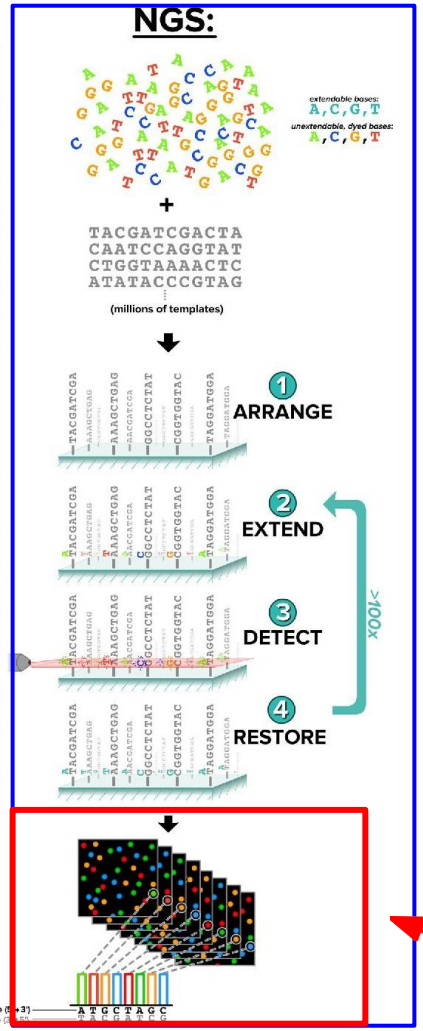
Sequenziamento massivo degli acidi nucleici

Next Generation Sequencing Data

- ❏ NGS data format (FASTQ)
- ❏ Paired end versus Single end reads
- ❏ Quality control of reads



Tipi di dati

Una analisi di dati biologici richiede di gestire e manipolare dati che possono essere classificati come:

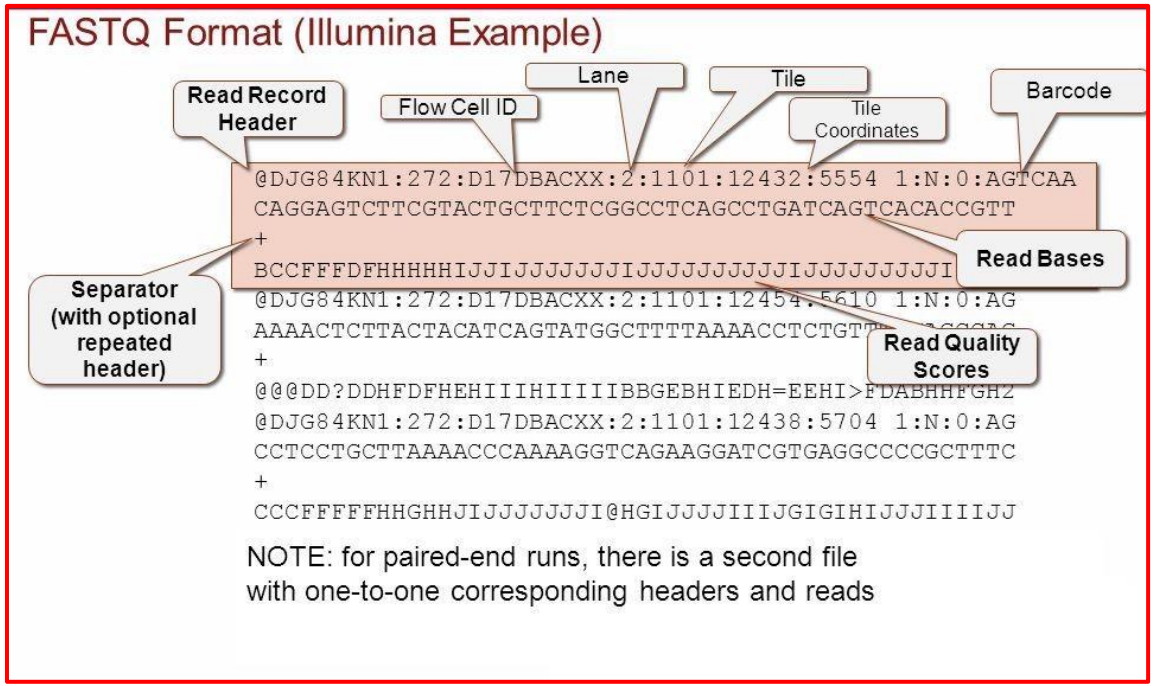
1. Dati grezzi (sequenze nucleotidiche o sequenze di amminoacidi **FASTA,** regioni genomiche come coordinate e annotazioni associate **BED,** geni e altre caratteristiche delle sequenze di DNA, RNA e proteine **GFF**)
2. Dati ottenuti sperimentalmente (noti anche come letture di sequenziamento ovvero *read*: **FASTQ**)
3. Dati derivati dalla analisi (**BAM**, **VCF**, formati dal punto 1 sopra e molti formati non standard)

Ogni formato lo rende adatto a obiettivi specifici

Si può riconoscere il formato dalla sua estensione

Qual è il nome della procedura software per estrarre le sequenze di lettura dalle immagini?

FORMATO FASTQ



Il formato FASTQ* è lo standard de facto usato per essere elaborato dagli strumenti di sequenziamento. Associa ad ogni base della sequenza una misura di qualità, FASTA with QUALITIES.

E' un formato in cui ad ogni base si associa una misura di affidabilità: la base è A e la probabilità che ci sbagliamo è 1/1000.

*** https://en.wikipedia.org/wiki/FASTQ_format**

FORMATO FASTQ: READS

Name	@ERR194146.1 HSQ1008:141:D0CC8ACXX:3:1308:20201:36071/1
Sequence	ACATCTGGTTCCTACTTCAGGGCCATAAAGCCTAAATAGCCCACACGTTCCCCTTAAAT
(ignore)	+
Base qualities	?@@FFBFFDDHHBCEAFGEGIIDHGH@GDHHHGEHID@C?GGDG@FHIGGH@FHBEG:G

file FASTQ

FORMATO FASTQ: caratteristiche READS

Il formato FASTQ è composto da 4 sezioni (e di solito vengono prodotte una sola riga ciascuna):

1. Un'intestazione che inizia con il simbolo @, un ID e un altro testo facoltativo.
2. La seconda sezione contiene la sequenza misurata (tipicamente su una singola riga), ma può essere spostata a capo .
3. La terza sezione è contrassegnata dal segno iniziale + e può essere facoltativamente seguita dallo stesso ID di sequenza e intestazione della prima sezione
4. L'ultima riga codifica i valori di qualità per la sequenza nella seconda sezione (le due righe hanno lunghezza uguale) secondo lo standard **Phred**.

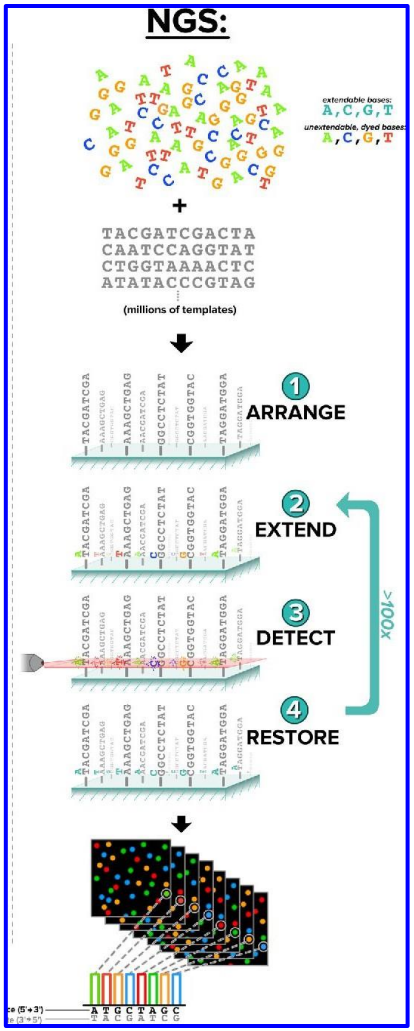
FORMATO FASTQ: esempio

@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!'"*((((*+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65**

I caratteri "strani" nella quarta riga sono i cosiddetti valori numerici “codificati”.

