BIOINFORMATICA II

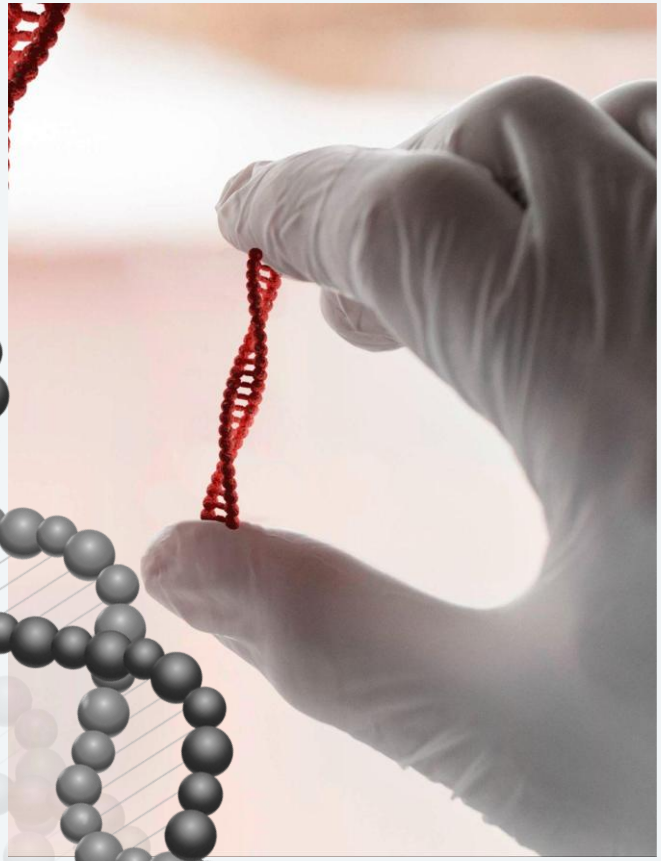
FORMATI

ARGOMENTI

01 FORMATO DIGITALE

02 FORMATI NSG

03 READS SINGLE e READS PAIRED



01

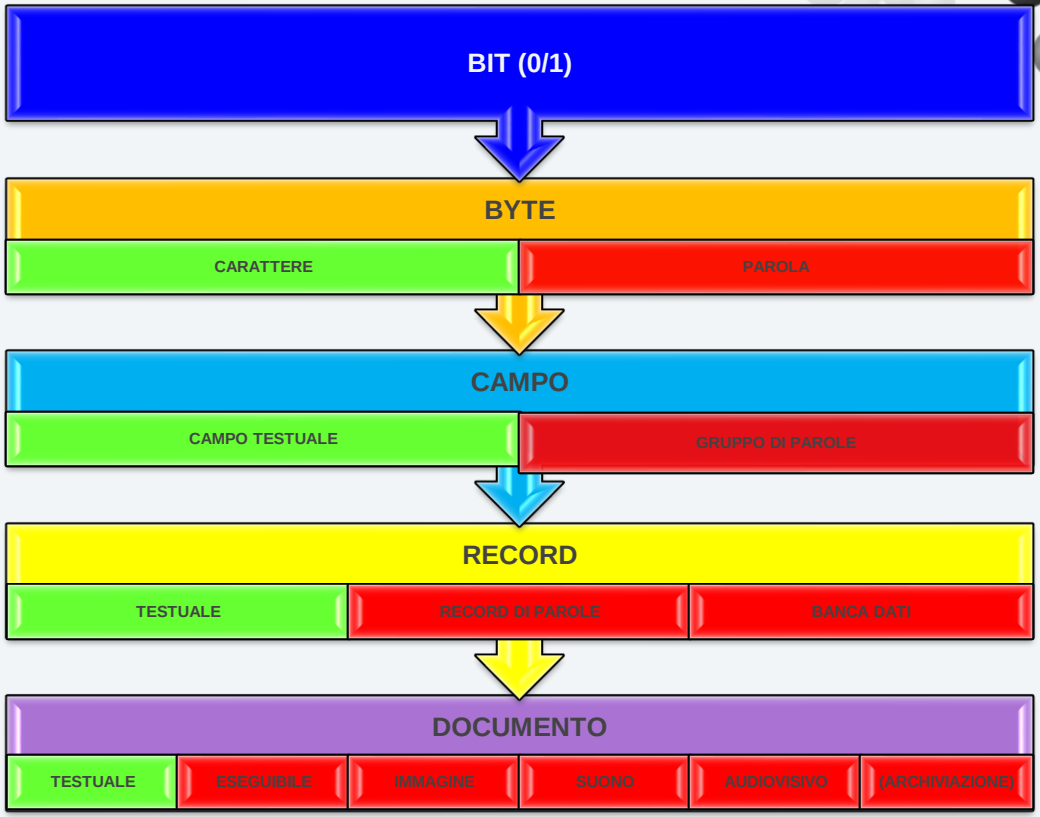
FORMATO DIGITALE

FORMATO DIGITALE: Definizione

Un **formato** è un modo per organizzare e archiviare dati binari su un supporto di memorizzazione digitale (es.: disco magnetico, memoria a stato solido, nastro magnetico, dischi ottici), con caratteristiche che variano a seconda del tipo di contenuto e con disposizione variabile in relazione al tipo di supporto (accesso randomico, SSD e HDD; sequenziale, Magnetic Tape; indicizzato Opictal Device: CD/DVD/Blu-ray).

FORMATO DIGITALE

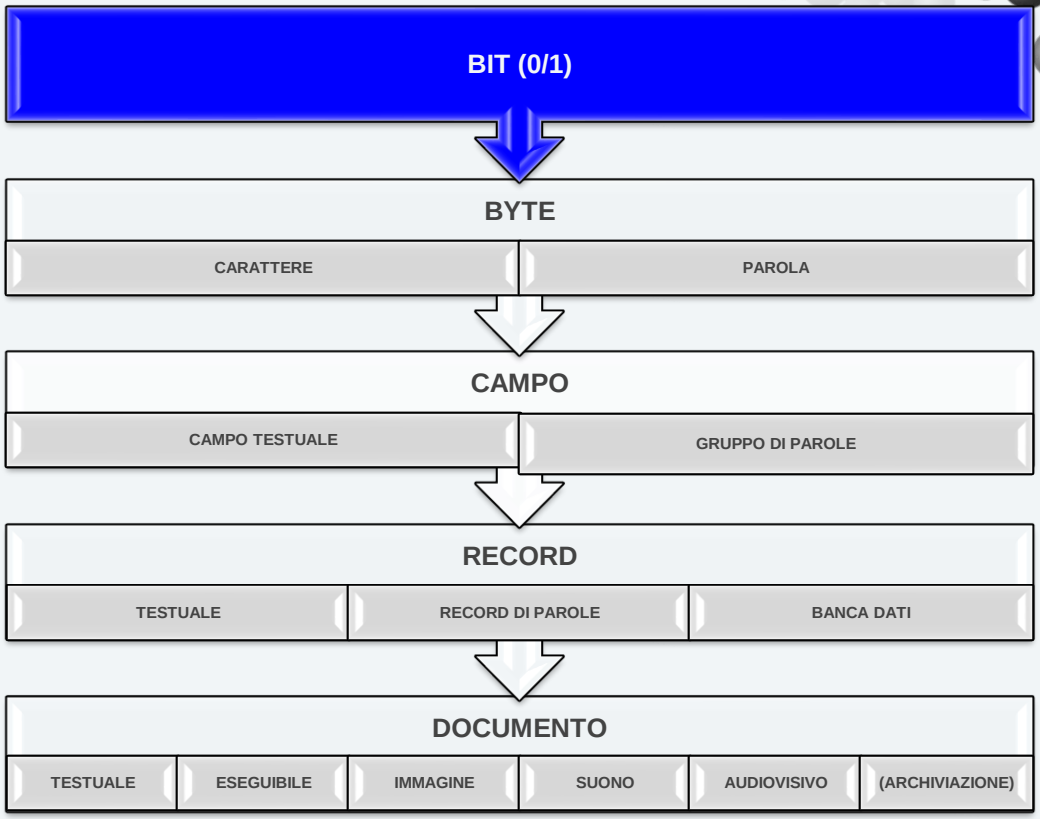
- I primi elaboratori operavano manipolando numeri (naturali ed interi)
- Le istruzioni ed i numeri sulle quali dovevano essere applicate erano prelavati dalle periferiche (tessera perforata, nastro magnetico, disco magnetico) e inviati in Memoria Centrale per essere elaborati
- Con il passare del tempo si gestirono messaggi più articolati tanto da definire una **gerarchia dell'informazione digitale**



FORMATO DIGITALE

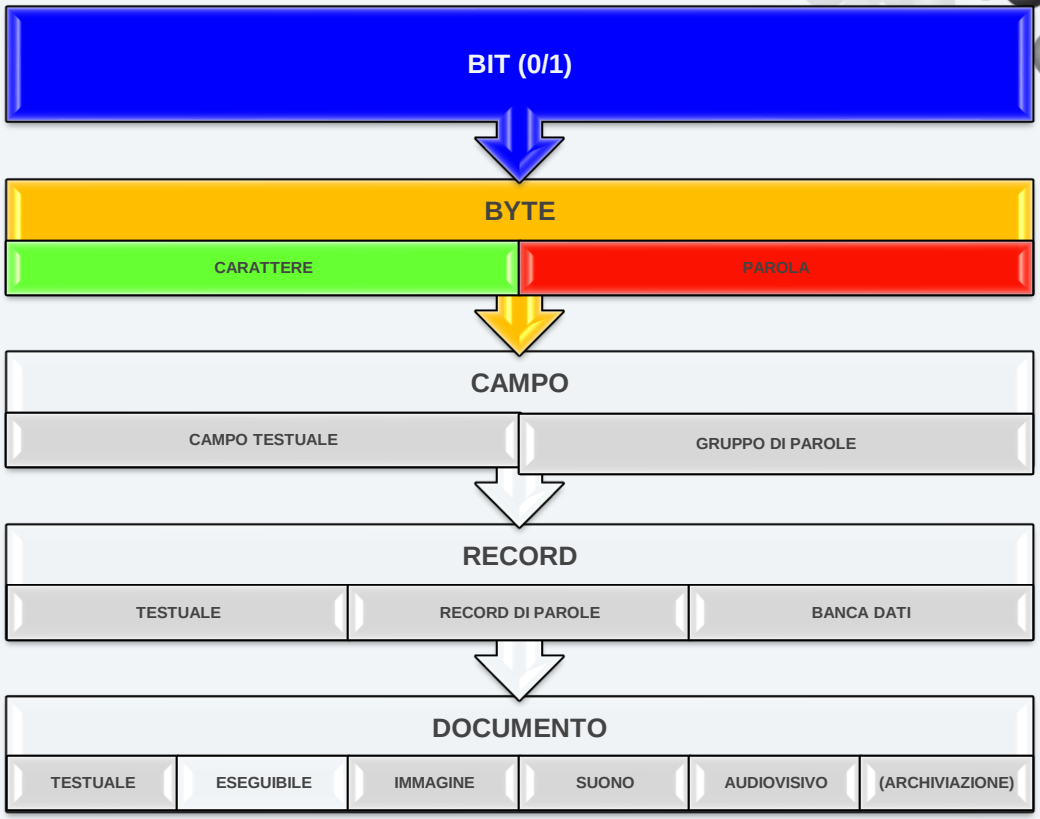
- Il **bit** è la più piccola unità di rappresentazione di informazione digitale ed assume il valore 0 o 1
- Otto bit formano un **byte** ed è la quantità minima standard che un elaboratore elettronico è in grado di gestire
- Altre grandezze utili sono:

