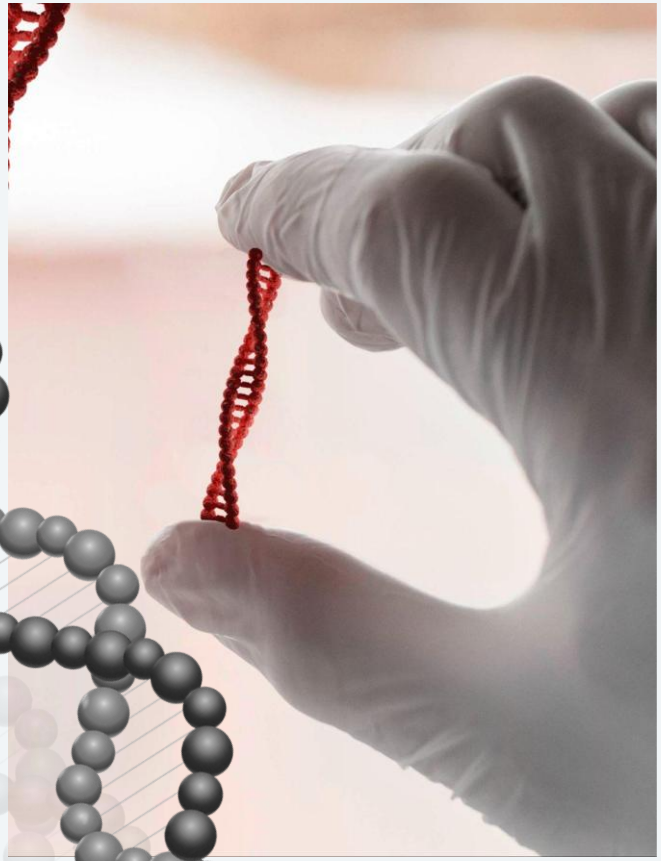
BIOINFORMATICA II

TRIMMING

ARGOMENTI

01 TRIMMING



01

TRIMMING

TRIMMING: Definizione

Il *trimming* è il processo di rimozione delle sequenze di bassa qualità o delle porzioni di una read che non sono rilevanti per l'analisi, come gli adattatori o le basi di qualità scarsa alle estremità.
Il trimming migliora la qualità dei dati che sono usati per analisi successive, come il mapping al genoma di riferimento o l'assemblaggio *de novo*.

TRIMMING: sequenze da eliminare

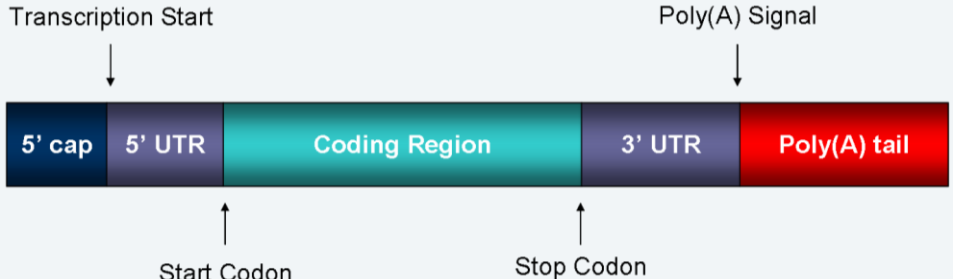
Le reads grezze prodotte dal sequenziamento possono contenere **sequenze indesiderate**, come:

- **Adattatori**: sequenze artificiali usate durante la preparazione della libreria.
- **Basi a bassa qualità**: nucleotidi con alta probabilità di errore di chiamata.
- **Poli-A o Poli-T tails**: le code omopolimeriche, che possono confondere l’analisi successiva.

