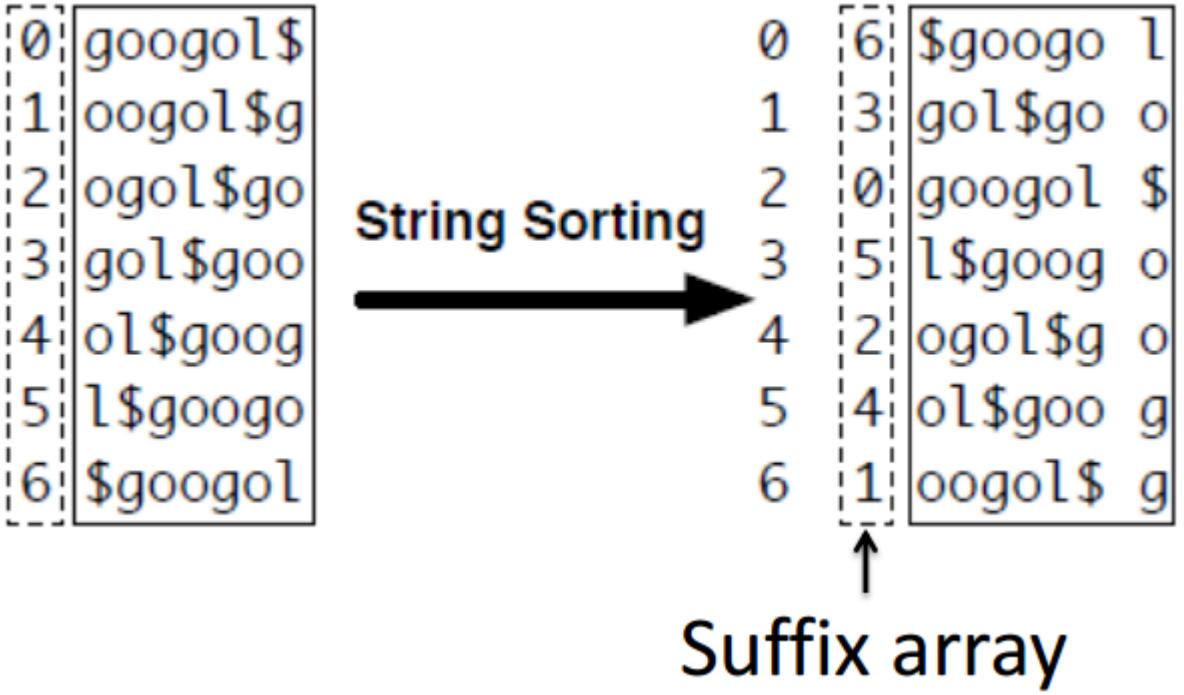
- Burrows M e Wheeler D (1994)
- Permutazione reversibile di caratteri testuali per permettere una compressione migliore (es: bzip2)
- Esistono algoritmi specifici per la ricerca rapida su dati trasformati tramite BW

Trasformata di Burrows-Wheeler

1. Testo originale = “googol”
2. Aggiungere ‘$’ alla fine = X = “googol$”
3. Mettere in ordine lessicografico tutte le rotazioni del testo X
4. Prendere l'ultima colonna



BW & suffix arrays



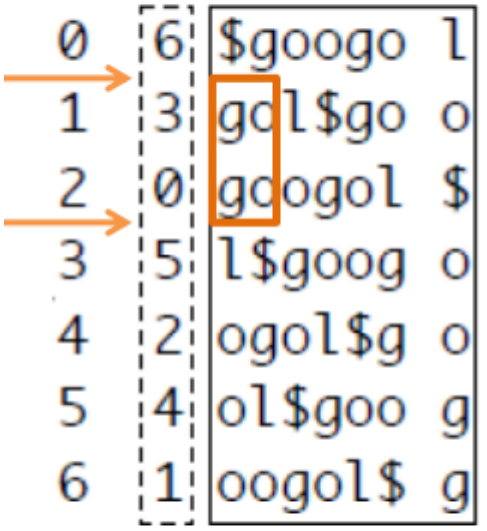
BW & suffix arrays interval

- Tutte le sottosequenze con un suffisso comune W appariranno vicine tra loro, definendo un intervallo nel suffix array