Quantità	Simbolo	Nome
0/1	b	bit
8 bit	B	Byte
1024 byte	KB	Kilobyte
1048576 byte	MB	Megabyte
1073741824byte	1GB	Gigabyte
1099511628000byte	1TB	Terabyte



FORMATO DIGITALE

- Un byte al quale è associato un segno testuale prende il nome di **carattere** (*character*)
  Un esempio è il valore numerico 102 espresso in un byte, cioè $(01100110)_2$, che rappresenta la lettera ‘f’ nel codice testuale ASCII
- Un byte che non ha una corrispondenza testuale è una **parola** il cui significato varia in relazione al dominio di applicazione: un operando, nel campo numerico; un colore nel caso di immagini digitali, un campione di un’onda sonora per un suono elettronico



FORMATO DIGITALE: Standard ASCII

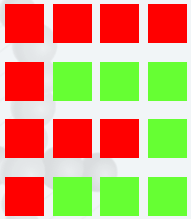
Standard che associa a gruppi di 7 bit simboli alfanumerici

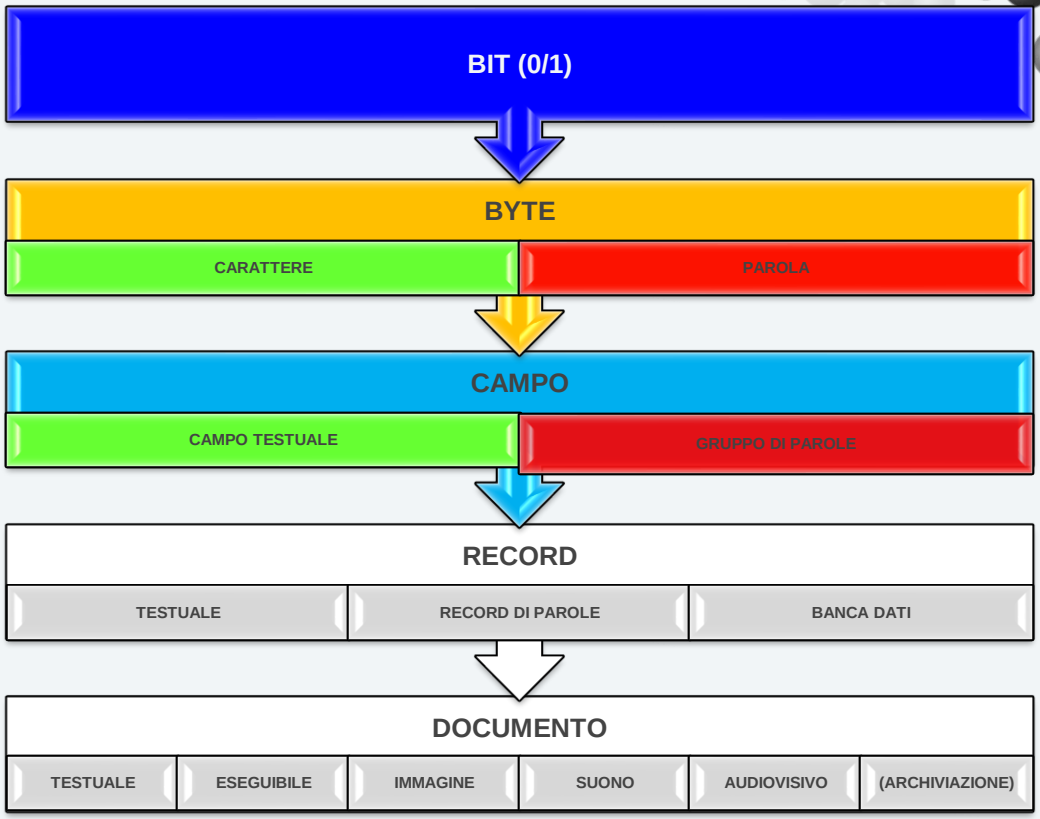
Decimal	Binary	Octal	Hex	ASCII	Decimal	Binary	Octal	Hex	ASCII	Decimal	Binary	Octal	Hex	ASCII	Decimal	Binary	Octal	Hex	ASCII
0	00000000	000	00	NUL	32	00100000	040	20	SP	64	01000000	100	40	@	96	01100000	140	60	`
1	00000001	001	01	SOH	33	00100001	041	21	!	65	01000001	101	41	A	97	01100001	141	61	a
2	00000010	002	02	STX	34	00100010	042	22	"	66	01000010	102	42	B	98	01100010	142	62	b
3	00000011	003	03	ETX	35	00100011	043	23	#	67	01000011	103	43	C	99	01100011	143	63	c
4	00000100	004	04	EOT	36	00100100	044	24	$	68	01000100	104	44	D	100	01100100	144	64	d
5	00000101	005	05	ENQ	37	00100101	045	25	%	69	01000101	105	45	E	101	01100101	145	65	e
6	00000110	006	06	ACK	38	00100110	046	26	&	70	01000110	106	46	F	102	01100110	146	66	f
7	00000111	007	07	BEL	39	00100111	047	27	'	71	01000111	107	47	G	103	01100111	147	67	g
8	00001000	010	08	BS	40	00101000	050	28	(	72	01001000	110	48	H	104	01101000	150	68	h
9	00001001	011	09	HT	41	00101001	051	29	)	73	01001001	111	49	I	105	01101001	151	69	i
10	00001010	012	0A	LF	42	00101010	052	2A	*	74	01001010	112	4A	J	106	01101010	152	6A	j
11	00001011	013	0B	VT	43	00101011	053	2B	+	75	01001011	113	4B	K	107	01101011	153	6B	k
12	00001100	014	0C	FF	44	00101100	054	2C	,	76	01001100	114	4C	L	108	01101100	154	6C	l
13	00001101	015	0D	CR	45	00101101	055	2D	-	77	01001101	115	4D	M	109	01101101	155	6D	m
14	00001110	016	0E	SO	46	00101110	056	2E	.	78	01001110	116	4E	N	110	01101110	156	6E	n
15	00001111	017	0F	SI	47	00101111	057	2F	/	79	01001111	117	4F	O	111	01101111	157	6F	o
16	00010000	020	10	DLE	48	00110000	060	30	0	80	01010000	120	50	P	112	01110000	160	70	p
17	00010001	021	11	DC1	49	00110001	061	31	1	81	01010001	121	51	Q	113	01110001	161	71	q
18	00010010	022	12	DC2	50	00110010	062	32	2	82	01010010	122	52	R	114	01110010	162	72	r
19	00010011	023	13	DC3	51	00110011	063	33	3	83	01010011	123	53	S	115	01110011	163	73	s
20	00010100	024	14	DC4	52	00110100	064	34	4	84	01010100	124	54	T	116	01110100	164	74	t
21	00010101	025	15	NAK	53	00110101	065	35	5	85	01010101	125	55	U	117	01110101	165	75	u
22	00010110	026	16	SYN	54	00110110	066	36	6	86	01010110	126	56	V	118	01110110	166	76	v
23	00010111	027	17	ETB	55	00110111	067	37	7	87	01010111	127	57	W	119	01110111	167	77	w
24	00011000	030	18	CAN	56	00111000	070	38	8	88	01011000	130	58	X	120	01111000	170	78	x
25	00011001	031	19	EM	57	00111001	071	39	9	89	01011001	131	59	Y	121	01111001	171	79	y
26	00011010	032	1A	SUB	58	00111010	072	3A	:	90	01011010	132	5A	Z	122	01111010	172	7A	z
27	00011011	033	1B	ESC	59	00111011	073	3B	;	91	01011011	133	5B	[	123	01111011	173	7B	{
28	00011100	034	1C	FS	60	00111100	074	3C	<	92	01011100	134	5C	\	124	01111100	174	7C	|
29	00011101	035	1D	GS	61	00111101	075	3D	=	93	01011101	135	5D	]	125	01111101	175	7D	}
30	00011110	036	1E	RS	62	00111110	076	3E	>	94	01011110	136	5E	^	126	01111110	176	7E	~
31	00011111	037	1F	US	63	00111111	077	3F	?	95	01011111	137	5F	_	127	01111111	177	7F	DEL

https://web.alfredstate.edu/faculty/weimandn/miscellaneous/ascii/ascii_index.html

FORMATO DIGITALE

- Il **campo** è un raggruppamento di caratteri o di parole
Ad esempio, il campo “1947-05-13” è associabile alla data di nascita di una persona
Il gruppo di parole
AA0000 AA0000 AA0000 AA0000 A0000 00FF00 00FF00 00FF00 AA0000 AA0000 AA0000 00FF00 AA0000 00FF00 00FF00 00FF00 può identificare una immagine che rappresenta la lettera F