In poche parole, ogni carattere !'"*((((rappresenta un valore numerico la cui visualizzazione avviene in codice ASCII

https://en.wikipedia.org/wiki/Phred_quality_score



FORMATO FASTQ

Il formato FASTQ ha lo scopo di memorizzare misurazioni di sequenze sperimentali prodotte da uno strumento di sequenziamento.

```
fastq-dump -X 100 -Z SRR1553607 | head
```

```
@SRR1553607.1 1 length=202
GTTAGCGTTGTTGATCGCGACGCAACAACTGGTAAAGAATCTGGAAGAAGGATATCAGTTCAAACGCTCAAG
+SRR1553607.1 1 length=202  BB
FFFFFHHHHHJJJJJJJJJJJJJJJJJJGHIJJJJJJJJJJHHHHHFFFFFEEEEEEEEEDDDDDDDDD
@SRR1553607.2 2 length=202
GGTGTAAGCACAGTACTCGGCCCACATCGCCTTTGTGTTAATGAAGTTTGGGTATCAACTTTCATCCCCAAT
+SRR1553607.2 2 length=202
BDDDDFHFHFFFGIIGHIIJJGJIGIJIIIIGDGGGHEIGJIIIGIIHJ5@FGHJJIEGGEEHHFFFFF
```

FORMATO FASTQ: PHRED

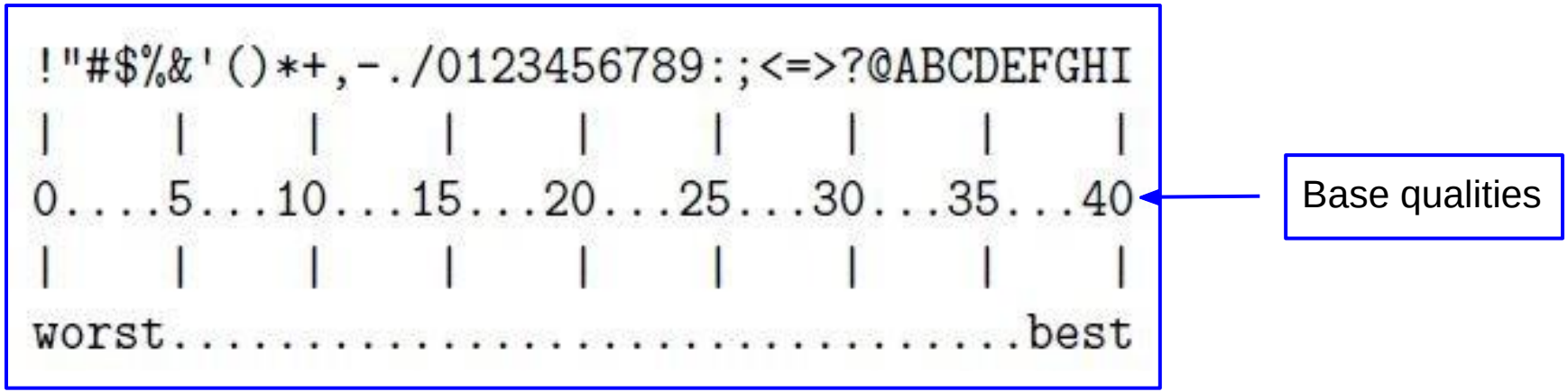
La metrica Phred mappa i numeri a due cifre che rappresentano la probabilità di errore su singoli caratteri in modo che la lunghezza della stringa di qualità è la stessa della lunghezza della sequenza

$$Q = -10 \log_{10} P \qquad P = 10^{\frac{-Q}{10}}$$

Ad esempio, se Phred assegna un punteggio di qualità pari a 30 a una base, le probabilità che questa base venga chiamata in modo errato sono 1 su 1000.

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 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

FORMATO FASTQ: PHRED



FORMATO FASTQ: controllo qualità

Quando i caratteri che rappresentano la qualità sembrano "imprecazioni" in un fumetto!"#$%&'()*+,-. significa che i dati sono di bassa qualità.

Immagina lo strumento di sequenziamento che ti dice: Ehi, guarda tutto questo!$!!$*#@! di dati

Quando i punteggi di qualità sono leggibili, magari con lettere confuse, ABAFEFGGII i dati hanno una discreta qualità.

Immagina lo strumento di sequenziamento che ti dice: Ehi, guarda tutti questi dati IAIAIAIAIA!

FORMATO BED

I formati GFF/GTF/BED hanno le posizioni di una regione appartenente ad un genoma (sono detti formati intervallo). Ogni campo è delimitato da tabulazione e contiene informazioni su coordinate cromosomiche, inizio, fine, filamento, valore e altri attributi.

Il BED a tre colonne ha il posizionamento <CROMOSOMA INIZIO FINE>

```
chr7    127471196  127472363
chr7    127472363  127473530
chr7    127473530  127474697
```

FORMATO BED

Colonna	Campo	Definizione	Obbligatorietà
1	**chrom**	Nome del cromosoma (chr3, chrY, chr2_random) o dello scaffold (scaffold10671)	Si
2	**chromStart**	Inizio sequenza (si conta da 0)	Si
3	**chromEnd**	Fine sequenza	Si
4	**name**	Nome della linea	No
5	**score**	Valore da 0 a 1000	No
6	**strand**	Orientamento strand DNA (positivo ["+"] o negativo ["-"] o "." se non c'è lo strand)	No
7	**thickStart**	Inizio coordinate dalle quali sono riportate le annotazioni (inizio di un codone di un gene)	No
8	**thickEnd**	Fine coordinate dalle quali sono riportate le annotazioni (stop codone di un gene)	No
9	**itemRgb**	Valore RGB (blue: <255,0,0,0>) per la visualizzazione cromatica del file BED	No
10	**blockCount**	Numero di blocchi (es: esoni)	No
11	**blockSizes**	Lista della dimensione dei blocchi	No
12	**blockStarts**	Lista della posizione inziale dei blocchi (relazione con il campo blockCount)	No

FORMATO GFF

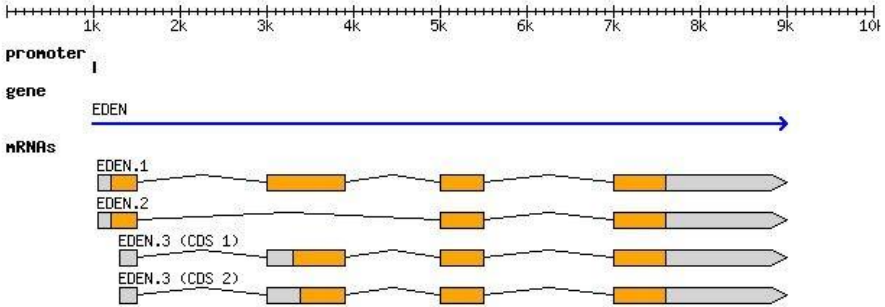
Il formato delle caratteristiche generali (gene-finding format, generic feature format, GFF) è organizzato per descrivere geni e altre caratteristiche di sequenze di DNA, RNA e proteine.

```
 0  ##gff-version 3.2.1
 1  ##sequence-region ctg123 1 1497228
 2  ctg123 . gene            1000  9000  .  +  .  ID=gene00001;Name=EDEN
 3  ctg123 . TF_binding_site 1000  1012  .  +  .  ID=tfbs00001;Parent=gene00001
 4  ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00001;Parent=gene00001;Name=EDEN.1
 5  ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00002;Parent=gene00001;Name=EDEN.2
 6  ctg123 . mRNA            1300  9000  .  +  .  ID=mRNA00003;Parent=gene00001;Name=EDEN.3
 7  ctg123 . exon            1300  1500  .  +  .  ID=exon00001;Parent=mRNA00003
 8  ctg123 . exon            1050  1500  .  +  .  ID=exon00002;Parent=mRNA00001,mRNA00002
 9  ctg123 . exon            3000  3902  .  +  .  ID=exon00003;Parent=mRNA00001,mRNA00003
10  ctg123 . exon            5000  5500  .  +  .  ID=exon00004;Parent=mRNA00001,mRNA00002,mRNA00003
11  ctg123 . exon            7000  9000  .  +  .  ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003
12  ctg123 . CDS             1201  1500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
13  ctg123 . CDS             3000  3902  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
14  ctg123 . CDS             5000  5500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
15  ctg123 . CDS             7000  7600  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
16  ctg123 . CDS             1201  1500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
17  ctg123 . CDS             5000  5500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
18  ctg123 . CDS             7000  7600  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
19  ctg123 . CDS             3301  3902  .  +  0  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
20  ctg123 . CDS             5000  5500  .  +  1  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
21  ctg123 . CDS             7000  7600  .  +  1  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
22  ctg123 . CDS             3391  3902  .  +  0  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
23  ctg123 . CDS             5000  5500  .  +  1  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
24  ctg123 . CDS             7000  7600  .  +  1  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
```

FORMATO GFF

Indice	Nome	Descrizione
1	seqid	Il nome delle sequenze
2	source	L'algoritmo che ha generato le caatteristiche (nome del software o del databas)
3	type	Il tipo (gene o esone). In GFF3, il tipo e le relazioni devon o rispettare lo standards released by the Sequence Ontology Project.
4	start	Inzio della caratteristica genomica (inizio a partire dal valore 1)
5	end	Fine della caratteristica genomica
6	score	Valore numeric che indica l'attendibilità della annotazione (il punto inidica una attendibilità nulla)
7	strand	Carattere che indica lo strand della caratteristica "+" (positivo, o 5'->3'), "-", (negativo, or 3'->5'), "." (non determinato), o "?" pe caratteristiche rilevanti ma il cui strand non è noto.
8	phase	Fase della funzionalità CDS
9	attributes	Coppie di marcatori per informazioni aggiuntive alle annotazioni (es: ID=…. ;NAME=….)