- Intervallo suffix array di

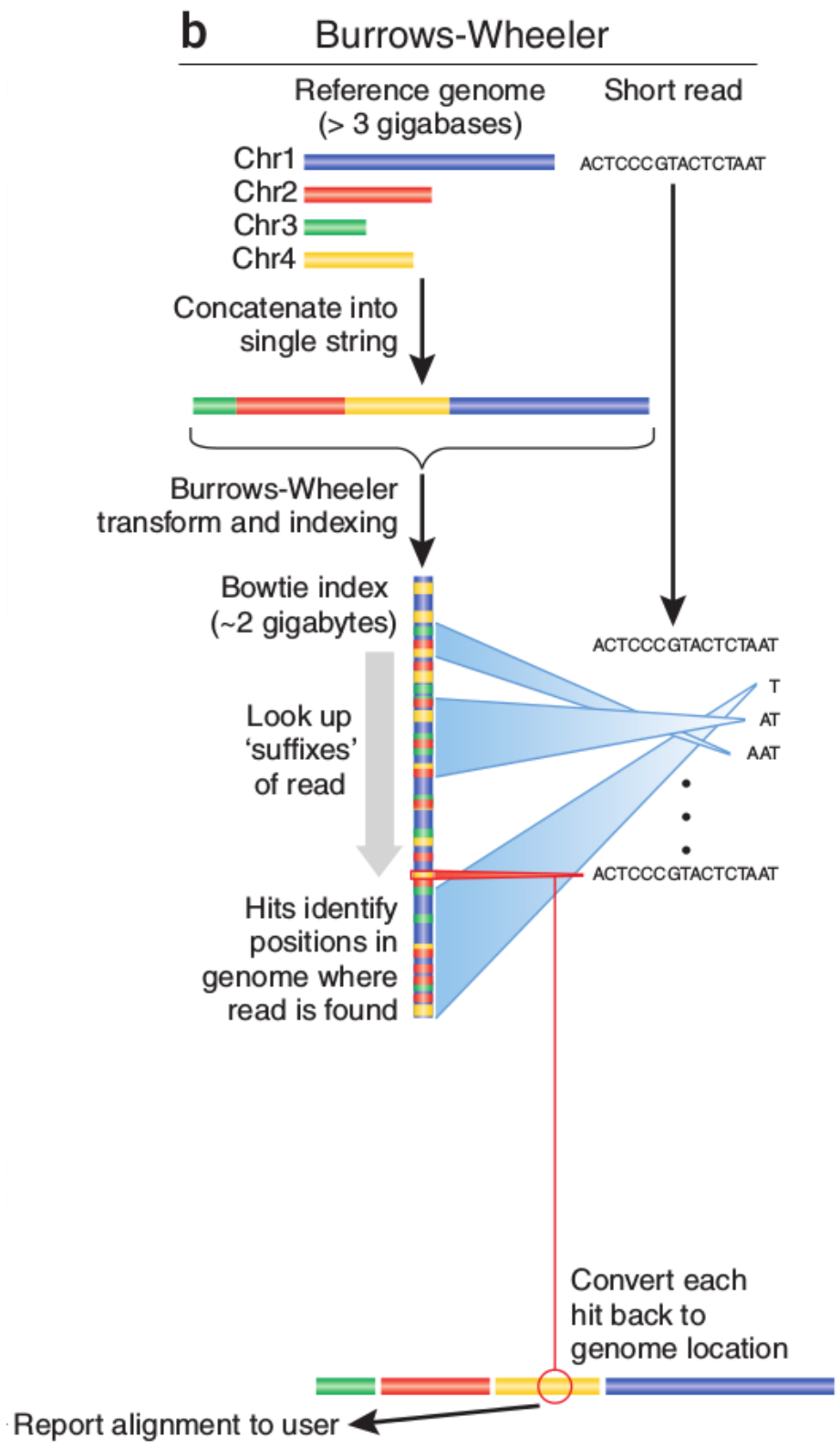
  “go” = [1,2]

W = “go”



X = googol$

0 3



Burrows-Wheeler

- Concatenare le sequenze di DNA di diversi cromosomi in un unica sequenza
- Applicare la trasformata di Burrows-Wheeler al genoma di riferimento (2-4GB)
- Per ogni read:
  - Calcolare il suffix array interval per un suffisso di lunghezza W (30pb) con al massimo x differenze
  - Estendere l'allineamento alla lunghezza della reads per identificare la regione specifica di miglior allineamento
- Riduzione del tempo di ricerca da ore a minuti (un milione di reads)

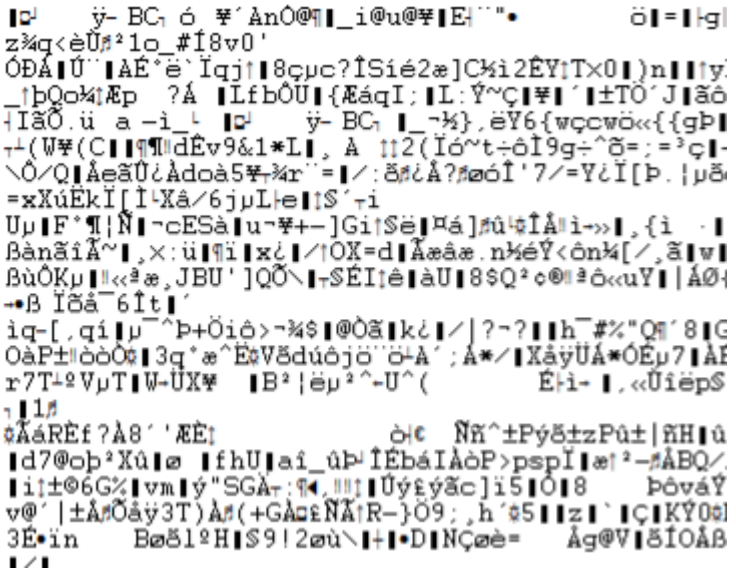
Formati per gli allineamenti

- Formato SAM/BAM
- SAM – sequence alignment map
- BAM – binary alignment map

Formato standard per conservare gli allineamenti (indipendente dal metodo utilizzato per allineare)

Il BAM è la versione binaria del SAM:

- dimensioni ridotte
- facilità di storage
- Accessibilità
- Non è un formato testuale → non è leggibile



Il formato SAM

Composto da due parti:

- Intestazione: contiene informazioni sui campioni
- Allineamenti: contiene informazioni sulla posizione e sulla qualità di allineamento di tutte le reads

Intestazione:

@HD – versione del formato

@SQ – Dizionario del genoma di riferimento (lista delle sequenze che compongono il genoma)

@RG – Codice di appartenenza delle reads

@PG – Programmi usati per generare e modificare l'allineamento

@CO – Commenti

- Nella parte contenente gli allineamenti, per ogni read abbiamo 11 colonne:

<table>
<tr><th>Col</th><th>Field</th><th>Type</th><th>Regexp/Range</th><th>Brief description</th></tr>
<tr><td>1</td><td>QNAME</td><td>String</td><td>[!-?A-~]{1,254}</td><td>Query template NAME</td></tr>
<tr><td>2</td><td>FLAG</td><td>Int</td><td>[0,2^{16}-1]</td><td>bitwise FLAG</td></tr>
<tr><td>3</td><td>RNAME</td><td>String</td><td>*|[!-()+-<>-~][!-~]*</td><td>Reference sequence NAME</td></tr>
<tr><td>4</td><td>POS</td><td>Int</td><td>[0,2^{31}-1]</td><td>1-based leftmost mapping POSition</td></tr>
<tr><td>5</td><td>MAPQ</td><td>Int</td><td>[0,2^{8}-1]</td><td>MAPping Quality</td></tr>
<tr><td>6</td><td>CIGAR</td><td>String</td><td>*|([0-9]+[MIDNSHPX=])+</td><td>CIGAR string</td></tr>
<tr><td>7</td><td>RNEXT</td><td>String</td><td>*|=|[!-()+-<>-~][!-~]*</td><td>Ref. name of the mate/next read</td></tr>
<tr><td>8</td><td>PNEXT</td><td>Int</td><td>[0,2^{31}-1]</td><td>Position of the mate/next read</td></tr>
<tr><td>9</td><td>TLEN</td><td>Int</td><td>[-2^{31}+1,2^{31}-1]</td><td>observed Template LENgth</td></tr>
<tr><td>10</td><td>SEQ</td><td>String</td><td>*|[A-Za-z=.]+</td><td>segment SEQuence</td></tr>
<tr><td>11</td><td>QUAL</td><td>String</td><td>[!-~]+</td><td>ASCII of Phred-scaled base QUALity+33</td></tr>
</table>

Esempio:
(1 read)

QNAME	FLAG	RNAME	POS	MAPQ
M00725:28:000000000-AJ72K:1:1101:18215:1102	99	cc_ref	1754677	60

CIGAR	MRNM	MPOS	ISIZE
300M	=	1754780	309

SEQ

CTCCTTCACCAGATGGATTCTCGCCTTACAGTCCTGAGGAAACTAACCGCAGAGTCAACAAAG
TAATGCGAGNNNNNNNGTACTTGCTACAGCNANNNGGTCCAAATNNNTNNNTTATTGGNNNAGATGTT

QUAL

CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGG#######::CFGGGGGGGGGG#:###:9BFGGGGG###:###::DFGDG###4+

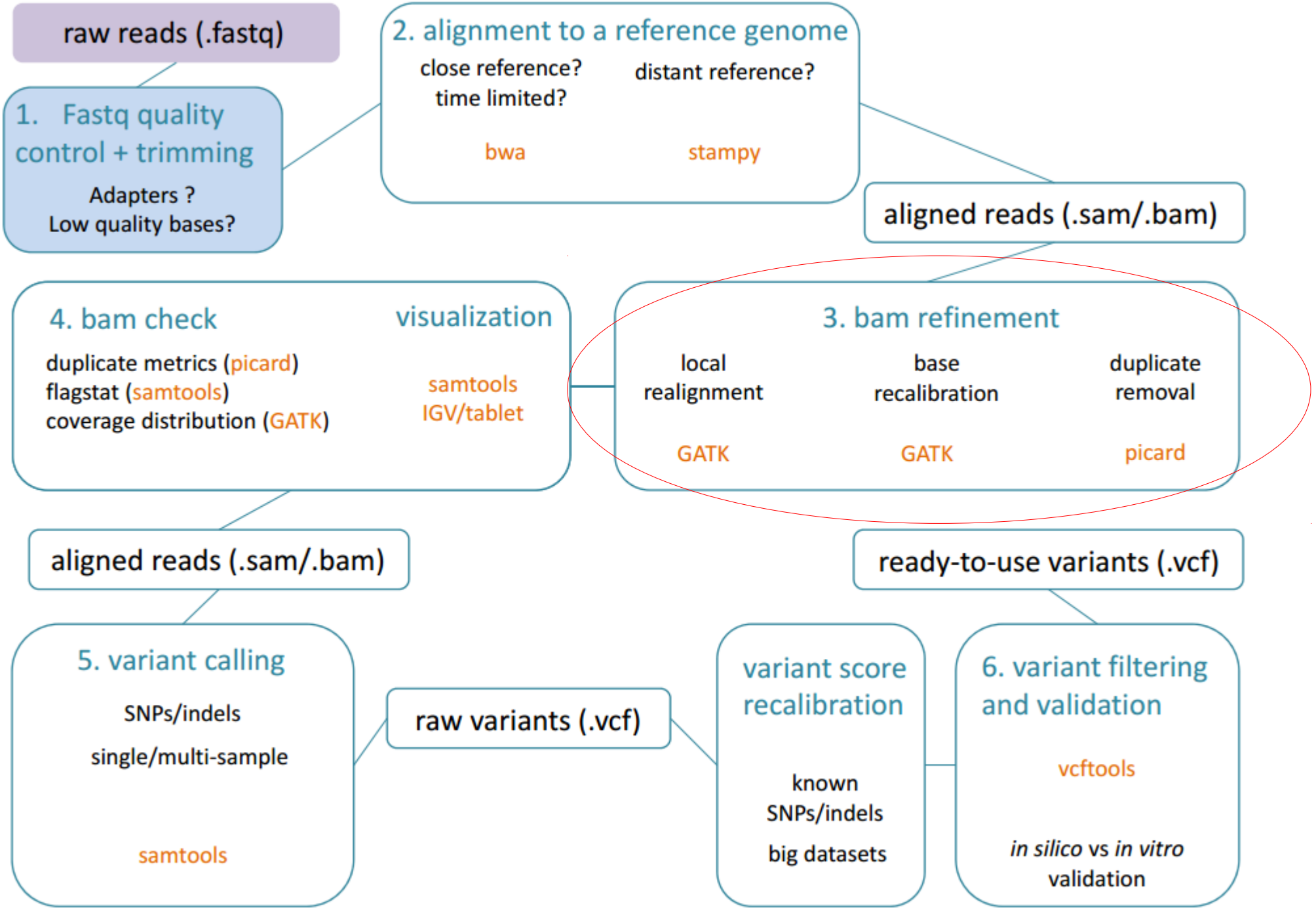
CIGAR string

- Rappresentazione compatta delle caratteristiche di allineamento. Include:
  - M, Match, corrispondenza tra basi nella reference e nella read
  - I, Insertion, inserzione di una o più basi nella read
  - D, Deletion, delezione di una o più basi nella read