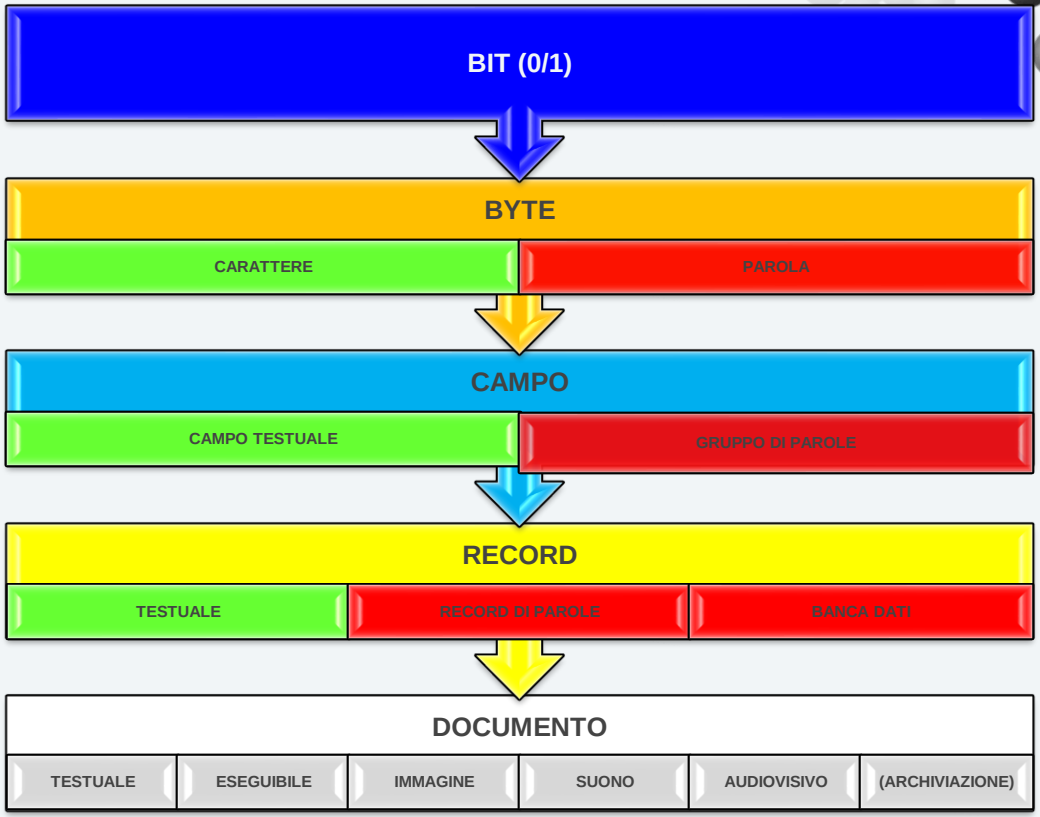
FORMATO DIGITALE

- Il **record** è una collezione di campi

- Ad esempio il record LIBRO è composto dai campi

<TITOLO, AUTORE, DATA, EDITORE>

e una istanza può essere <"Il Codice da Vinci ", "Dan Brown", "2000", "Random House">



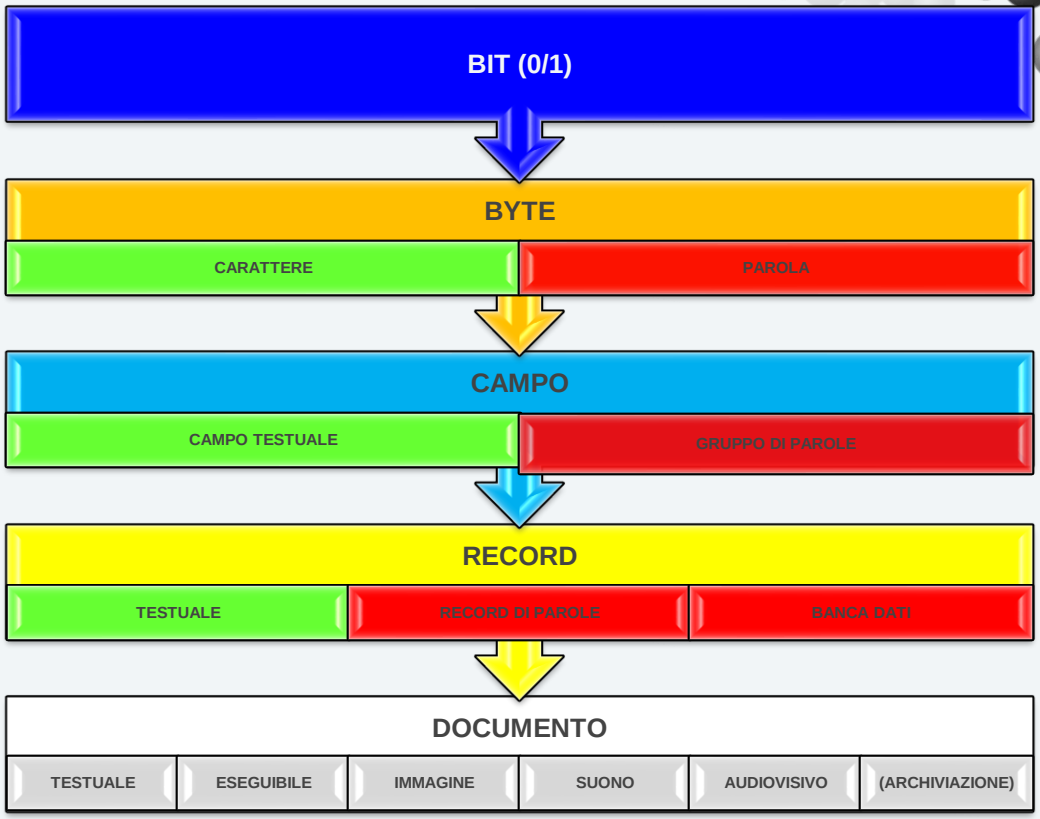
FORMATO DIGITALE

- Più record con relazioni tra di loro formano una **tabella di una banca dati**

STUDENTE
Matricola, Nome, Cognome, Indirizzo, Città, Facoltà

ESAMI STUDENTE
Matricola, Id_Corso, Voto

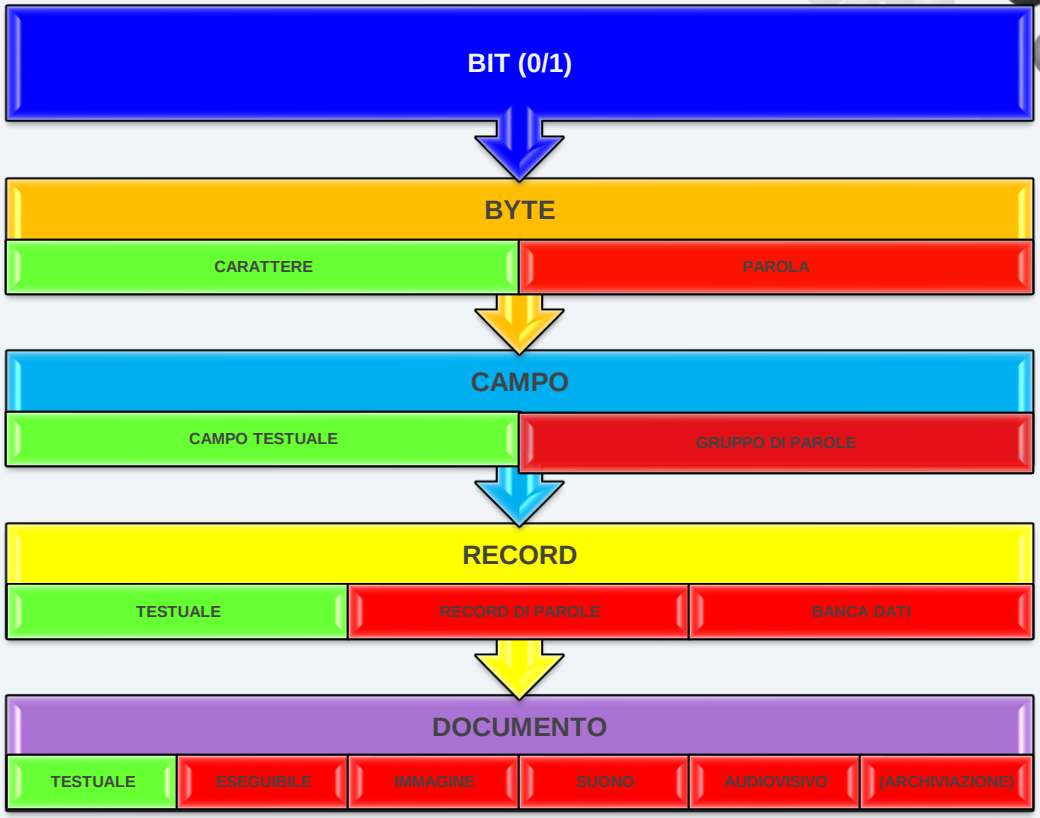
CORSO
CodiceCorso, Nome, Sinossi, Docente



FORMATO DIGITALE

- Un **documento digitale** (o file) è una collezione di record omogeni, correlati e con un ordine prestabilito

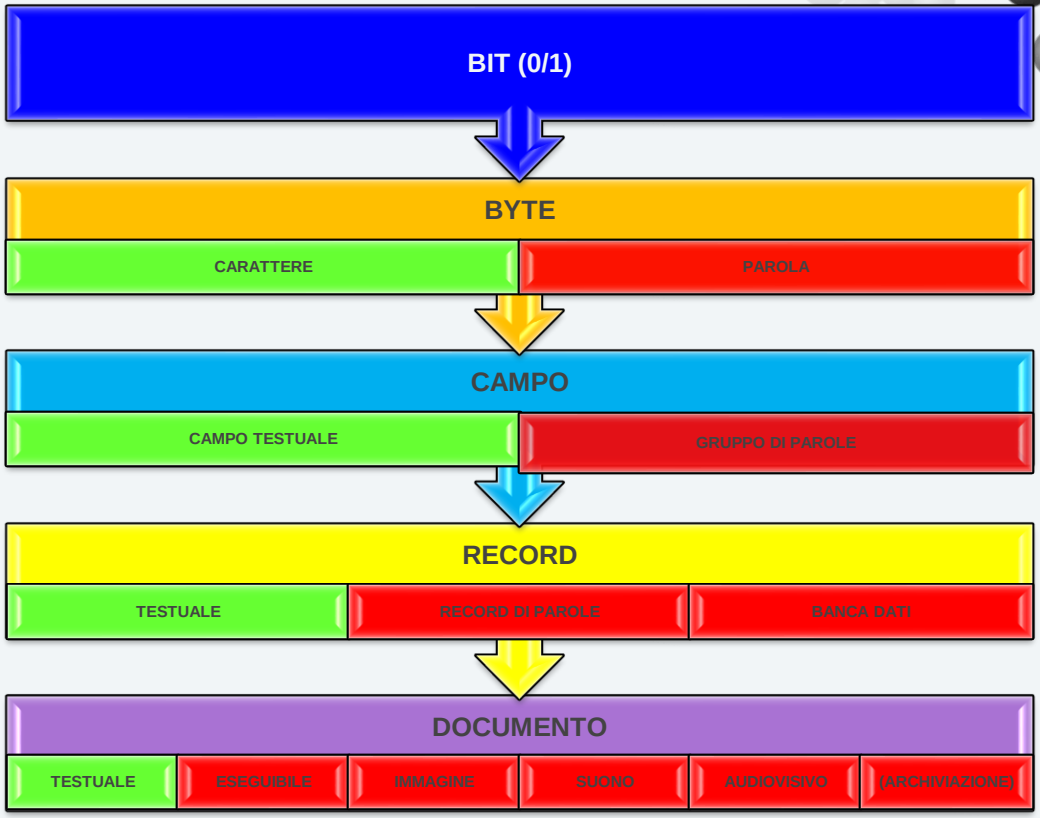
- Un documento digitale in informatica ha due caratterizzazioni: **testuale** e **binario**. Nel primo caso si tratta di una collezione di record testuali; nel secondo è un insieme di record avente un significato (numerico, immagine, suono, audiovisivo, multimediale) esplicitato mediante campi aggiuntivi essenziali (metadati tecnici)