FORMATO GFF




```
0  ##gff-version 3.2.1
1  ##sequence-region ctg123 1 1497228
2  ctg123 . gene            1000  9000  .  +  .  ID=gene00001;Name=EDEN
3  ctg123 . TF_binding_site 1000  1012  .  +  .  ID=tfbs00001;Parent=gene00001
4  ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00001;Parent=gene00001;Name=EDEN.1
5  ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00002;Parent=gene00001;Name=EDEN.2
6  ctg123 . mRNA            1300  9000  .  +  .  ID=mRNA00003;Parent=gene00001;Name=EDEN.3
7  ctg123 . exon            1300  1500  .  +  .  ID=exon00001;Parent=mRNA00003
8  ctg123 . exon            1050  1500  .  +  .  ID=exon00002;Parent=mRNA00001,mRNA00002
9  ctg123 . exon            3000  3902  .  +  .  ID=exon00003;Parent=mRNA00001,mRNA00003
10 ctg123 . exon            5000  5500  .  +  .  ID=exon00004;Parent=mRNA00001,mRNA00002,mRNA00003
11 ctg123 . exon            7000  9000  .  +  .  ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003
12 ctg123 . CDS             1201  1500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
13 ctg123 . CDS             3000  3902  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
14 ctg123 . CDS             5000  5500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
15 ctg123 . CDS             7000  7600  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
16 ctg123 . CDS             1201  1500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
17 ctg123 . CDS             5000  5500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
18 ctg123 . CDS             7000  7600  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
19 ctg123 . CDS             3301  3902  .  +  0  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
20 ctg123 . CDS             5000  5500  .  +  1  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
21 ctg123 . CDS             7000  7600  .  +  1  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
22 ctg123 . CDS             3391  3902  .  +  0  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
23 ctg123 . CDS             5000  5500  .  +  1  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
24 ctg123 . CDS             7000  7600  .  +  1  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
```

FORMATO SAM

Il file SAM contiene informazioni sugli allineamenti.

Il formato SAM è un formato di testo delimitato da tabulati costituito da:

- Sezione di intestazione, in cui ogni riga contiene alcuni metadati
- Sezione di allineamento, in cui ciascuna riga fornisce informazioni su un allineamento

In generale, la qualità delle informazioni all'interno di un file SAM determina il successo dell'analisi. Pertanto, è importante produrre questo file in modo che contenga le informazioni di cui si ha bisogno per indagare sui dati.

FORMATO SAM

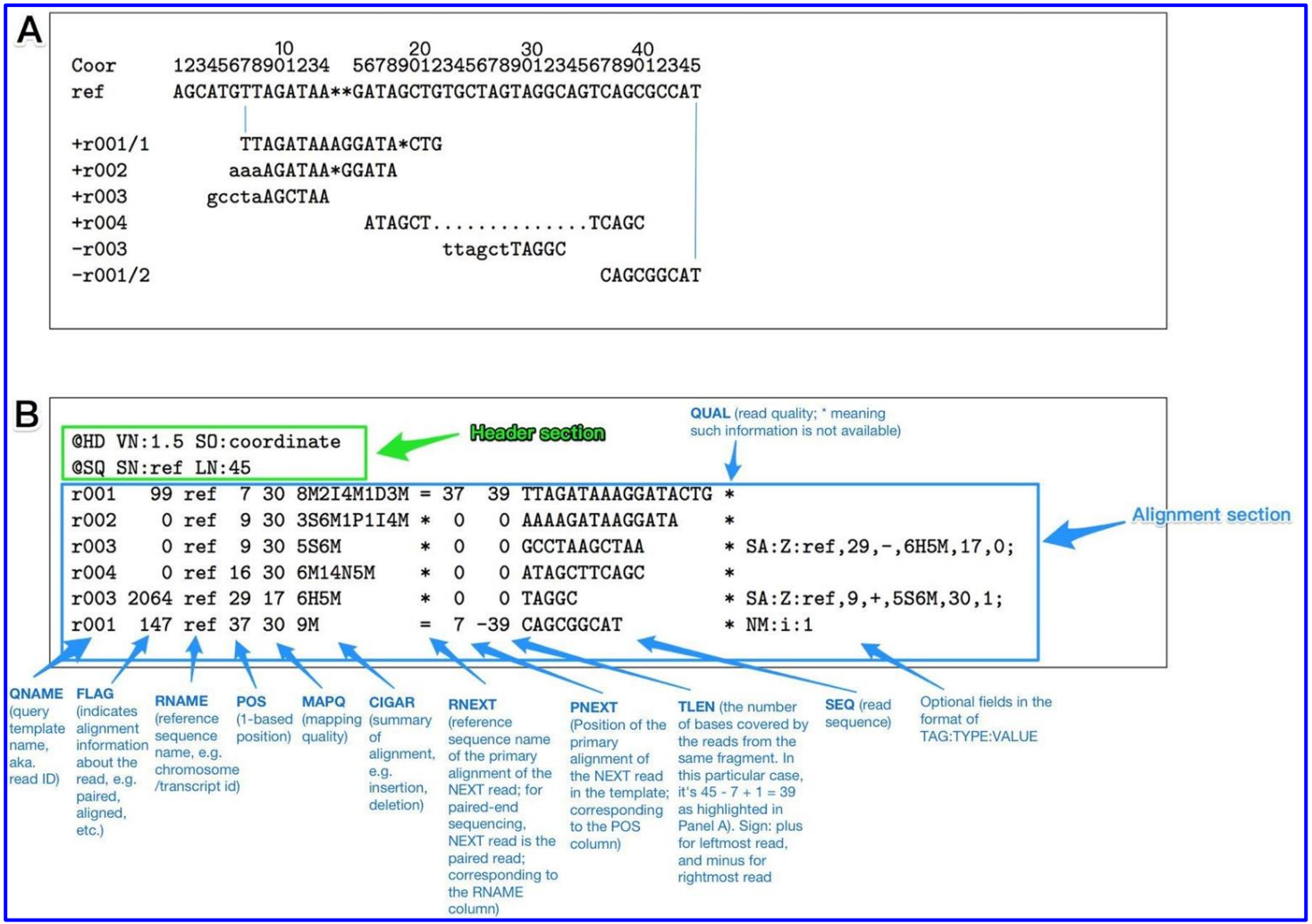
Il formato SAM (come il BAM) è detto Sequence Alignment Maps Format.

Rappresenta i risultati dell'allineamento di un file FASTQ a un file FASTA di riferimento e descrivono i singoli allineamenti a coppie trovati. Algoritmi diversi possono creare allineamenti diversi (e quindi file SAM).

```
[tcastign@r033c01s03 ~]$ samtools view http://data.biostarhandbook.com/bam/demo.bam | head -5
```

```
SRR1553425.13617   163 AF086833.2 46 60 101M   =  541 596 GAATAACTATGAGGAAGATTAATAATTTTC
SRR1553425.13755   99  AF086833.2 46 60 101M   =  46  101 GAATAACTATGAGGAAGATTAATAATTTTCC
SRR1553425.13755   147 AF086833.2 46 60 101M   =  46  -101   GAATAACTATGAGGAAGATTAATAATTT
SRR1553425.11219   2227   AF086833.2 47 60 71H30M =  146 171 AATAACTATGAGGAAGATTAATAATTT
```

FORMATO SAM


A
10 20 30 40
Coor 12345678901234 5678901234567890123456789012345
ref AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
+r001/1 TTAGATAAAGGATA*CTG
+r002 aaaAGATAA*GGATA
+r003 gcctaAGCTAA
+r004 ATAGCT..............TCAGC
-r003 ttagctTAGGC
-r001/2 CAGCGGCAT
B
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
Header section
QUAL (read quality; * meaning such information is not available)
r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;
r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1
Alignment section
QNAME (query template name, aka. read ID)
FLAG (indicates alignment information about the read, e.g. paired, aligned, etc.)
RNAME (reference sequence name, e.g. chromosome /transcript id)
POS (1-based position)
MAPQ (mapping quality)
CIGAR (summary of alignment, e.g. insertion, deletion)
RNEXT (reference sequence name of the primary alignment of the NEXT read; for paired-end sequencing, NEXT read is the paired read; corresponding to the RNAME column)
PNEXT (Position of the primary alignment of the NEXT read in the template; corresponding to the POS column)
TLEN (the number of bases covered by the reads from the same fragment. In this particular case, it's 45 - 7 + 1 = 39 as highlighted in Panel A). Sign: plus for leftmost read, and minus for rightmost read
SEQ (read sequence)
Optional fields in the format of TAG:TYPE:VALUE

FORMATO SAM

Colonna	Campo	Tipo di dato	Descrizione
1	QNAME	String	Nome
2	FLAG	Int	Flag bit per bit
3	RNAME	String	Nome della sequenza
4	POS	Int	1- Posizione di mappatura più a sinistra
5	MAPQ	Int	MAPping Quality
6	CIGAR	String	CIGAR string
7	RNEXT	String	Nome della read precedente/successive
8	PNEXT	Int	Posizione della read precedente/successive
9	TLEN	Int	Lunghezza
10	SEQ	String	Sequenza
11	QUAL	String	Phred

FORMATO SAM

I file BAM prodotti da strumenti diversi NON contengono una quantità comparabile di informazioni e le differenze NON risiedono principalmente nella precisione o nelle caratteristiche prestazionali degli allineamenti.