  NB: i polimorfismi tra read e reference non sono indicati

```
read:   ACTCA–TGCAGT
ref:    ACTCAGTG—GT
cigar   5M1D2M2I2M
```

```
read:   ACGTCATG———CAGT
ref:    ACG–CATGCGGCAGT
cigar   3M1I4M3D4M
```


raw reads (.fastq)
1. Fastq quality control + trimming
Adapters ?
Low quality bases?
2. alignment to a reference genome
close reference?
time limited?
bwa
distant reference?
stampy
aligned reads (.sam/.bam)
3. bam refinement
local realignment
GATK
base recalibration
GATK
duplicate removal
picard
4. bam check
duplicate metrics (picard)
flagstat (samtools)
coverage distribution (GATK)
visualization
samtools
IGV/tablet
aligned reads (.sam/.bam)
5. variant calling
SNPs/indels
single/multi-sample
samtools
raw variants (.vcf)
variant score recalibration
known SNPs/indels
big datasets
6. variant filtering and validation
vcftools
in silico vs in vitro validation
ready-to-use variants (.vcf)

Rifinitura degli allineamenti

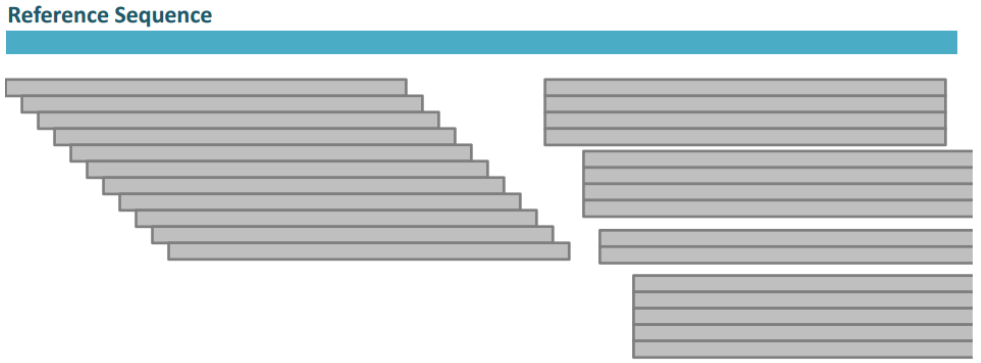
- Una volta ottenuto un allineamento in formato SAM conviene convertirlo nella sua versione binaria BAM
- Successivamente bisogna rifinire il BAM:
  - Rimozione dei duplicati di PCR (altamente consigliato)
  - Riallineamento in prossimità di INDELs (consigliato)
  - Ricalibrazione della qualità delle basi (opzionale)

Duplicati di PCR

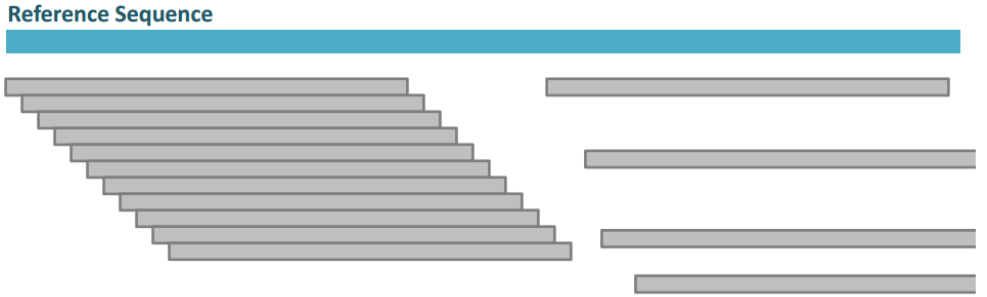
- Alcuni cicli di PCR sono previsti durante la preparazione delle library (non sempre vero).
- Introduzione di frammenti di DNA duplicati nelle library

- Il processo di rimozione dei duplicati di PCR cerca di:
  - Mantenere le informazioni contenute nei frammenti indipendenti
  - Eliminare le copie multiple di uno stesso frammento