FORMATO DIGITALE

- Il **record binario** necessita di informazioni ausiliarie (metadati tecnici) per stabilire la classe di appartenenza del documento (immagine, video, audio) e ne consente la riproduzione con il corretto applicativo
- La totalità dei record e i metadati tecnici costituiscono il **formato** del documento

Formato	Nome	Tipo di parola
Immagine	JPEG	Binario
Audio	MP3	Binario
Video	MP4	Binario
Testo	TXT	Testuale
Testo con marcatori	HTML	Testuale



FORMATO DIGITALE

780 x 749

30C x 2ED

https://hexed.it

Nuovo file | Apri file | Salva con nome | Annulla | Ripeti | Strumenti | Traduci | Impostazioni | Aiuto

Informazioni File

Nome File	Viterbo.bmp
Dim. File	2.336.950 Bytes (2.283 KiB)

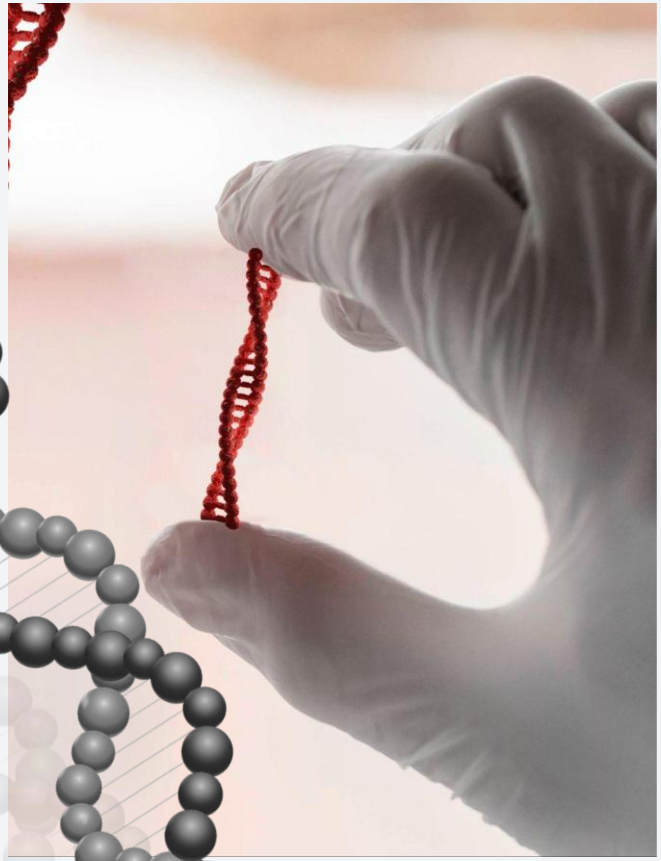
Ispettore dati (Little-endian)

Tipo	Senza segno (+)	Con segno (±)
Intero a 8-bit	66	66
Intero a 16-bit	19778	19778
Intero a 24-bit	11947330	-4829886
Intero a 32-bit	2830519618	-1464447678
Intero a 64-bit (+)	153154374978	
Intero a 64-bit (±)	153154374978	
Virg. mob. a 16-bit	21,03125	
Virg. mob. a 32-bit	-2,0239564e-14	
Virg. mob. 64-bit	7,5668315186918019e-313	
LEB128 (+)	66	
LEB128 (±)	-62	
Data/Ora MS-DOS	22/05/2064 09:42:04 Local	
Data/Ora OLE 2.0	30/12/1899 00:00:00.000 UTC	
Data/Ora UNIX	11/09/2059 15:26:58 UTC	
Data/Ora Macintosh HFS	10/09/1993 17:26:58 Local	
Data/Ora Macintosh HFS+	10/09/1993 15:26:58 UTC	
UTF-8 Character	B	
Binario	○ ● ○ ○ ○ ○ ● ○	

Ispettore dati (Big-endian)

-Senza Titolo- | Viterbo.jpg | Viterbo.bmp

```
00000000  42 4D B6 A8 23 00 00 00 00 00 46 00 00 00 38 00  BM¶¿#.....F...8.
00000010  00 00 0C 03 00 00 ED 02 00 00 01 00 20 00 03 00  ......φ..... ...
00000020  00 00 70 A8 23 00 13 0B 00 00 13 0B 00 00 00 00  ..p¿#...........
00000030  00 00 00 00 00 00 00 00 FF 00 00 FF 00 00 FF 00  ........ .. .. .
00000040  00 00 00 00 00 FF 44 5A 5F FF 56 6C 71 FF 60 7D  ..... DZ_ Vlq `}
00000050  82 FF 7C 99 9E FF 85 A2 A9 FF 7D 99 A0 FF 6E 88  é |ÖÞ àó⌐ }Öá nê
00000060  8F FF 75 8F 95 FF 6E 8B 8F FF 72 91 92 FF 51 67  Å uÅò nïÅ ræÆ Qg
00000070  6C FF 3F 51 58 FF 56 6C 72 FF 55 6F 76 FF 5D 72  l ?QX Vlr Uov ]r
00000080  7A FF 54 6B 73 FF 5D 78 82 FF 61 77 82 FF 58 7B  z Tks ]xé awé X{
00000090  7F FF 5E 7C 81 FF 51 6B 71 FF 55 70 74 FF 52 71  ⌂ ^|ü Qkq Upt Rq
000000A0  74 FF 4D 70 74 FF 4D 73 77 FF 59 82 85 FF 61 87  t Mpt Msw Yéà aç
000000B0  8C FF 65 85 8B FF 48 62 68 FF 34 48 4D FF 2C 3D  î eàï Hbh 4HM ,=
000000C0  40 FF 27 36 39 FF 29 38 3A FF 2B 3A 3C FF 25 2E  @ '69 )8: +:< %.
000000D0  32 FF 1E 2A 2E FF 2C 3B 3E FF 2B 3A 3D FF 2A 39  2 .*. ,;> +:= *9
000000E0  3C FF 3D 4B 51 FF 24 37 3C FF 3D 50 57 FF 81 A0  < =KQ $7< =PW üá
000000F0  A9 FF 94 B0 B7 FF 93 AB B1 FF 9A B0 B6 FF 94 AC  ⌐ ö╲╖ ▒½▒ Ü╲╢ ö¼
00000100  B2 FF 92 AA B0 FF 8B A2 AA FF 82 97 9F FF 81 94  ▓ Æ¬╲ ïó¬ éùƒ üö
00000110  9C FF 83 96 9E FF 7C 8F 96 FF 75 87 8E FF 84 96  £ âûÞ |ÅÅ uçÃ äû
00000120  9D FF 82 94 9B FF 85 97 9E FF 75 87 8E FF 5A 66  ¥ éö¢ àùÞ uçÃ Zf
00000130  6C FF 38 44 4A FF 36 44 4A FF 33 40 48 FF 36 43  l 8DJ 6DJ 3@H 6C
00000140  4B FF 34 42 48 FF 32 42 48 FF 3A 4A 50 FF 35 49  K 4BH 2BH :JP 5I
00000150  4E FF 39 4C 51 FF 38 48 4E FF 36 44 4A FF 38 46  N 9LQ 8HN 6DJ 8F
00000160  4C FF 33 40 48 FF 3A 47 4F FF 3C 49 51 FF 25 38  L 3@H :GO <IQ %8
00000170  3F FF 44 5A 60 FF 5E 77 7B FF 6A 84 8A FF 80 9D  ? DZ` ^w{ jäè Ç¥
00000180  A2 FF B1 CC D6 FF AA C4 D2 FF 88 A2 B2 FF 84 99  ó ▒╠╓ ¬─┬ êó▓ ãÖ
00000190  A8 FF 86 9E AA FF 8F AA B4 FF 93 AF B6 FF 88 A2  ¿ åÞ¬ Å¬┤ ô»╢ êó
000001A0  A8 FF 84 9C A2 FF 84 9C A2 FF 85 9D A3 FF 85 9D  ¿ ä£ó ä£ó à¥ú à¥
000001B0  A3 FF 83 9B A1 FF 7D 95 9B FF 7A 90 96 FF 71 84  ú â¢í }ò¢ zÉû qä
000001C0  8B FF 7D 90 97 FF 76 88 8F FF 6B 7D 84 FF 75 88  ï }Éù vêÅ k}ä uê
000001D0  8D FF 7C 8F 94 FF 55 68 6D FF 35 48 4D FF 37 47  ì |Åö Uhm 5HM 7G
000001E0  4D FF 34 44 4A FF 35 43 49 FF 35 43 49 FF 39 49  M 4DJ 5CI 5CI 9I
```



02

FORMATI NSG

FORMATI NSG: Generalità

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

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

Ogni formato lo rende adatto a obiettivi specifici
Si può riconoscere il formato dalla sua estensione (le ultime tre lettere del nome del file)

FORMATI NSG: FASTQ

Il formato FASTQ (*FAST-Alignment Quality*) **è** lo standard *de facto* **usato per conservare sequenze sperimentali (reads) prodotte da uno strumento di sequenziamento**

È un **formato testuale** in cui ad ogni base si associa una misura di affidabilità/qualità

È una evoluzione del FASTA (i primi dati ottenuti dal metodo Sanger erano dei file FASTA senza Quality)

L'estensione è **fastq** o **fq**

FORMATI NSG: FASTQ

Il formato FASTQ è composto da 4 campi (ciascuno disposto su una riga singola):

1. **Intestazione** (*Header*): inizia con il simbolo @, prosegue con un ID e un altro testo facoltativo
2. **Sequenza** (*sequence*): la sequenza dei nucleotidi (può essere disposta su una sola riga o più righe (per facilitare la visione))
3. **Delimitatore** (*space*): la terza sezione è contrassegnata dal segno iniziale + e può essere (facoltativamente) seguito dalla Intestazione della prima sezione
4. **Punteggio di qualità** (*quality score*): l'ultima riga codifica i valori di qualità per la sequenza nella seconda sezione (ad ogni nucletodite è associato un valore numercio: le due righe hanno lunghezza uguale) secondo lo standard Phred