L'equivoco è causato dalla stessa specifica SAM che prevede 11 colonne e ciascuna colonna contiene le informazioni descritte in precedenza

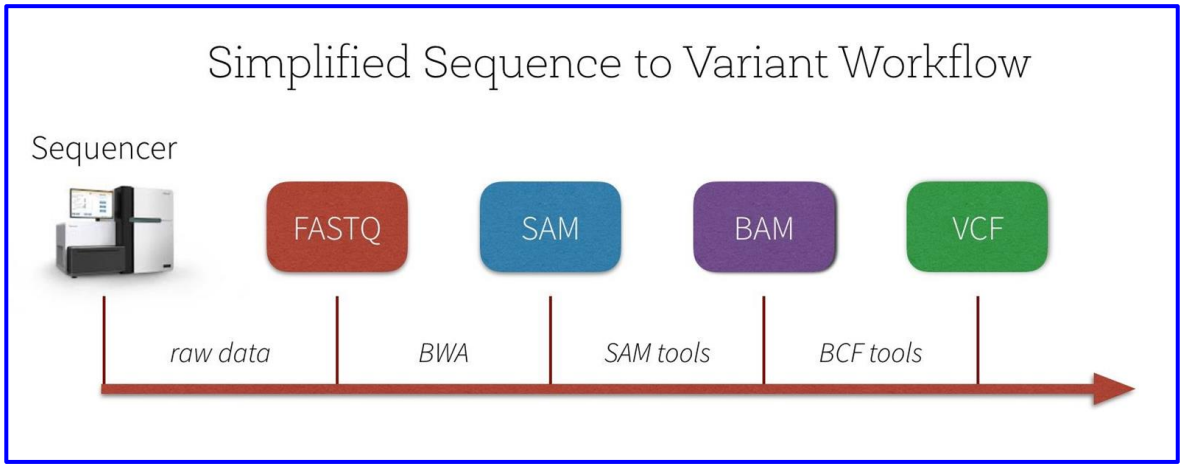
I tag “opzionali” e altri campi del file SAM contengono tutte le informazioni necessarie per l'analisi. Pertanto potrebbero esserci differenze sostanziali tra le informazioni di allineamento prodotte da strumenti diversi

Il messaggio da portare a casa è che il formato SAM dipende dall'allineatore

FORMATO SAM

Un file BAM è una rappresentazione binaria, compressa (e quasi sempre ordinata) delle informazioni SAM.

Generalmente, i file BAM sono ordinati in base alla coordinata di allineamento e più raramente in relazione ai nomi di lettura nel caso di coppie di lettura con nome identico



FORMATO SAM: uso

I file SAM sono stati progettati per due casi d'uso principali:

1. archiviare gli allineamenti in modo standardizzato ed efficiente
2. consentire un rapido accesso agli allineamenti tramite le loro coordinate

Ad esempio, se in un file sono presenti 100 milioni di allineamenti e si desidera che gli allineamenti si sovrappongano alla coordinata 1.200.506, il formato BAM può restituire tali informazioni in breve tempo (millisecondi), senza dover leggere l'intero file

Il formato SAM NON è opzionale malgrado la specifica iniziale sia stato rovinata da ottimizzazioni premature, bloccata da una progettazione concettuale imperfetta, compromessa da specifiche incomplete.

FORMATO SAM

Il formato SAM è stato introdotto per supportare i casi d'uso presentati dalla strumentazione di sequenziamento ad alto rendimento:

1. **Accesso rapido agli allineamenti che si sovrappongono a una coordinata.** Ad esempio, seleziona allineamenti che si sovrappongono alla coordinata 323.567.334 sul cromosoma 2
2. **Selezione rapida e filtraggio delle letture in relazione agli attributi.** Ad esempio, se si vuole essere in grado di selezionare rapidamente gli allineamenti che si allineano sul filo inverso
3. **Archiviazione e distribuzione efficienti dei dati.** Ad esempio, avere un singolo file compresso contenente i dati per tutti i campioni, ciascuno etichettato in qualche modo.

FORMATO SAM e BAM: creazione

Di solito si genera un file SAM, quindi si ordina il file e lo si converte in un formato BAM.

Infine, si deve indicizzare il file BAM risultante.

```
# Crea un file SAM
bwa mem $REF $R1 $R2 > alignments.sam
# Converte SAM in BAM ordinato.
samtools sort alignments.sam > alignments.bam
# Indicizza il file BAM.
samtools index myfile.bam
```

Il pacchetto samtools dalla versione 1.3 converte e ordina un file SAM in BAM in un solo passaggio:

```
# Crea un file BAM ordinato in una riga.
bwa mem $REF $R1 $R2 | samtools sort > alignments.bam
# Indicizza il file BAM.
samtools index alignments.bam
```

FORMATO VCF

Il VCF (Variant Call Format) è il formato che descrive la variazione degli allineamenti rispetto ad un riferimento.

Un file VCF è in genere creato da un file BAM, mentre il file BAM è stato creato da un file FASTQ e un file FASTA.

Pertanto, il file VCF va considerato come un formato che cattura le differenze di ciascuna delle sequenze nel file FASTQ rispetto al genoma nel file FASTA

```
[tcastign@r033c01s03 ~] bcftools view -H http://data.biostarhandbook.com/variant/subset_hg19.vcf.gz | head -5
```

```
19  400410  rs540061190 CA  C   100 PASS   AC=0;AF=0.00179712;AN=12;NS=2504;DP=7773;EAS_AF=(
19  400666  rs11670588  G   C   100 PASS   AC=5;AF=0.343251;AN=12;NS=2504;DP=8445;EAS_AF=0.3
19  400742  rs568501257 C   T   100 PASS   AC=0;AF=0.000199681;AN=12;NS=2504;DP=15699;EAS_AF
19  400819  rs71335241  C   G   100 PASS   AC=0;AF=0.225839;AN=12;NS=2504;DP=10365;EAS_AF=0.
19  400908  rs183189417 G   T   100 PASS   AC=1;AF=0.0632987;AN=12;NS=2504;DP=13162;EAS_AF=(
```

FORMATO VCF

Types of variants

SNPs

Alignment	VCF representation		
ACGT	POS	REF	ALT
ATGT	2	C	T

Insertions

Alignment	VCF representation		
AC-GT	POS	REF	ALT
ACTGT	2	C	CT

Deletions

Alignment	VCF representation		
ACGT	POS	REF	ALT
A--T	1	ACG	A

Complex events

Alignment	VCF representation		
ACGT	POS	REF	ALT
A-TT	1	ACG	AT

Large structural variants

VCF representation

POS	REF	ALT	INFO
100	T	<DEL>	SVTYPE=DEL;END=300

Example

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">
```

Body

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
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

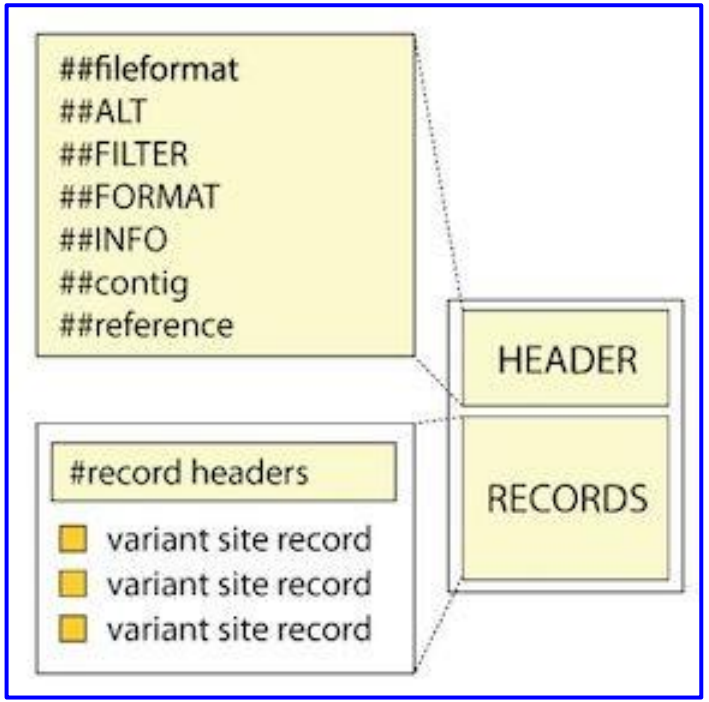
Insertion

Other event

Phased data (G and C above are on the same chromosome)

FORMATO VCF: record

La sezione record di un file VCF è composta da colonne delimitate da tabulazioni, dove le prime otto colonne descrivono una variante e le restanti colonne descrivono le proprietà di ciascun campione. La nona colonna è il FORMATO e ciascuna colonna oltre la nona rappresenta un campione



Come accennato in precedenza, ma vale la pena ribadirlo, un file VCF può contenere un numero qualsiasi di colonne campione, anche migliaia, e può essere pensato come un unico database che rappresenta tutte le variazioni in tutti i campioni

FORMATO VCF

HEADER