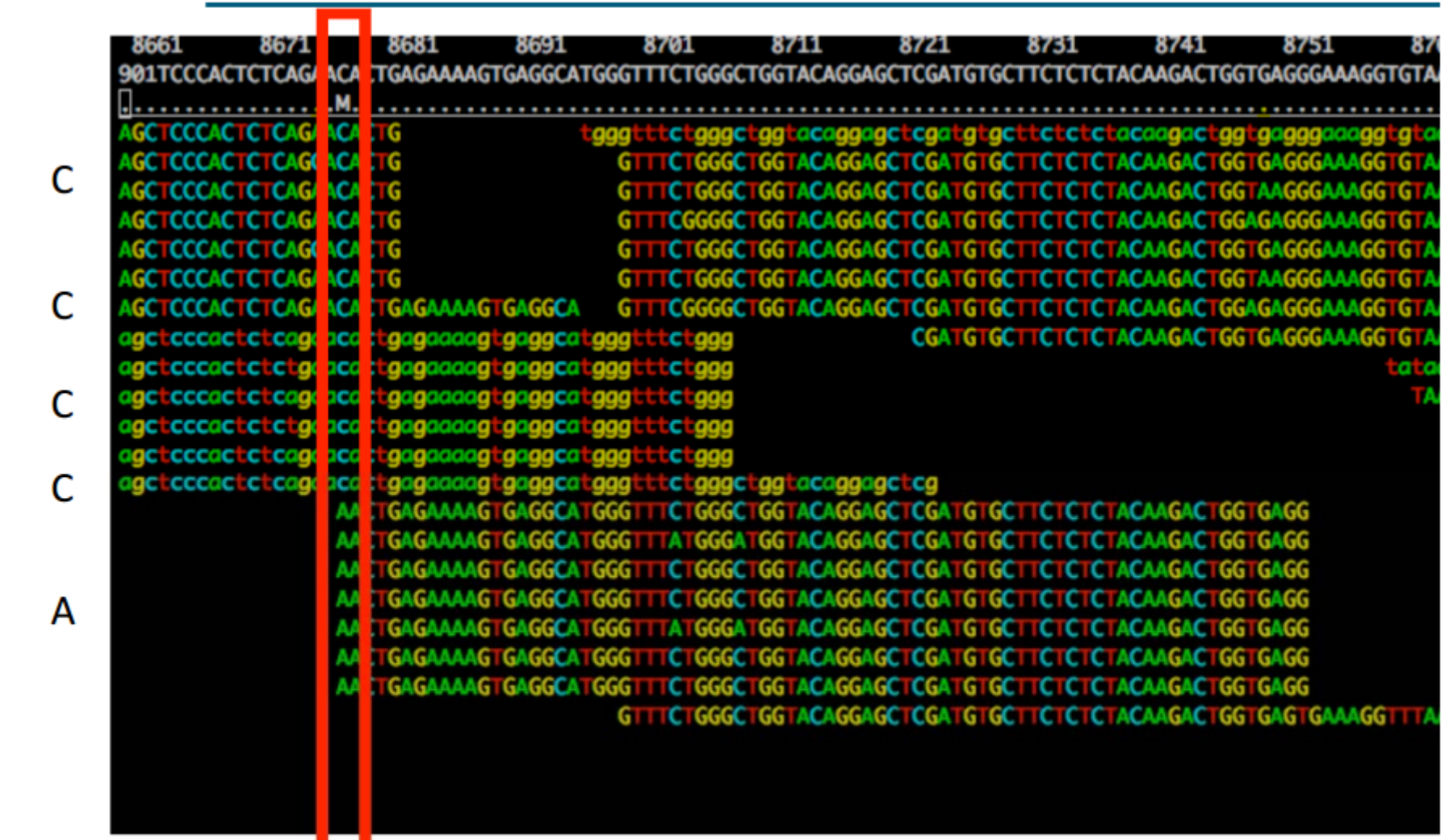
Prima della rimozione



Dopo la rimozione

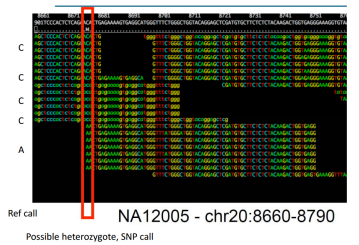


Duplicati di PCR



Possible heterozygote, SNP call

Duplicati di PCR

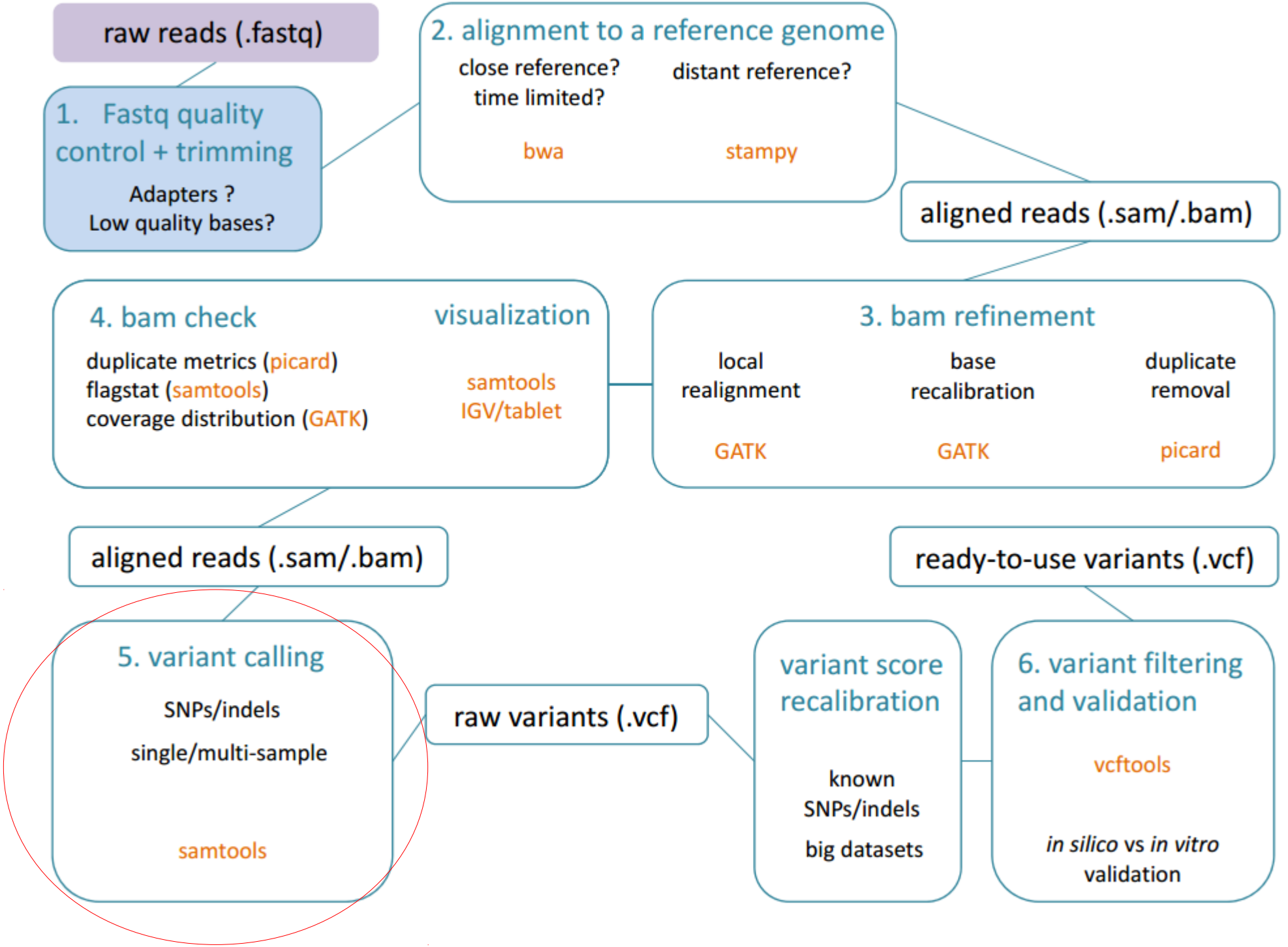


- Può provocare chiamate di SNPs false
- Influenza la profondità di sequenziamento effettiva
- Esistono protocolli PCR-free ma richiedono una notevole quantità di DNA di partenza.
- **Rimozione**: identificare se coppie di reads (pairs) mappano esattamente nella stessa posizione genomica e rimuoverle tutte eccetto una