FORMATI NSG: FASTQ

Header Sequence Quality

```
@HWI-ST227:389:C4WA2ACXX:7:1204:2272:59979
GGAGGAAGGTCCTCGCTCCTCTTTCATATAAGGGAAATGGCTGAAT
+
FFFFHHHHHHJIJJJJJJJJIJJJIGIGIGGIJJIJIJJJJJJIII
@HWI-ST227:389:C4WA2ACXX:7:1205:15214:42893
GAGGATCCCAGGGAGGAAGGTCCTCGCTCCTCTTTCATCTAAGGGA
+
12BAFB?A:3<AE1@<FF;1*@EG*)?0?DBD>9BF9B*?######
@HWI-ST227:389:C4WA2ACXX:8:2208:2467:44624
AAAGAGGAGAGAGGACCATCCTCCCTGGGATCCTCAGAAGTCTACT
+
BDDA:DB?2AA@FC>F?EEGC<FED>GFD;?GBB?<?F99*/9?9?
```

FORMATI NSG: FASTQ

Comando Linux per vedere la prima parte del file:

head *nomefile* -n 20

```
fliberat@raganella:/data/LIBERATI/ARTEMIA/RAW$ head C1_1.fq -n 16
@LH00587:21:227YCTLT4:7:1101:36509:1028 1:N:0:NGTGACAACC+NTCATGCGTT
CNCACCCCTCTTTTATTTATAGTGATGACTTTTTGATTGCTCCCCTCACGAGCTCAATTCTTGACTAGAAGAGTCTTAAACAGATTNAACCGAGAGATGTCCCTGCTTATAGGGACAGCTAGAATCGGAACCAACATTCTAGTAATTGCT
+
I#IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII#IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IIIIIII9I
@LH00587:21:227YCTLT4:7:1101:36525:1028 1:N:0:NGTGACAACC+NTCATGCGTT
CNCACCCCTCTTTTATTTATAGTGATGACTTTTTGATTGCTCCCCTCACGAGCTCAATTCTTGACTAGAAGAGTCTTAAACAGATTNAACCGAGAGATGTCCCTGCTTATAGGGACAGCTAGAATCGGAACCAACATTCTAGTAATTGCT
+
I#IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII#II9IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@LH00587:21:227YCTLT4:7:1101:41672:1028 1:N:0:NGTGACAACC+NTCATGCGTT
GNGAATTCGGCTCATATTAGTGGCAACAAATCCAGGTGTCACACATTGGACTACGACACCGCATGGCTTATATTCCAAGGCTAAATCCTGGCTAAACTTGAGCACAAACGCCTTAGTGCTGGAATACATAGTCATGAGAGGAGTAGGCAG
+
I#IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IIIIIIIIIIIIIIIIIIIIIIIII-IIIIIIIIIIIIIIIIIIIIIIIIIIII9IIIIIIIIIIIIIIIIIIIIIII9IIIIIIIIIII
@LH00587:21:227YCTLT4:7:1101:41720:1028 1:N:0:NGTGACAACC+NTCATGCGTT
GNTACCTCATCAAAAACAAATGAGCCTGCCTGAATAGATCCAATATTTTTATTTCCTTTAAAGTTTATTTCTTCTAGCCTCTGACACTCTTCAAATGGAAAATCCTGGCTGATGCCGATTAGTCCTGAAGCTTGAGCTGAAATTATCCGT
+
I#IIIIIIIIIIIIIIIIIIII-IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
fliberat@raganella:/data/LIBERATI/ARTEMIA/RAW$ |
```

Comando Linux per vedere l'intero file (si consiglia di applicare il more per una visione parziale, riga per riga):

cat *nomefile* | more

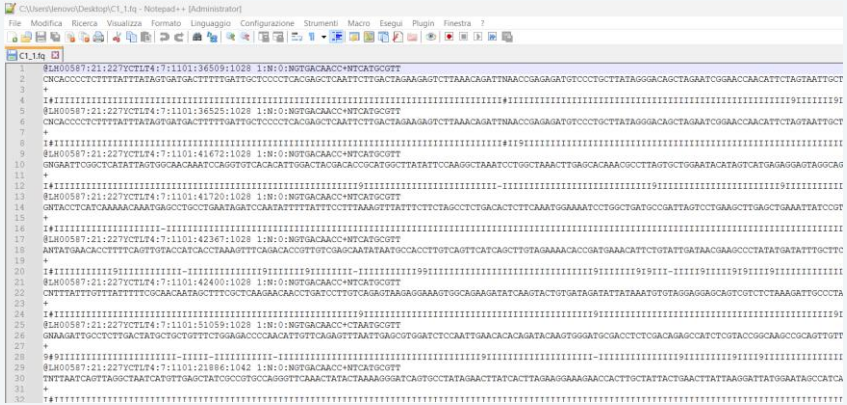
FORMATI NSG: FASTQ

Programma Windows per vedere il file:

NotePad

Programma Windows per vedere file di grande dimensione:

NotePad++ **MEGA**



FORMATI NSG: FASTQ - PHRED

La metrica Phred mappa i numeri che rappresentano la probabilità di errore su singoli nucleotidi

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

Valore PHRED (PhredQuality Score)	Probabilità di errore (Probability of incorrect base call)	Precisione della misurazione (Base call accuracy)
10	1 base errata su 10	90%
20	1 base errata su 100	99%
30	1 base errata su 1000	99.9%
40	1 base errata su 10000	99.99%
50	1 base errata su 100000	99.999%
60	1 base errata su 1000000	99.9999%

FORMATI NSG: BED

Il **BED è un formato testuale nel quale sono riportate le posizioni di una regione appartenente ad un genoma** (questa tipologia è detta *formato intervallo)*.
Ogni campo è delimitato da una tabulazione e contiene informazioni su coordinate cromosomiche quali inizio, fine, filamento, valore e altri attributi

L'estensione è **.bed**

Il BED ha tre colonne obbligatorie <CROMOSOMA INIZIO FINE> e altre nove opzionali

Chr7	**127471196**	**127472363**
Chr7	**127472363**	**127473530**
Chr7	**127473530**	**127474697**

FORMATI NSG: BED

Colonna	Campo	Definizione	Obbligatorietà
1	**chrom**	Nome del cromosoma (es.: chr3, chrY) o dello scaffold (es.: 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 riportatele annotazioni (inizio di un codone o di un gene)	No
8	**thickEnd**	Fine coordinate dalle quali sono riportatele annotazioni (fine di un codone o 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 iniziale dei blocchi (relazione con il campo blockCount)	No

FORMATI NSG: GFF

Il **GFF è un formato testuale** delle caratteristiche generali (gene-finding format, generic feature format, GFF) ed è utile **per archiviare informazioni descriventi geni e altre caratteristiche di sequenze di DNA, RNA e proteine**

È costituto da nove campi.

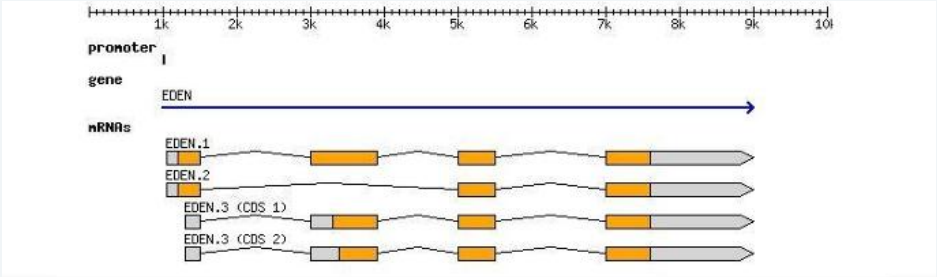
L'estensione è **.bed**

```
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
```

FORMATI NSG: GFF

Colonna	Campo	Definizione	Obbligatorietà
1	**seqid**	Il nome delle sequenze	Si
2	**source**	L'algoritmo che ha generato le sequenze (nome del software o del database)	Si
3	**type**	Il tipo (gene o esone). In GFF3, il tipo e le relazioni devono rispettare lo standards rilasciato da Sequence Ontology Project	Si
4	**start**	Inizio della caratteristica genomica (inizio a partire dal valore 1)	Si
5	**end**	Fine della caratteristica genomica	Si
6	**score**	Valore numerico che indica l'attendibilità della annotazione (il punto '.' significa attendibilità nulla)	Si
7	**strand**	Carattere che indica lo strand della caratteristica (il filamento) "+" (positivo, o 5'->3'), "-", (negativo, or 3'->5'), "." (non determinato), o "?" per caratteristiche rilevanti ma il cui strand non è noto	Si
8	**phase**	Specifico per le annotazioni delle regioni di codifica di proteine (CDS)	Si
9	**attributes**	Coppie di marcatori per informazioni aggiuntive alle annotazioni (es: ID=…. ; NAME=….)	Si