TRIMMING: approfondimento PolyA/PolyT

Le poli-A e le poli-T sono sequenze di basi nucleotidiche specifiche che si trovano nel DNA o nell'RNA :

- **Poli-A (Poly-A):**
  Una sequenza costituita da molte basi adenina (A) consecutive. In RNA messaggero (mRNA) degli eucarioti, una coda poli-A viene aggiunta all'estremità 3' durante la maturazione dell'mRNA.
- **Poli-T (Poly-T):**
  Una sequenza costituita da molte basi timina (T) consecutive. Nel DNA, le sequenze di poli-T possono apparire come complementari alle regioni poli-A su un filamento opposto.
  Sono spesso usate in laboratorio, ad esempio: Nelle colonne di purificazione di mRNA, dove oligonucleotidi di poli-T legano la coda poli-A dell'mRNA per separarlo dagli altri tipi di RNA.

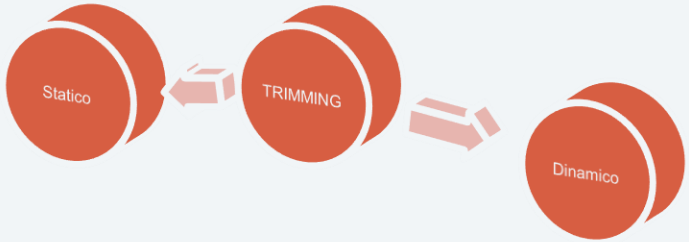


https://2013.igem.org/Team:Berkeley/Methods/Protocols

TRIMMING: Tipologia di elaborazione

Rimuovere queste sequenze migliora l'accuratezza delle analisi bioinformatiche, riducendo gli errori durante il mapping e l'interpretazione dei dati.

Il trimming può essere statico o dinamico

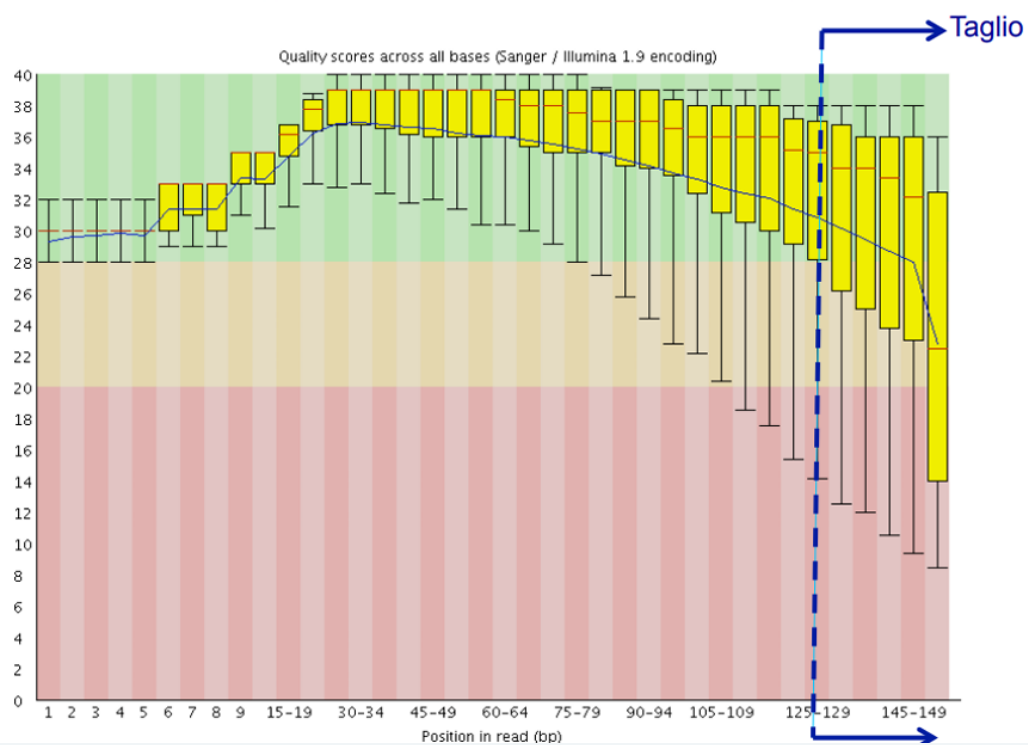


TRIMMING: Statico

Trimming statico:

- Nel trimming statico, le reads sono tagliate seguendo regole predefinite. Questo significa che sono rimosse sempre le stesse posizioni della sequenza, indipendentemente dalla qualità specifica di quella porzione.
  Ad esempio, si può decidere di rimuovere sempre i primi 10 nucleotidi e gli ultimi 5, oppure eliminare una porzione fissa identificata come adattatore.
- È poco flessibile perché non considera la variabilità della qualità della sequenza e potrebbe rimuovere informazioni utili o lasciare parti di bassa qualità se queste non si trovano nelle posizioni preimpostate.