```
##fileformat=VCFv4.1
##fileDate=20090805
##tcgaversion=1.1
##vcfProcessLog=<InputVCF=<file1.vcf>,InputVCFSource=<caller1>,InputVCFVer=<1.0>,InputVCFParam=<a1,b>,InputVCFgeneAnno=<anno1.gaf>>
##reference=ftp://ftp.ncbi.nih.gov/genbank/genomes/Eukaryotes/vertebrates_mammals/Homo_sapiens/GRCh37/special_requests/GRCh37-lite.fa
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##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=A,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">
```

INFO meta-information

```
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
```

FILTER meta-information

```
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

FORMAT meta-information

```
##SAMPLE=<ID=NORMAL,Individual=TCGA-01-1000,File=TCGA-01-1000-1.bam,Platform=Illumina,Source=dbGAP,Accession=1234>
##SAMPLE=<ID=TUMOR,Individual=TCGA-01-1000,File=TCGA-01-1000-2.bam,Platform=Illumina,Source=dbGAP,Accession=4567>
##PEDIGREE=<Name_0=TUMOR,Name_1=NORMAL>
```

Fixed fields

Optional: FORMAT field specifying data type + Per-sample genotype data

BODY

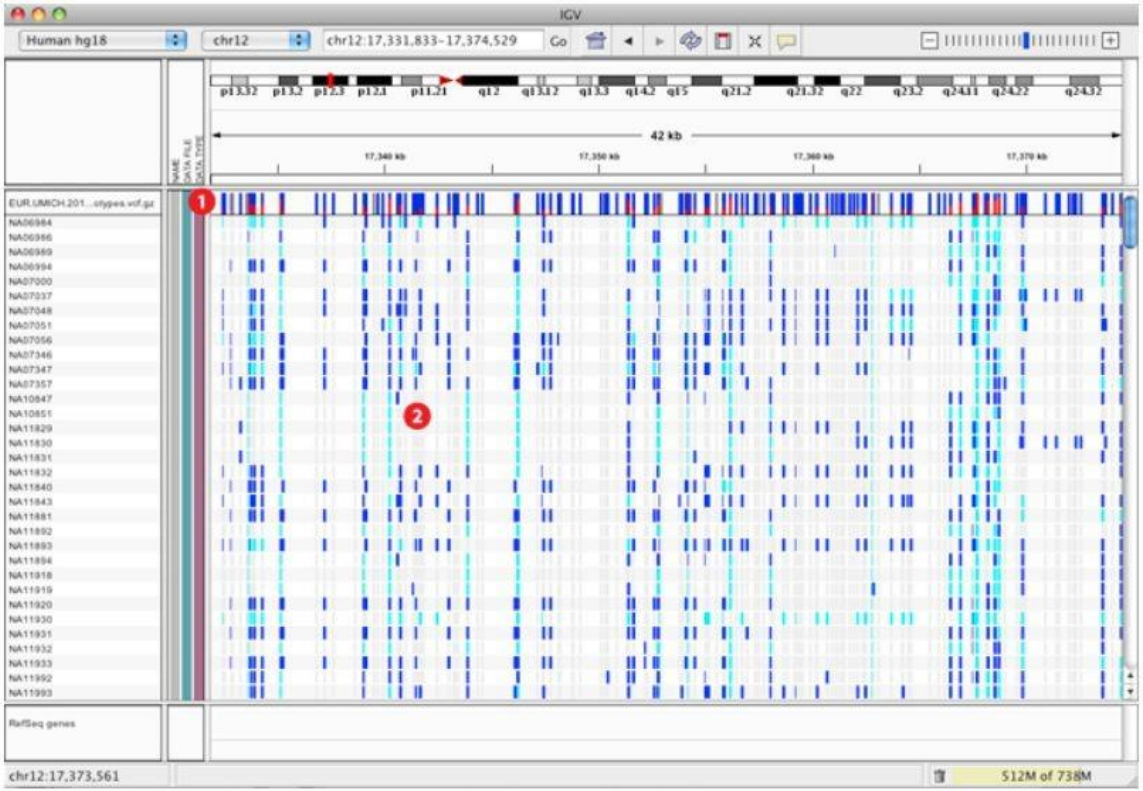
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NORMAL	TUMOR
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51
20	1234567	microsat1	GTC	G,GTCTC	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2

FORMATO VCF: record

Colonna	Campo	Descrizione
1	CHROM	Cromosoma (o conting) su cui si verifica la variante
2	POS	Le coordinate genomiche su cui si verifica la variante
3	ID	un identificatore per la variante (se esiste). In genere un database dbSNP se noto
4	REF	L'allele di riferimento sul filamento anteriore
5	ALT	L'allele(i) alternativo(i) sul filamento anteriore. Potrebbero essere presenti più di uno
6	QUAL	Probabilità che la variante REF/ALT esista in questo sito. È in scala Phred
7	FILTER	il nome dei filtri che la variante non riesce a superare o il valore PASS se la variante ha superato tutti i filtri. Se il valore FILTER è ., al record non è stato applicato alcun filtro
8	INFO	contiene le annotazioni specifiche del sito rappresentate nel formato ID=VALORE
9	FORMAT	annotazioni a livello di campione come TAG separati da due punti

FORMATO VCF: record

Viewing a VCF File with Genotypes



1. Each bar across the top of the plot shows the allele fraction for a single locus.
2. The genotypes for each locus in each sample. Dark blue = heterozygous, Cyan = homozygous variant, Grey = reference. Filtered entries are transparent.

REFERENCE DATA

I dati di riferimento (reference data) rappresentano l'istantanea della conoscenza accumulata in un preciso momento. Vale la pena notare che le informazioni sul genoma umano, a causa della loro importanza per la società, sono trattate in modo molto diverso rispetto alle informazioni su quasi tutti gli altri genomi.
Ai dati relativi al genoma umano è stata applicata la maggior parte della “standardizzazione”.

Per altri organismi che hanno un ampio seguito, diverse organizzazioni sono intervenute per standardizzarne la rappresentazione.
Gli standard “locali” sono più dettagliati perché la quantità di dati accumulata è poca ed è più facile “sorvegliarla”

GENOMIC BUILD

Quando vengono scoperte ulteriori informazioni, potrebbe essere necessario correggere e riorganizzare le precedenti rappresentazioni del genoma.

Le genomic build rappresentano una “edizione”, un’istantanea delle informazioni nel tempo.

Soprattutto quando si ottengono dati da fonti disparate, è essenziale garantire che tutto si riferisca alle stesse informazioni di base.

GENOMIC BUILD

Quando si cambia un genomic build, anche le informazioni associate al genoma devono essere modificate. Ad esempio, aggiungere semplicemente una singola base all'inizio di un genoma significa che le coordinate di tutti gli elementi successivi devono essere modificate, spostate di uno.

Poiché i genomi sono riorganizzati in modi vari e complessi - inversioni, inserzioni, delezioni, ricollocazioni, a volte in modo sovrapposto, rimappatura di una coordinata in una nuova - la localizzazione si rivela un'operazione impegnativa: anche se la sequenza non dovesse cambiare sostanzialmente, le coordinate potrebbero risultare alterate in un modo che potrebbe essere difficile o addirittura impossibile da riconciliare con i dati precedenti. Inoltre alcune località nella versione precedente di un genoma potrebbero “non esistere” nel nuovo genoma

iGENOMES

Gli iGenomes sono una raccolta di sequenze di riferimento e file di annotazioni per organismi comunemente analizzati

I file sono stati scaricabili da Ensembl, NCBI o UCSC

I nomi dei cromosomi sono stati modificati per essere semplici e coerenti con la fonte di download.