FORMATI NSG: 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
```

FORMATI NSG: SAM

Il **file SAM** (*Sequence Alignment Maps Format*) è un formato testuale che **contiene informazioni sugli allineamenti** Reads/genoma o Reads/trascrittoma

Rappresenta i risultati dell'allineamento di un file FASTQ a un file FASTA di riferimento e descrivono i singoli allineamenti a coppie trovati.

L’estensione è **.sam**

Algoritmi diversi possono creare allineamenti diversi (e quindi file SAM differenti)

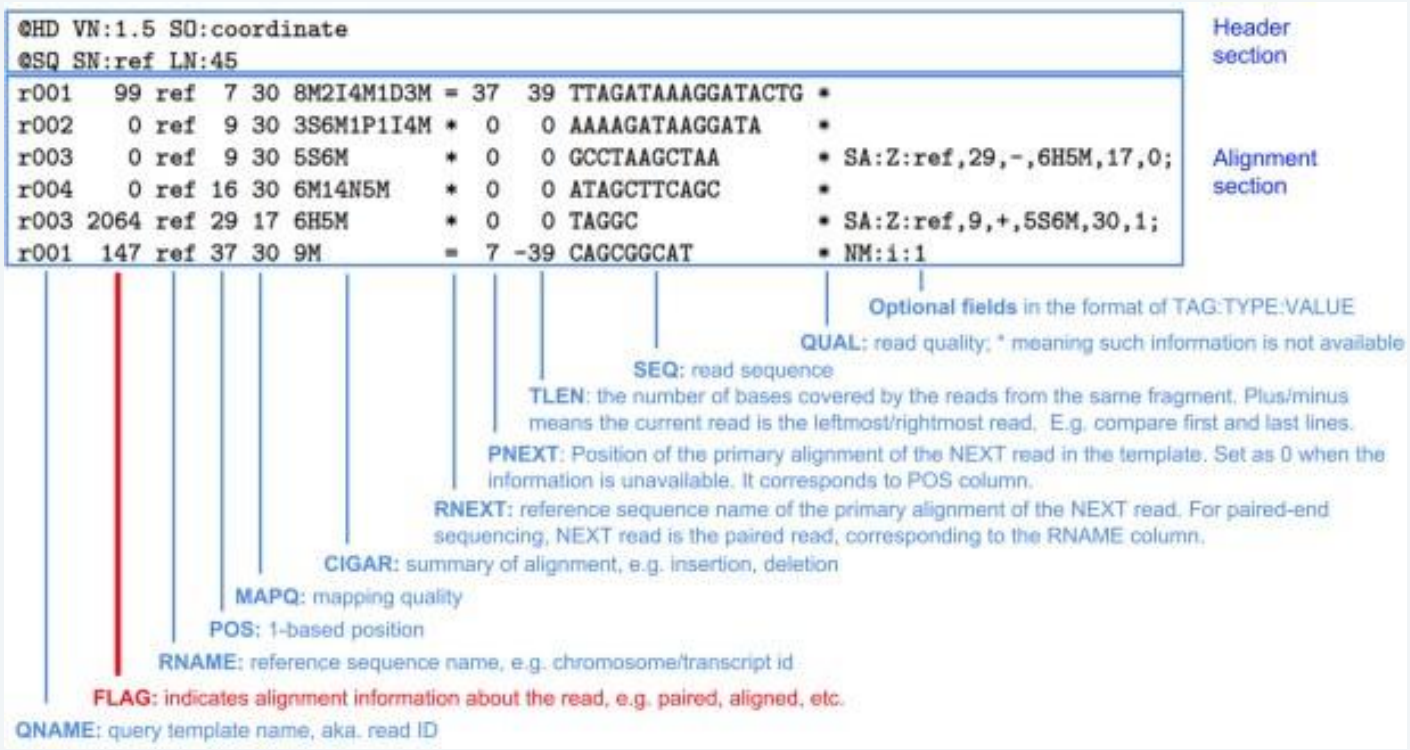
FORMATI NSG: SAM

Il formato SAM è un formato di testo delimitato da tabulati e costituito da due campi obbligatori:

- **Intestazione** (*Header*): contiene alcuni metadati (può occupare più righe)
- **Allineamento** (*Allignment*): informazioni sull’allineamento (può occupare più righe)

In generale, la qualità delle informazioni presenti 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

FORMATI NSG: SAM



FORMATI NSG: SAM

Il sotto campi SAM del campo Allineamento sono:

Colonna	Campo	Definizione	Tipo
1	**QNAME**	Nome	String
2	**FLAG**	Informazioni sull'allineamento tra reads (sono dei flag bit che assumo valore 0/1 se la proprietà non si verifica o se si verifica). Il valore va da 0 a 4095 https://broadinstitute.github.io/picard/explain-flags.html	Int
3	**RNAME**	Nome della reads	String
4	**POS**	Posizione di mappatura più a sinistra	Int
5	**MAPQ**	MAPping Quality	Int
6	**CIGAR**	Informazioni sull'allineamento (cancellazione, inserimento,..)	String
7	**RNEXT**	Nome della read/mate precedente/successive	String
8	**PNEXT**	Posizione della read/mate precedente/successive	Int
9	**TLEN**	Lunghezza delle basi che coincidono	Int
10	**SEQ**	La sequenza read	String
11	**QUAL**	Phred	String

FORMATI NSG: SAM

Il sotto campi SAM del campo Allineamento sono:

Colonna	Campo	Sottocampo	Significato
6	**CIGAR**		
		M	Corrispondenza (colonna di allineamento contenente due lettere). Potrebbe contenere due lettere diverse (mancata corrispondenza) o due lettere identiche
		D	Assenza
		Io	Inserimento
		S	Segmento che non viene visualizzato nell'allineamento
		H	Segmento della sequenza non è visualizzato nell'allineamento. I valori H specificano i segmenti all'inizio e/o alla fine della query che non sono visualizzati nel record SAM
		=	Colonna di allineamento contenente due lettere identiche.
		X	Colonna di allineamento contenente una mancata corrispondenza, ovvero due lettere diverse

FORMATI NSG: SAM

Il formato SAM garantisce di:

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 consente di restituire tali informazioni in breve tempo (millisecondi), senza dover leggere l'intero file

FORMATI NSG: SAM (esempio)

read1 99 chr1 10000 60 100M = 10100 200 ATCG... IIII...

read1: Nome della lettura.
99: FLAG (lettura allineata, paired-end).
chr1: Cromosoma di riferimento.
10000: Posizione iniziale.
60: Qualità dell'allineamento.
100M: CIGAR string (100 basi allineate perfettamente).
=: Mate si trova sullo stesso cromosoma.
10100: Posizione del mate
200: Lunghezza del template.
ATCG...: Sequenza allineata.
IIII...: Valori di qualità della sequenza.

FORMATI NSG: SAM

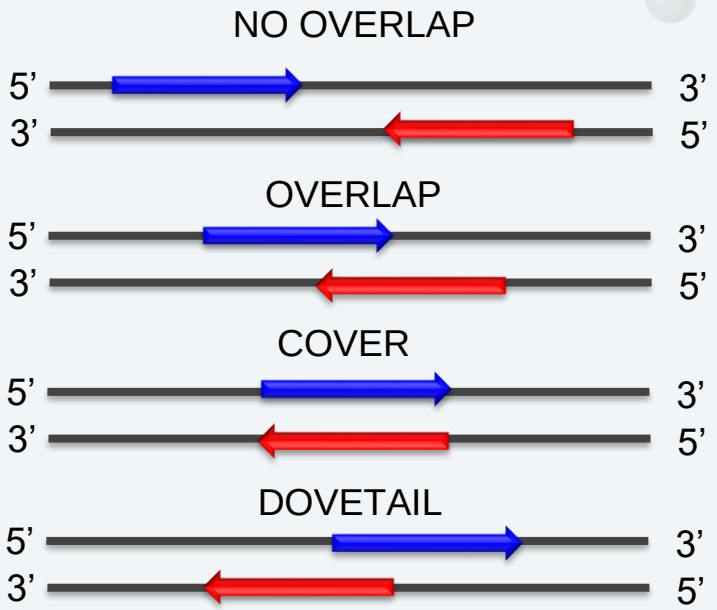
Topologia delle mappature dei mate (accoppiamenti di coppia):

(a) gli accoppiamenti si mappano su regioni distinte del genoma di riferimento, con l'accoppiamento diretto "a monte" dell'accoppiamento del complemento inverso (no overlap);

(b) le mappature di accoppiamento si sovrappongono in una o più posizioni di base (overlap);

(c) un accoppiamento mappa interamente all'interno della regione coperta dall'accoppiamento opposto (cover);

(d) l'accoppiamento in complemento inverso mappa "a monte" dell'accoppiamento in avanti (dovetail)



FORMATI NSG: SAM

Il formato SAM inoltre consente

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

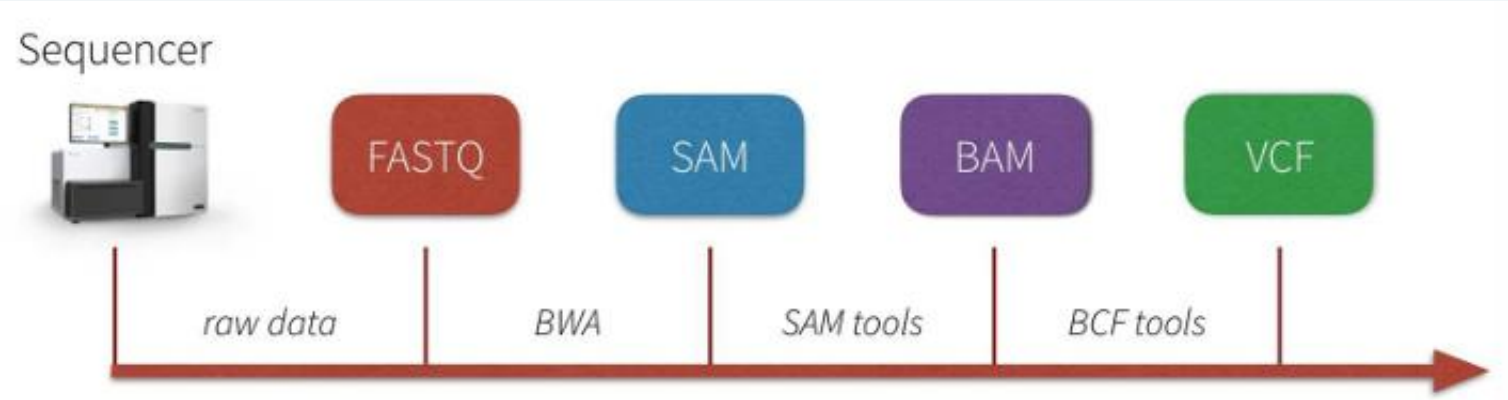
    Ad esempio, se si vuole essere in grado di selezionare rapidamente gli allineamenti che si allineano sul filo inverso

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

FORMATI NSG: BAM

Un file BAM è una rappresentazione binaria, compressa (e quasi sempre ordinata) delle informazioni SAM.
Generalmente, i file BAM sono ordinati in relazione alla coordinata di allineamento e più raramente in relazione ai nomi di lettura nel caso di coppie di lettura con nome identico
L'estensione è **.bam**



FORMATI NSG: dal SAM al BAM

Di solito si genera un file SAM, quindi si ordina il file e lo si converte in un formato BAM. Infine, si indicizza il file BAM risultante per svolgere operazioni di confronto più rapide.

Creare 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

FORMATI NSG: VCF

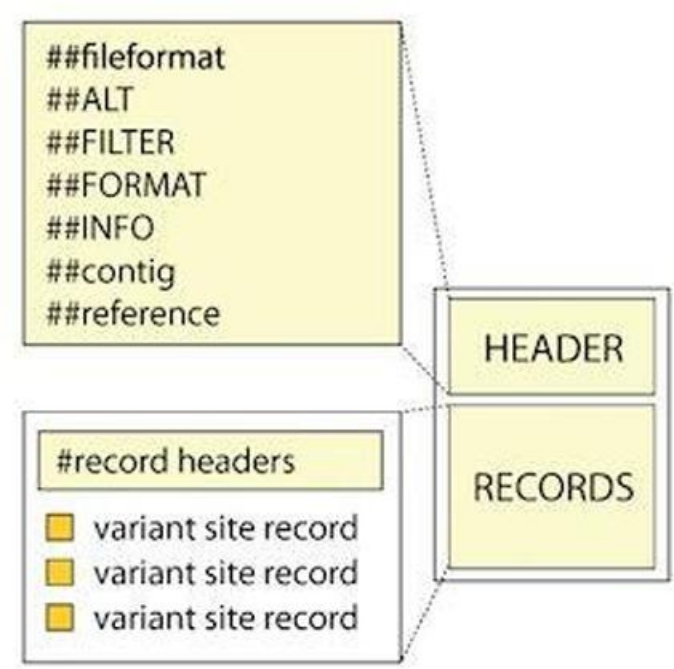
Il VCF (VariantCall Format) è un formato testuale che descrive la variazione degli allineamenti rispetto ad un riferimento.
Un file VCF è in genere creato da un file BAM (mentre il file BAM è stato creato da un file FASTQ e un file FASTA). Pertanto, il file VCF va considerato come un formato che cattura le differenze di ciascuna delle sequenze nel file FASTQ rispetto al genoma nel file FASTA
L'estensione è **.vcf**