Riallineamento di INDELs

- Gli algoritmi di allineamento hanno **difficoltà nell'allineare piccole INDELs** presenti nel campione (rispetto alla seqeuenza di riferimento), soprattutto se presenti verso la fine delle reads
- Se non trattate, questi errori di allineamento possono introdurre **falsi SNP**.
- Soluzione: identificare le regioni che contengono INDELS e **ri-allineare le reads** al genoma di riferimento solo in queste regioni

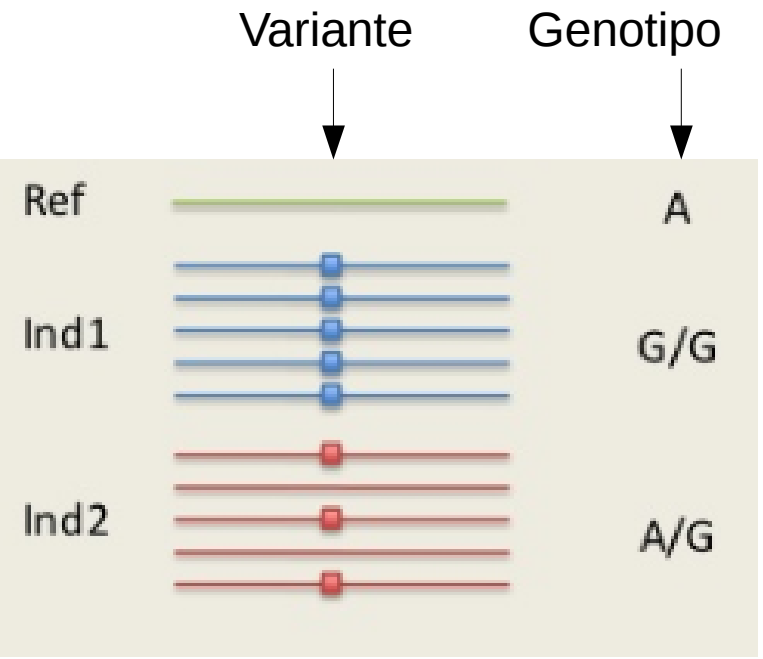
```
Ref: ACTTTCGGATGCTGATCGGGATGCTTTAGCTGATGCTGATGGGCTTTCGATCGATTTAAAAGCT
     ACTTTCGGATGCTGATC____ATGCTTTAGCTGA
          TCGGATGCTGATC____T_GCTTTAGCTGATGCT
                    CTGATC____ATGCTTTAGCTGATGCTGATGG
                         TC_____T_GCTTTAGCTGATGCTGATGGGCTT

Ref: ACTTTCGGATGCTGATCGGGATGCTTTAGCTGATGCTGATGGGCTTTCGATCGATTTAAAAGCT
     ACTTTCGGATGCTGATC____ATGCTTTAGCTGA
          TCGGATGCTGATC____ATGCTTTAGCTGATGCT
                    CTGATC____ATGCTTTAGCTGATGCTGATGG
                         TC____ATGCTTTAGCTGATGCTGATGGGCTT
```


raw reads (.fastq)
1. Fastq quality control + trimming
Adapters ?
Low quality bases?
2. alignment to a reference genome
close reference?
time limited?
bwa
distant reference?
stampy
aligned reads (.sam/.bam)
3. bam refinement
local realignment
GATK
base recalibration
GATK
duplicate removal
picard
4. bam check
duplicate metrics (picard)
flagstat (samtools)
coverage distribution (GATK)
visualization
samtools
IGV/tablet
aligned reads (.sam/.bam)
5. variant calling
SNPs/indels
single/multi-sample
samtools
raw variants (.vcf)
variant score recalibration
known SNPs/indels
big datasets
6. variant filtering and validation
vcftools
in silico vs in vitro validation
ready-to-use variants (.vcf)

Variant & genotype calling

- Chiamata delle varianti: Processo nel quale si identificano le differenze tra il genoma di riferimento e il genoma del campione sulla base delle reads allineate
  - SNP bi- e multi-allelici
  - Polimorfismi complessi (es: tre SNP consecutivi)
  - Inserzioni e delezioni (INDELs, <50bp circa)

- Chiamata dei genotipi: Processo nel quale, per ognuno dei polimorfismi identificati si cerca di attribuire il genotipo ad ogni individuo sulla base della ploidia (es: RefRef/RefAlt/AltAlt, n=2)