Ogni iGenome è disponibile come file compresso che contiene sequenze e file di annotazioni per una singola build genomica di un organismo

iGENOMES

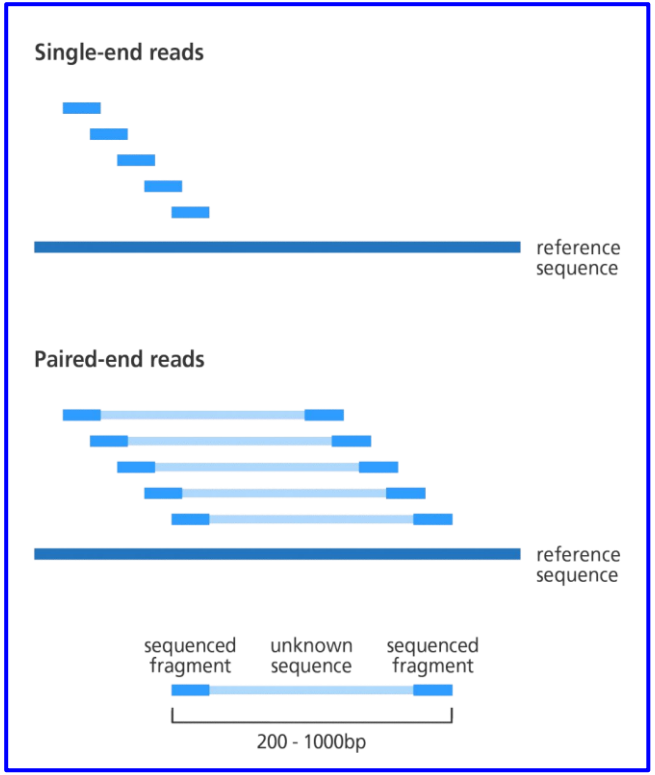
illumina Search SIGN IN VIEW CART CONTACT US

Species	Source	Build(s)			
Arabidopsis thaliana	Ensembl	TAIR10	TAIR9		
	NCBI	TAIR10	build9.1		
Bacillus_cereus strain ATCC 10987	NCBI	2003-02-13			
Bacillus_subtilis strain 168	Ensembl	EB2			
Bos taurus (Cow)	Ensembl	UMD3.1	Btau_4.0		
	NCBI	UMD_3.1.1	UMD_3.1	Btau_4.6.1	Btau_4.2
	UCSC	bosTau8	bosTau7	bosTau6	bosTau4
Caenorhabditis elegans	Ensembl	WBcel235	WBcel215	WS220	WS210
	NCBI	WS195	WS190		
	UCSC	ce10	ce6		
Canis familiaris (Dog)	Ensembl	CanFam3.1	BROADD2		
	NCBI	build3.1	build2.1		

SEQUENZE PAIRED-END

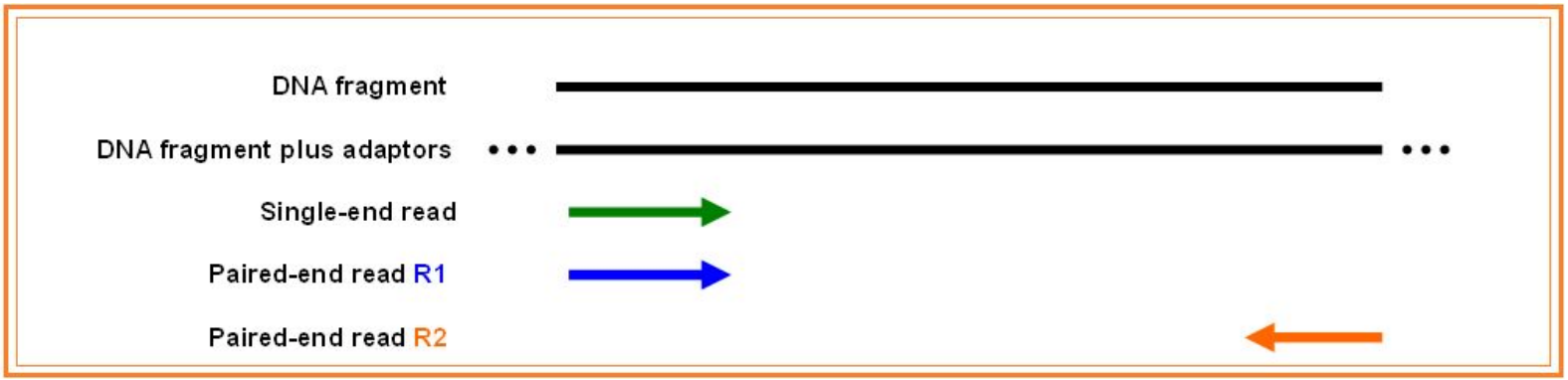
Il sequenziamento accoppiato (paired-end, PE) è un metodo per sequenziare entrambe le estremità di un frammento di DNA e crearne le informazioni di abbinamento disponibili nei dati.

I frammenti di DNA sono in genere più lunghi delle lunghezze di lettura misurate.



SEQUENZE PAIRED-END

Per molte applicazioni è vantaggioso poter misurare (se non l'intero pezzo) almeno entrambe le estremità. Molti processi biologici iniziano in punti specifici del genoma, sapere dove inizia e finisce il frammento può fornire informazioni di fondamentale importanza. Per questo motivo, alcuni strumenti offrono la possibilità di far funzionare il dispositivo in modo diverso per ottenere due misurazioni da un singolo filamento di DNA.



SEQUENZE PAIRED-END

Nel protocollo di sequenziamento Illumina, ad esempio, il primo ciclo di letture è chiamato “single end” (SE) o “prime” letture.

---->

AAAATTTTGGGGCCCC

Se viene utilizzato il protocollo peer-end, dopo aver prodotto le "prime" letture, la stessa sequenza viene invertita all'interno dello strumento, viene complementata in modo inverso e viene eseguita una seconda misurazione per produrre un'altra serie di letture. Queste sono chiamate le “seconde” letture.

<----

GGGGCCCCAAAATTTT

L'effetto finale è che otteniamo due misurazioni da un frammento a singolo filamento:

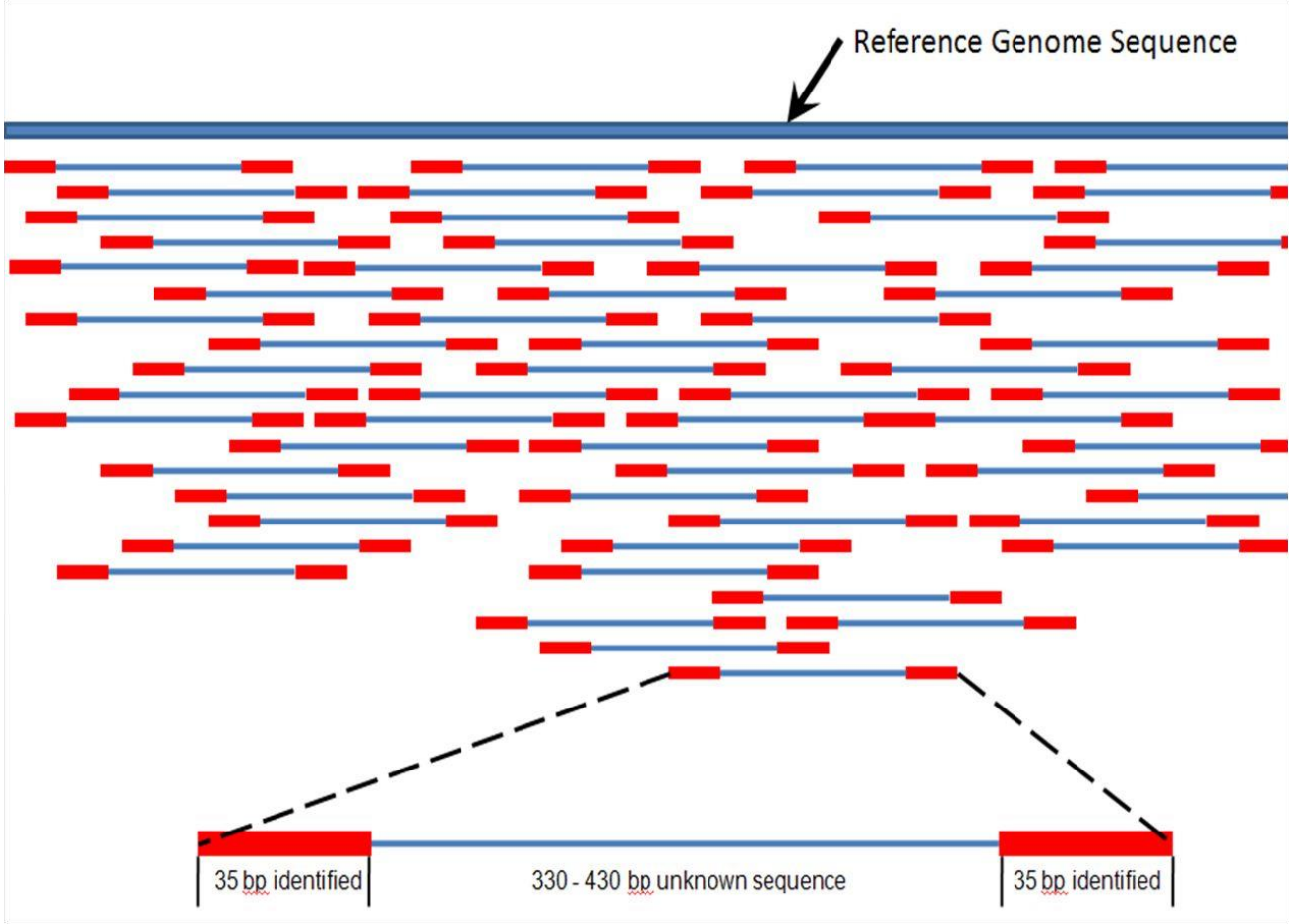
---->

AAAATTTTGGGGCCCC TTTTAAAACCCCGGGG

<----

Le due letture sono generalmente archiviate in file FASTQ separati e sincronizzate per nome e ordine. Ogni lettura nel file 1 ha una voce corrispondente nel file 2.