```
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=(
```

FORMATI NSG: VCF

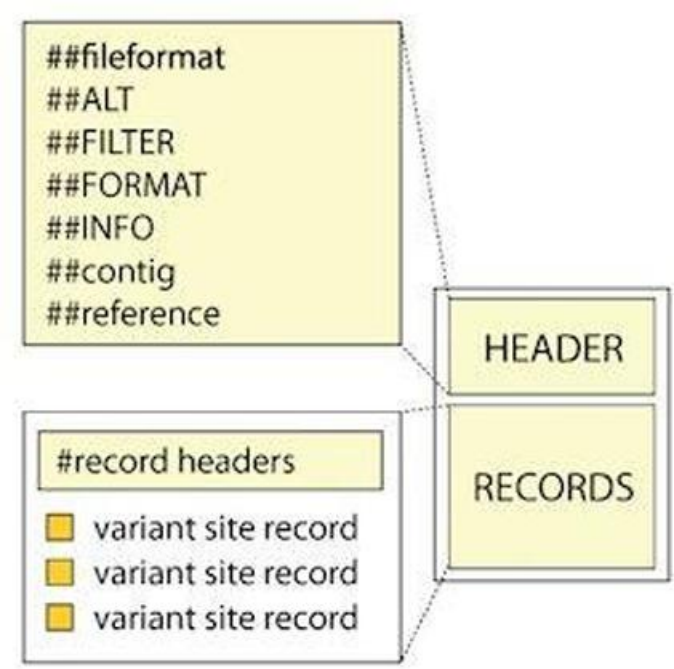
Un file VCF è composto da:

- **Intestazione** (*header*) mantiene informazioni su come sono strutturati i record
- **Record** (*record*) riportano le informazioni in forma tabellare

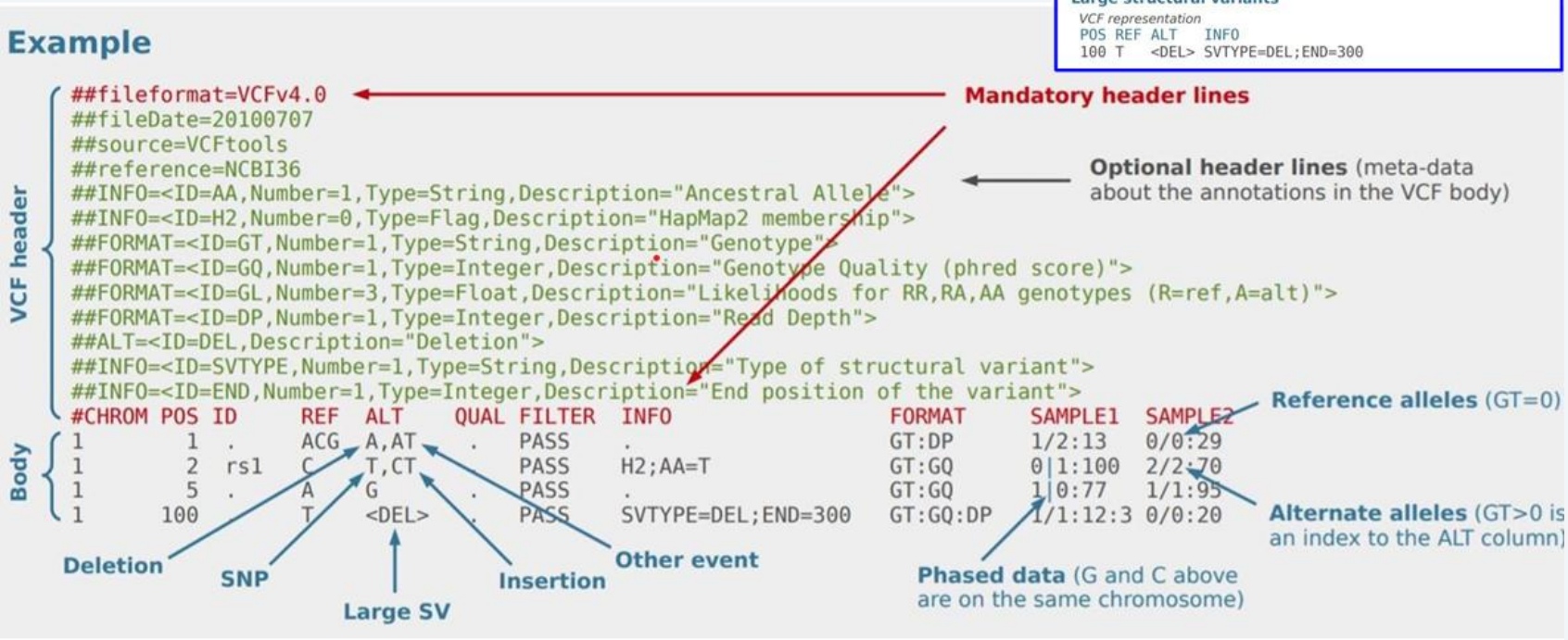


FORMATI NSG: VCF

I campi dei record sono delimitati da tabulazioni, dove le prime otto colonne descrivono una variante e le restanti colonne indicano le proprietà di ciascun campione. La nona colonna è il FORMATO e ciascuna colonna oltre alla nona rappresenta un campione.
Un file VCF può contenere un numero qualsiasi di colonne campione, anche migliaia, e può essere pensato come una tabella di database che rappresenta tutte le variazioni in tutti i campioni



FORMATI NSG: VCF



FORMATI NSG: VCF

Colonna	Campo	Definizione
1	**CHROM**	Cromosoma (o conting) su cui si verifica la variante
2	**POS**	Le coordinate genomiche su cui si verifica la variante
3	**ID**	identificatore per la variante (se esiste). In genere un database dbSNP se noto
4	**REF**	L'allele di riferimento sul filamento anteriore
5	**ALT**	Lo (Gli) 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

FORMATI NSG: VCF

```
##fileformat=VCFv4.2
##reference=GRCh38
##INFO=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
#CHROM  POS  ID  REF  ALT  QUAL  FILTER  INFO  FORMAT  SAMPLE1