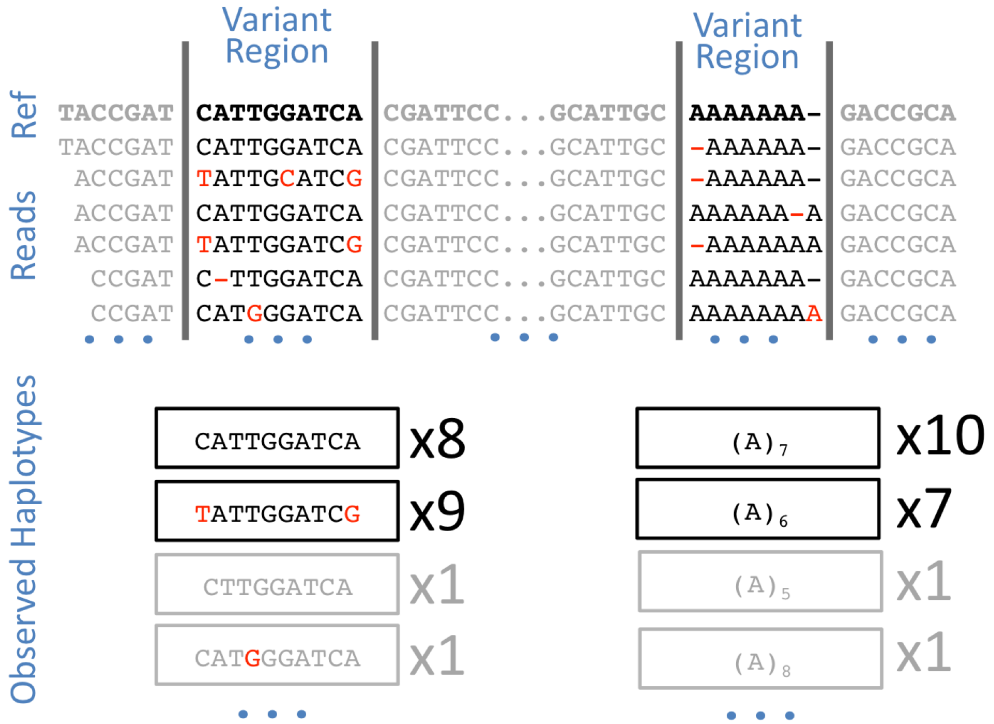


Metodi per effettuare le chiamate (SNP & Geno.)

- Frequenza degli alleli: conteggio per ogni posizione genomica del numero di alleli Ref e Alt (PhredQbase > 20). SNP con genotipo eterozigote se Alt {20-80%}
  - Soglie arbitrarie
  - Problemi con basso coverage
- Metodi probabilistici (es: modello bayesiano): per ogni posizione genomica è possibile calcolare la probabilità di osservare ogni genotipo possibile (Ref/Ref;Ref/Alt;Alt/Alt) dati gli alleli presenti nelle reads.
  - Chiamata di SNP e Geno. in un unico passaggio
  - Avendo una probabilità dei genotipi si tiene in considerazione l'incertezza nelle chiamate
  - Problema nella definizione di una prior sulla frequenza degli alleli (uniforme, database di variabilità (ES: dbSNP, uomo), analisi di più individui)

Metodi per effettuare le chiamate (SNP & Geno.)

- Metodi basati sugli aplotipi
  - Non trattano ogni posizione genomica in maniera indipendente
  - Dividono il genoma in moltissime piccole regioni (5-10pb)
  - In ognuna vengono identificati gli aplotipi più frequenti e vengono chiamati gli SNP e Geno. sulla base delle combinazioni alleliche osservate

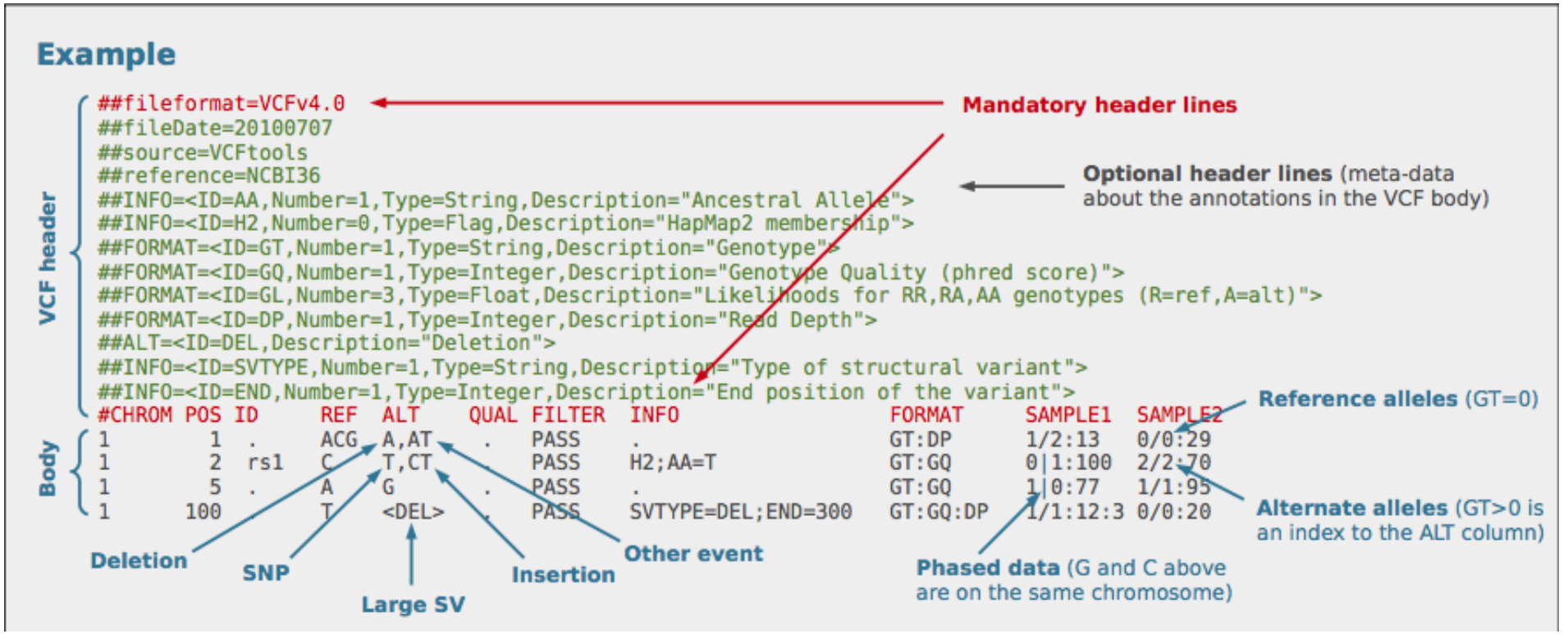


Controllo delle chiamate

- E' buona norma controllare alcune caratteristiche degli SNP chiamati:
  - Non ci siano troppe basi con bassa qualità (PhredQbase <20)
  - Le reads mappino in proporzioni simili su entrambe le eliche (+ e -)
  - Prossimità ad INDELS, omopolimeri o regioni ripetute
  - Densità degli SNP (troppi SNP vicini sono sintomo di errori)
  - Non ci siano troppe reads con bassa qualità di allineamento (PhreadQmap < 20)
  - Il numero di reads che coprono il polimorfismo sia adeguato
  - L'allele alternativo abbia una frequenza minima tra gli individui analizzati (es: eterozigote in un singolo individuo su N analizzati)