SEQUENZE PAIRED-END



CONTROLLO QUALITA': FASTQC

Il software più accreditato per il controllo qualità è FastQC sviluppato dal Babraham Institute, un istituto coinvolto nella ricerca biomedica.
Si tratta di uno standard di visualizzazione de facto, ma i suoi risultati non sono sempre i più semplici da interpretare. Il lato positivo è che lo strumento è facile da eseguire (richiede solo Java), semplice, ragionevolmente efficiente e produce grafici esteticamente gradevoli.

FASTQC produce file HTML per ogni file FASTQ analizzato.

FASTQC non esegue il controllo qualità: visualizza solo la qualità dei dati.

REPORT FASTQC BUONO

Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

REPORT FASTQC NON BUONO

Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

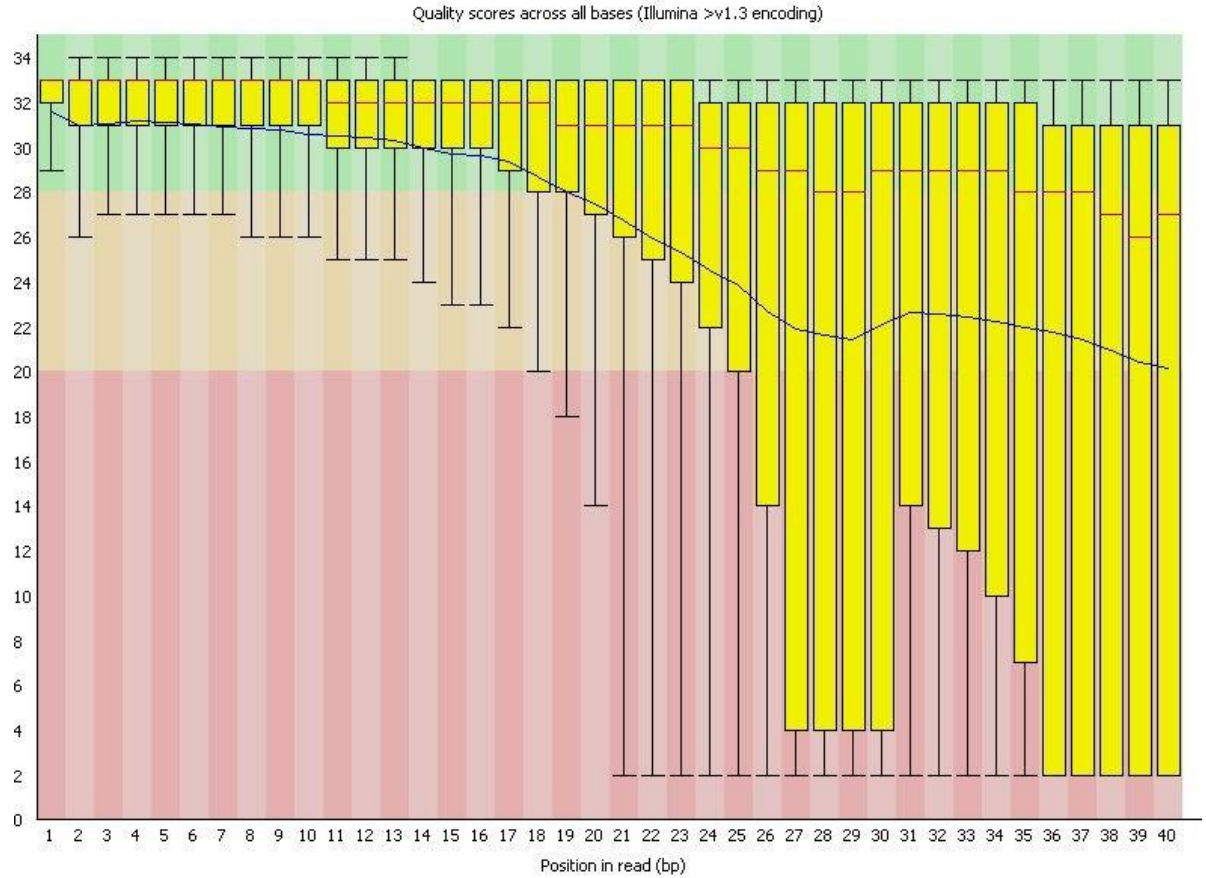
FASTQC: BASE SEQUENCE QUALITY

Questa visualizzazione mostra una panoramica dell'intervallo di valori di qualità su tutte le basi in ciascuna posizione nel file FastQ

Per ogni posizione viene disegnato un grafico di tipo BoxWhisker. Gli elementi della trama sono i seguenti:

- La linea rossa centrale è il valore mediano
- La casella gialla rappresenta l'intervallo interquartile (25-75%)
- I baffi superiori e inferiori rappresentano i punti 10% e 90%.
- La linea blu rappresenta la qualità media

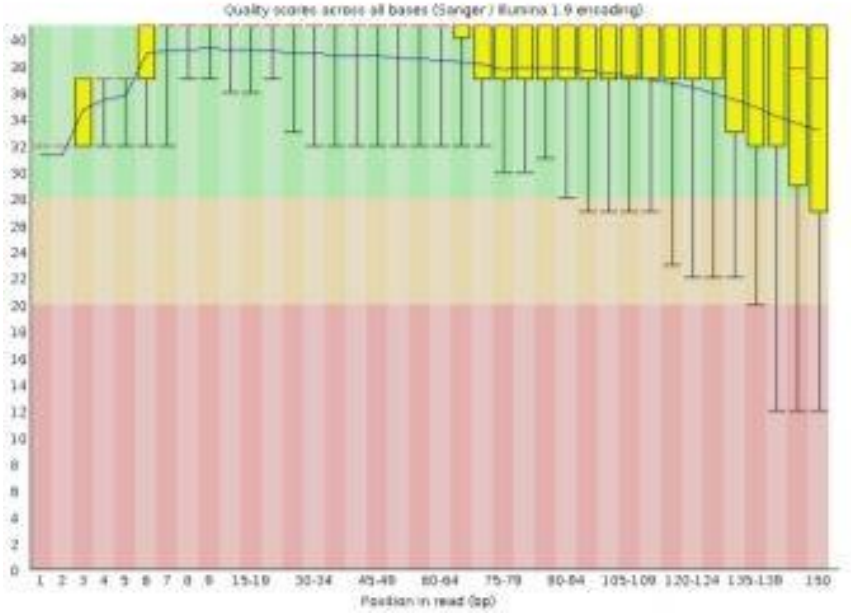
FASTQC: BASE SEQUENCE QUALITY



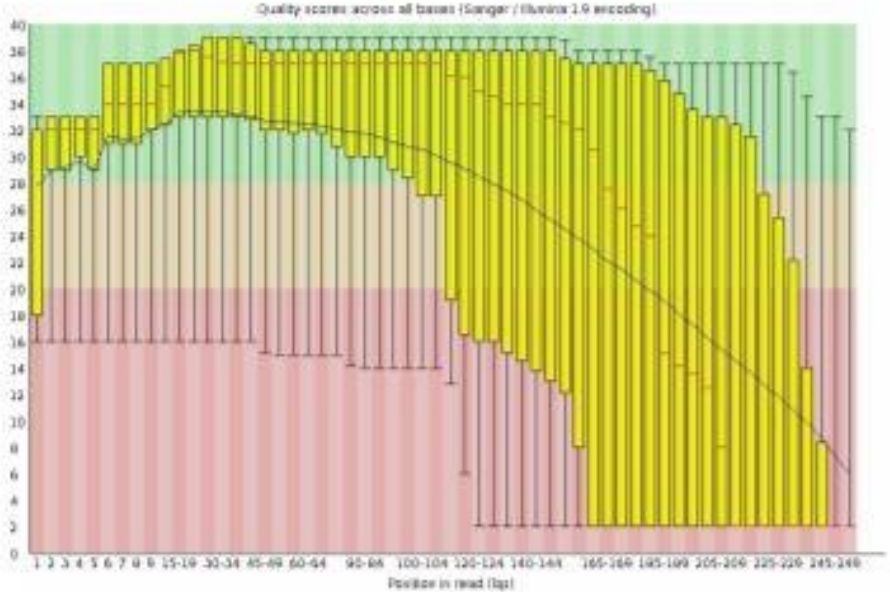
L'asse Y del grafico mostra i punteggi di qualità. Più alto è il punteggio, migliore è l'identificazione della base. Lo sfondo del grafico divide l'asse y in chiamate di qualità molto buona (verde), chiamate di qualità ragionevole (arancione) e chiamate di scarsa qualità (rosso). La qualità delle identificazioni sulla maggior parte delle piattaforme peggiorerà man mano che la corsa procede, quindi è comune vedere le identificazioni delle basi cadere nell'area arancione verso la fine di una lettura.