chr1  123456  .  A  G  50  PASS  DP=100  GT  0/1
```

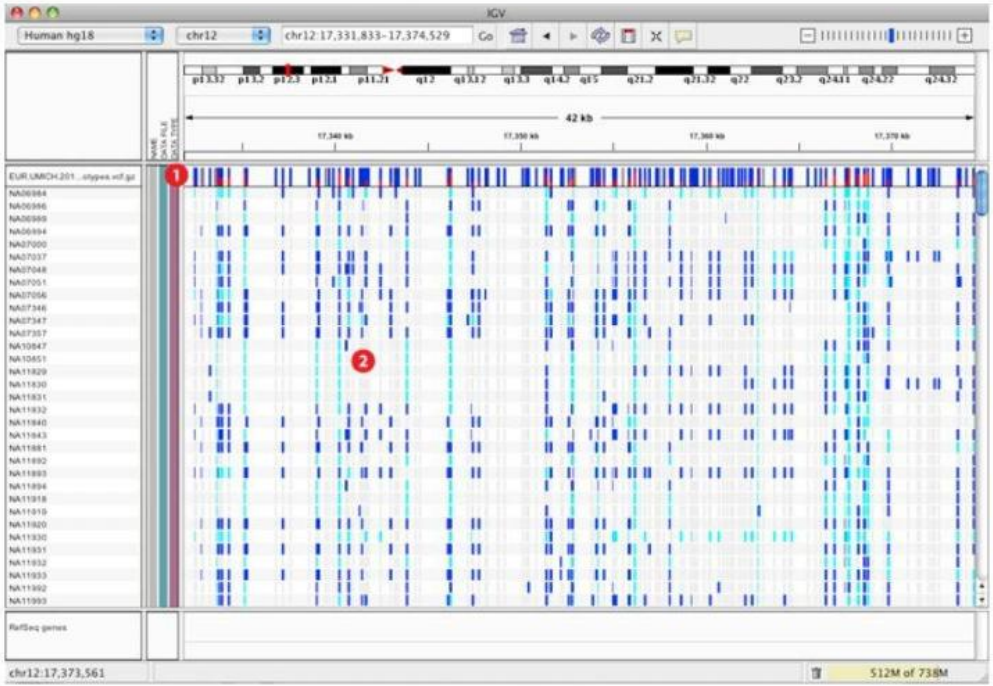
chr1: Cromosoma 1.
123456: Posizione 123456.
.: Nessun identificatore specifico per la variante.
A: Allele di riferimento.
G: Allele alternativo.
50: Qualità della variante (phred-scaled).
PASS: Variante ha passato i filtri.
DP=100: Profondità di lettura totale per questa posizione.
GT: Formato del genotipo.
0/1: Genotipo del campione (0: allele di riferimento, 1: allele alternativo).

INFO:
DP: profondità di lettura.
AF: frequenza dell'allele alternativo.
ANN: annotazioni funzionali (es. impatto della variante).

FORMAT:
GT: genotipo (0/0, 0/1, 1/1).
AD: profondità di lettura per ogni allele.
PL: probabilità genotipiche.

FORMATI NSG: VCF

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 delle informazioni in un preciso momento**.

Ai dati relativi al genoma umano sono stati adottati standard; per altri organismi, che hanno un ampio seguito, sono intervenute più organizzazioni ognuna proponendo uno standard.
Quando sono scoperte ulteriori informazioni, il formato va aggiornato, corretto e riorganizzato.
La genomic build è un esempio di dati di riferimento perché rappresentano una “edizione”, un’istantanea delle informazioni in un preciso momento

RIORGANIZZAZIONE DEL FORMATO

L’aggiornamento di un genomic build comporta spesso una modifica sostanziale.

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

RIORGANIZZAZIONE DEL FORMATO

Gli standard “specifici”, cioè quelli dedicati ad organismi generici, sono più dettagliati perché la quantità di dati accumulata è poca ed è più facile “sorvegliarla”
Soprattutto quando si ottengono dati da fonti disparate, è essenziale garantire che tutto si riferisca alle stesse informazioni di base.

iGENOMES

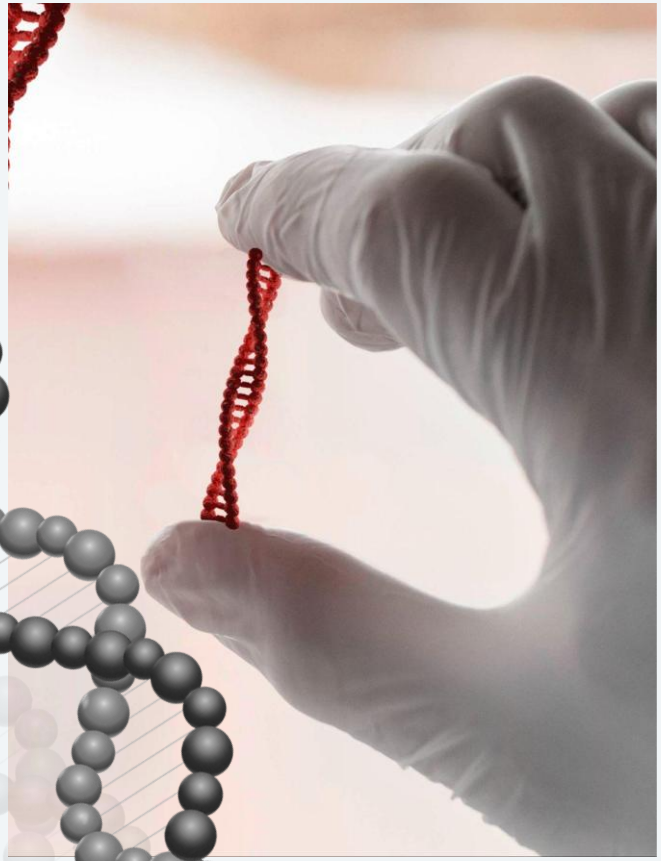
Gli iGenomes sono una raccolta di dati di riferimento e si file di annotazioni per organismi comunemente analizzati

Ogni iGenome è disponibile (da Ensembl, NCBI o UCSC) 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		



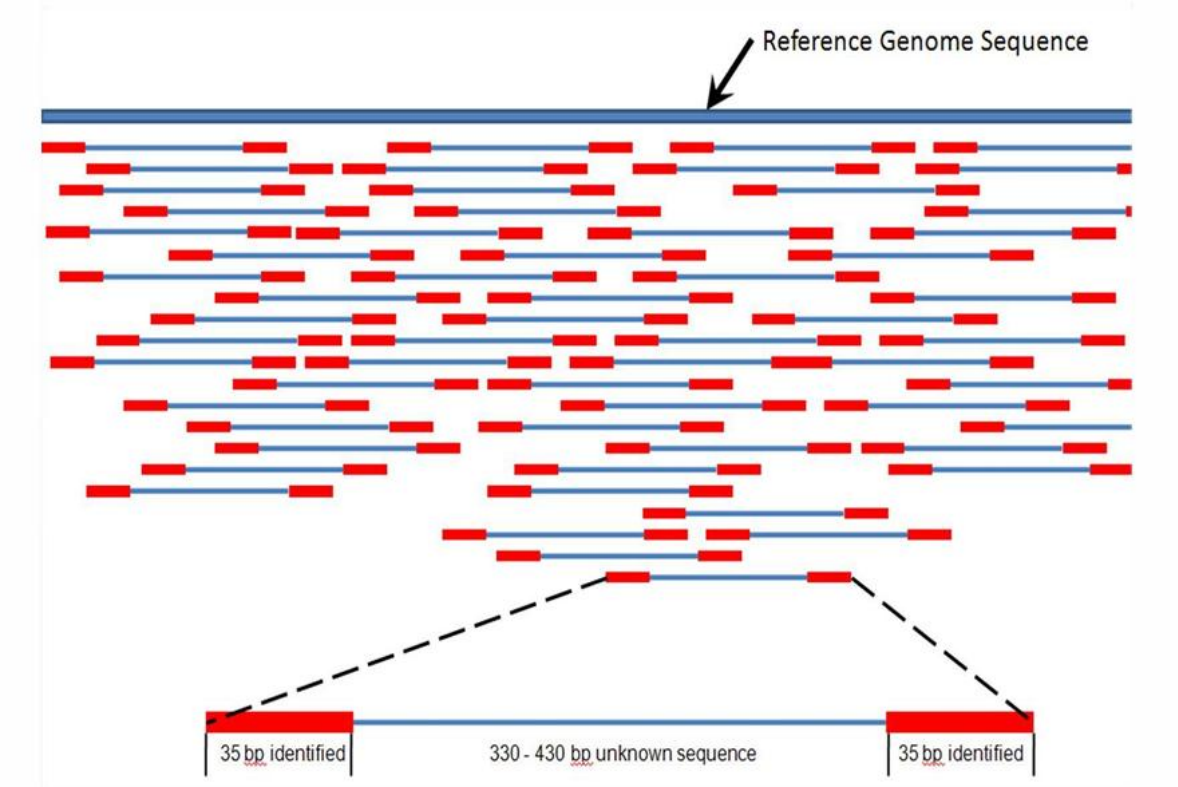
02

READS SINGLE e READS PAIRED

READS

I reads paired-end e i reads single-end sono due modalità di sequenziamento utilizzate per leggere frammenti di DNA o RNA in tecniche di sequenziamento di nuova generazione (NGS). Si differenziano principalmente per come vengono letti i frammenti e per l'utilizzo che ne consegue

READS


Reference Genome Sequence
35 bp identified
330 - 430 bp unknown sequence
35 bp identified

READS SINGLE-END (SE)

1. Reads Single-End (SE)

- **Cosa sono:**
  - In questa modalità, ogni frammento di DNA viene sequenziato a partire da un'unica estremità. Si ottiene quindi una sola sequenza per frammento.
- **Caratteristiche:**
  - Più semplice da eseguire.
  - Meno costoso rispetto al paired-end.
  - Genera una quantità inferiore di dati.
- **Quando si usa:**
  - Studi con budget limitato.
  - Analisi di regioni genomiche semplici, ad esempio in genomi batterici o per RNA-seq quando non è necessario individuare accoppiamenti.
  - Situazioni in cui si preferisce una pipeline di analisi semplice.

READS PAIRED-END (PE)

1. Reads Paired-End (PE)

- **Cosa sono:**
  - In questa modalità, entrambi gli estremi di ogni frammento di DNA sono sequenziati, ottenendo due letture complementari (forward e reverse) per ciascun frammento.
- **Caratteristiche:**
  - Maggiore accuratezza: avere due letture dello stesso frammento consente di correggere errori o incertezze di sequenziamento.
  - Permette una mappatura più precisa delle letture su un genoma di riferimento, soprattutto in regioni ripetitive.
  - Consente di stimare la lunghezza degli inserti (distanza tra le due letture).
- **Quando si usa:**
  - Studi su genomi complessi con regioni ripetitive; Assemblaggi genomici de novo; RNA-seq per individuare giunzioni di splicing o trascritti fusi e sequenziamento metagenomico, dove la diversità e la complessità sono elevate.

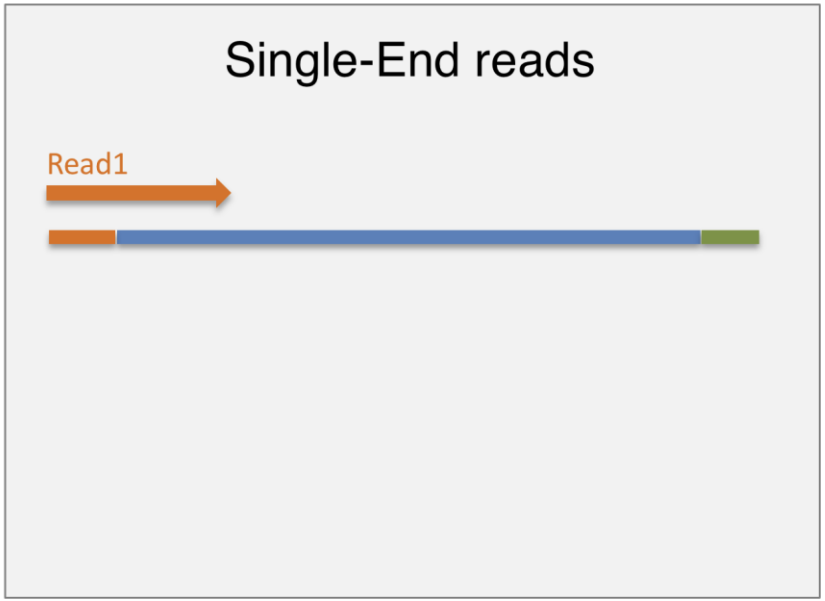
READS PAIRED-END e SINGLE-END

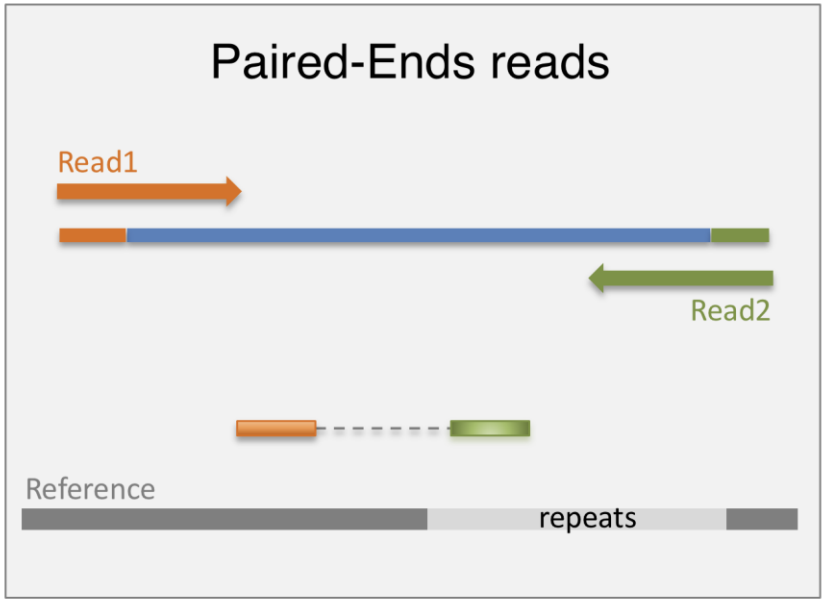
Nel protocollo il single-end (SE), o “prime” letture, è rappresentato da una singola lettura delle sequenze:
----> AAAATTTTGGGGCCCC
Nel protocollo paired-end viene eseguita una seconda misurazione per produrre un'altra serie di letture, cioè le “seconde” letture, che avrà andamento speculare alla prima.
<--- GGGGCCCCAAAATTTT

L'effetto finale è due misurazioni di 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

READS PAIRED-END e SINGLE-END





https://uclouvain-cbio.github.io/WSBIM2122/

Grazie!

Domande?

franco.liberati@unitus.it
deb.scienceontheweb.com