Variant calling format (VCF)

- Presenza di una “intestazione” contenente la descrizione delle informazioni registrare per ogni variante
- Presenza di un “corpo” contenete le informazioni riguardo le varianti



Intestazione del vcf

CHROM → cromosoma

POS → posizione iniziale della variante

ID → identificativo univoco della variante (se esiste)

REF → allele referenza

ALT → lista degli alleli alternativi (uno o più)

QUAL → PhreadQ associata alla variante

FILTER → informazioni sul filtro

INFO → informazioni relative alla variante (dipende dal metodo di chiamata)

FORMAT → informazioni registrate per il genotipo di ogni individuo

```
##bcftools_callCommand=call -m -
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT
```

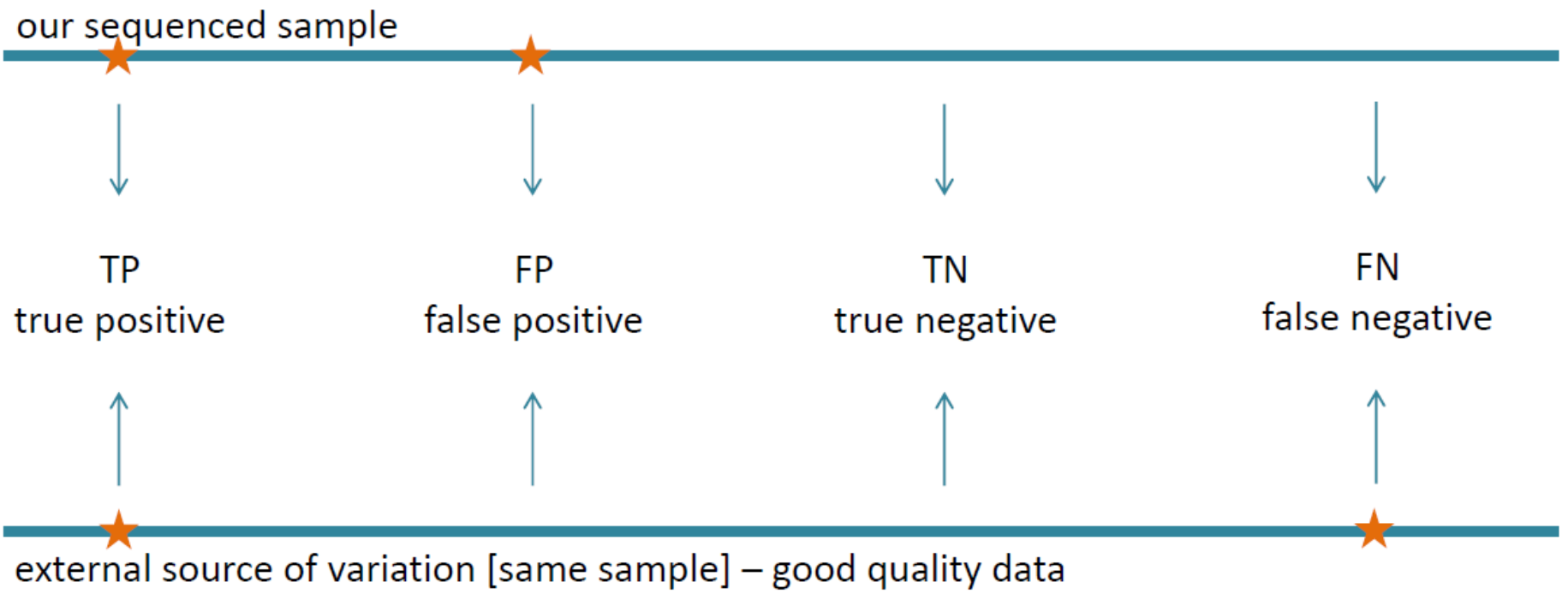
Corpo del vcf

```
##fileformat=VCFv4.0
##fileDate=20110705
##reference=1000GenomesPilot-NCBI37
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS    ID        REF  ALT     QUAL FILTER INFO                              FORMAT      Sample1        Sample2        Sample3
2      4370   rs6057    G    A       29   .      NS=2;DP=13;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:52,51 1|0:48:8:51,51 1/1:43:5:.,.
2      7330   .         T    A       3    q10    NS=5;DP=12;AF=0.017               GT:GQ:DP:HQ 0|0:46:3:58,50 0|1:3:5:65,3   0/0:41:3
2      110696 rs6055    A    G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
2      130237 .         T    .       47   .      NS=2;DP=16;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:56,51 0/0:61:2
```

Validazione degli SNP

variant calling – evaluating

Specificity vs Sensitivity = False Positive vs False Negative



high specificity → low FP

high sensitivity → low FN

Validazione degli SNP

- Usare **database** di variabilità: sequenziare campioni di cui si conoscono già i polimorfismi ci aiuta a capire quanto le procedure di identificazione sono accurate

- Validazione **sperimentale**: selezionare un gruppo di varianti appena scoperte e risequenziarle con un'altra piattaforma (Sanger) . Il tasso di false scoperte (Falsi positivi e Falsi negativi) ci informa riguardo la bontà della procedura di chiamata.