FASTQC: BASE SEQUENCE QUALITY

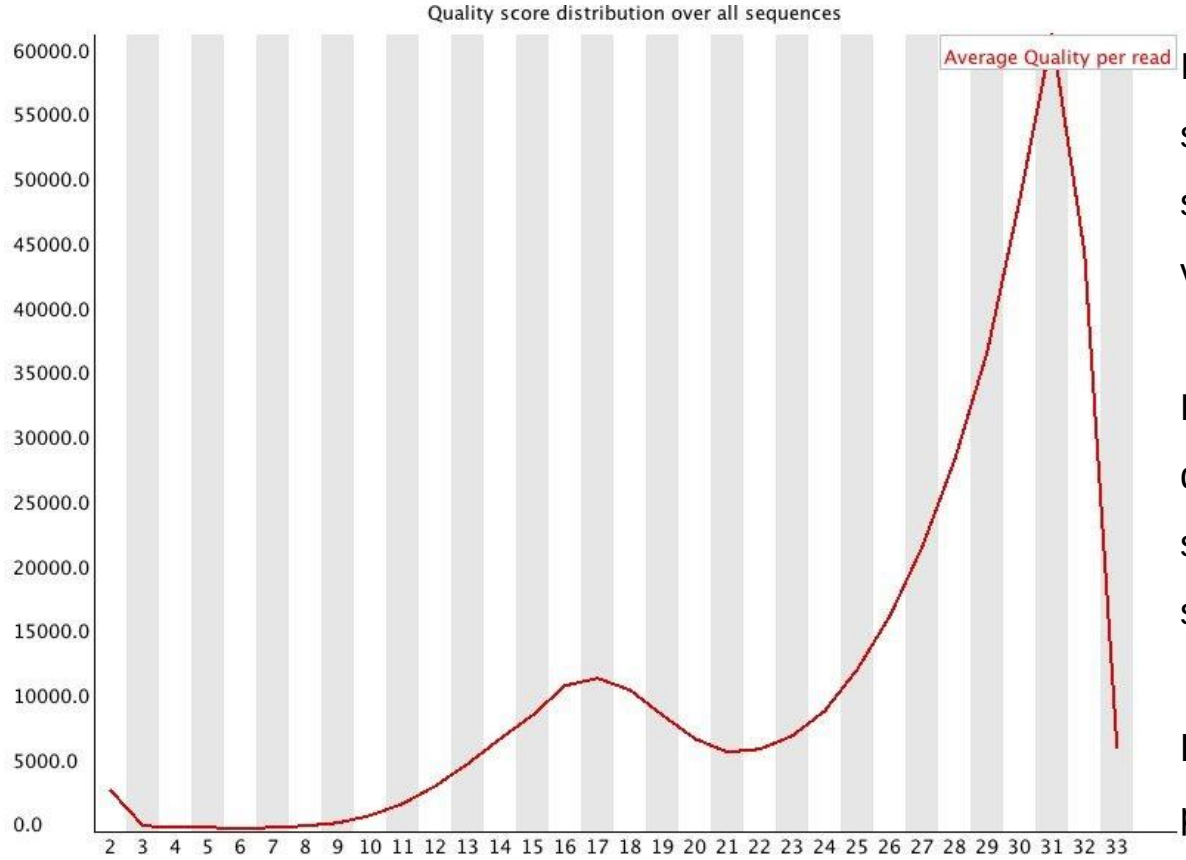
A good per base quality graph



A bad per base quality graph



FASTQC: BASE SEQUENCE QUALITY

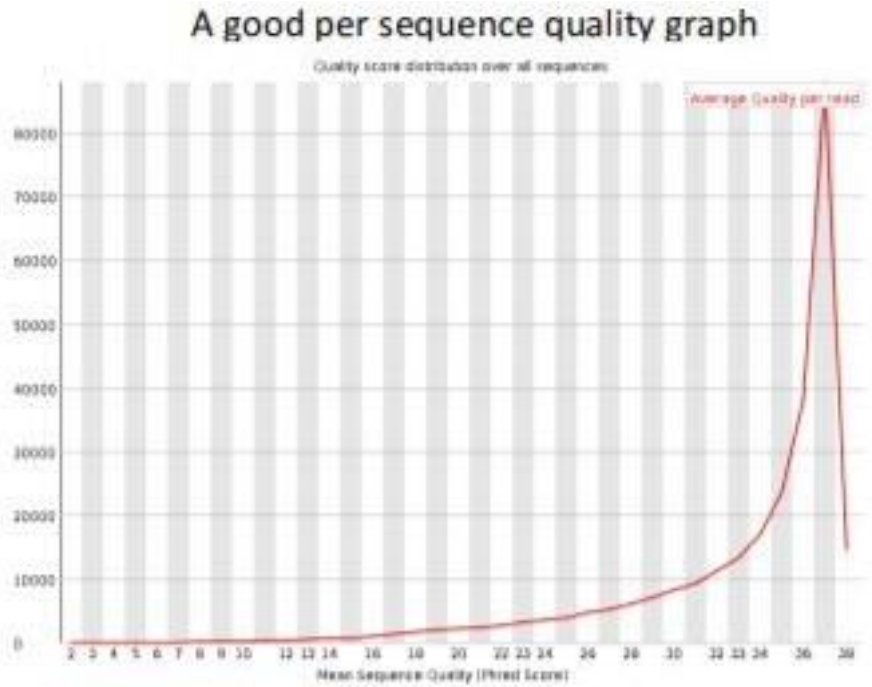


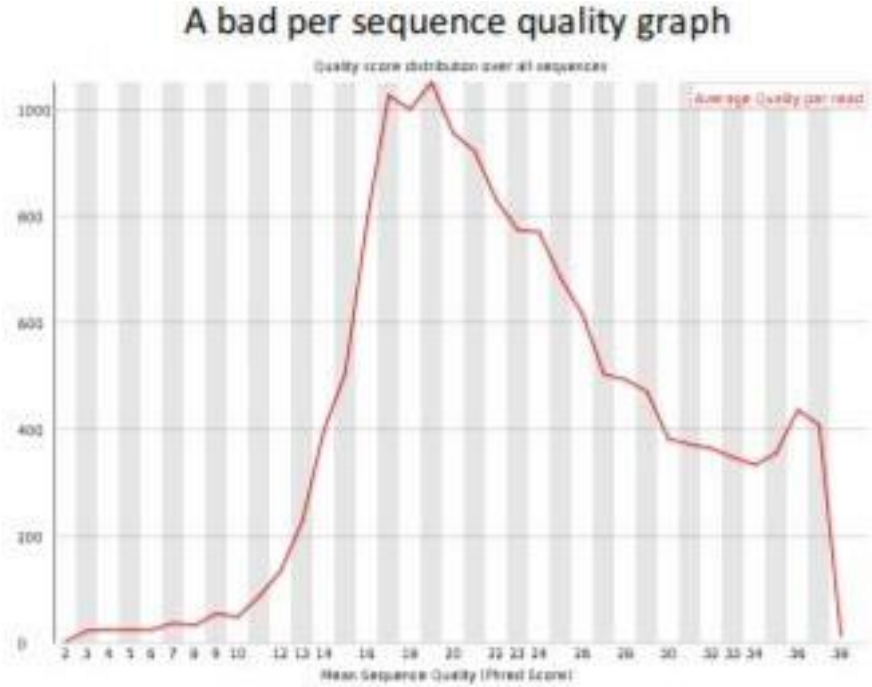
Il report sul punteggio di qualità per sequenza consente di vedere se un sottoinsieme delle sequenze (run) ha valori di qualità universalmente bassi

Può accadere che un run abbia una qualità universalmente scarsa, perché scarsamente riprodotte o per un errore sistematico.

E’ importante che nel grafico ci siano pochi picchi

FASTQC: BASE SEQUENCE QUALITY





END