TRIMMING: statico (esempio)



TRIMMING: Dinamico

Trimming dinamico:

- Il trimming dinamico adatta il taglio delle reads in relazione alla qualità effettiva di ciascuna lettura. Generalmente, un algoritmo valuta ogni posizione della sequenza e taglia la read quando rileva che la qualità scende sotto una certa soglia.
  Ad esempio, può eliminare le basi dalla fine di una read fino a quando la qualità non raggiunge un livello accettabile, oppure rimuovere adattatori rilevati sulla base di algoritmi di allineamento specifici.
- È più preciso e adattabile, perché analizza ogni sequenza individualmente, garantendo che solo le porzioni effettivamente di bassa qualità vengano rimosse, senza perdere informazioni utili.

TRIMMING: Dinamico (esempio)

Adattatore

READS
pessima qualità

READS
scarsa qualità

Adattatore

TRIMMING: confronto

Caratteristica	Trimming Statico	Trimming Dinamico
Flessibilità	Fisso, indipendente dalla qualità	Adattivo, dipende dalla qualità della read
Precisione	Potenzialmente meno accurato	Maggiore precisione
Efficienza computazionale	Più rapido e meno dispendioso	Più complesso e dispendioso
Rischio di perdita dati	Rischio di rimuovere basi utili	Mantiene solo le porzioni necessarie
Uso comune	Rimozione di adattatori specifici	Controllo qualità per rimuovere basi degradate

In sintesi, il trimming statico è utile quando si conoscono esattamente le porzioni da tagliare e si vuole un approccio rapido e standard, mentre il trimming dinamico è preferibile quando si desidera ottenere la massima qualità possibile dalle reads, poiché permette di adattarsi alle variazioni intrinseche di ciascuna sequenza.

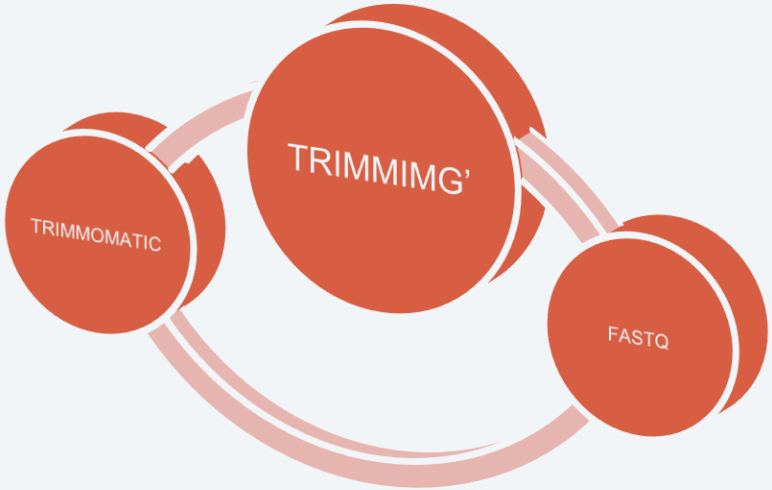
BIOINFORMATICA

TRIMMING

Il "trimming" in bioinformatica si riferisce al processo di rimozione di basi indesiderate o di bassa qualità dalle estremità delle sequenze nei dati di sequenziamento, come quelli nel formato FASTQ.
Può essere personalizzato in relazione alle specifiche esigenze dell'analisi.
Ad esempio, è possibile impostare soglie di qualità per definire quali basi devono essere tagliate in base ai loro punteggi di qualità Phred.

In generale, il trimming è una fase cruciale nel pre-processing dei dati di sequenziamento per garantire che le sequenze siano pulite, accurate e pronte per le successive analisi di bioinformatica

TRIMMING



TRIMMOMATIC

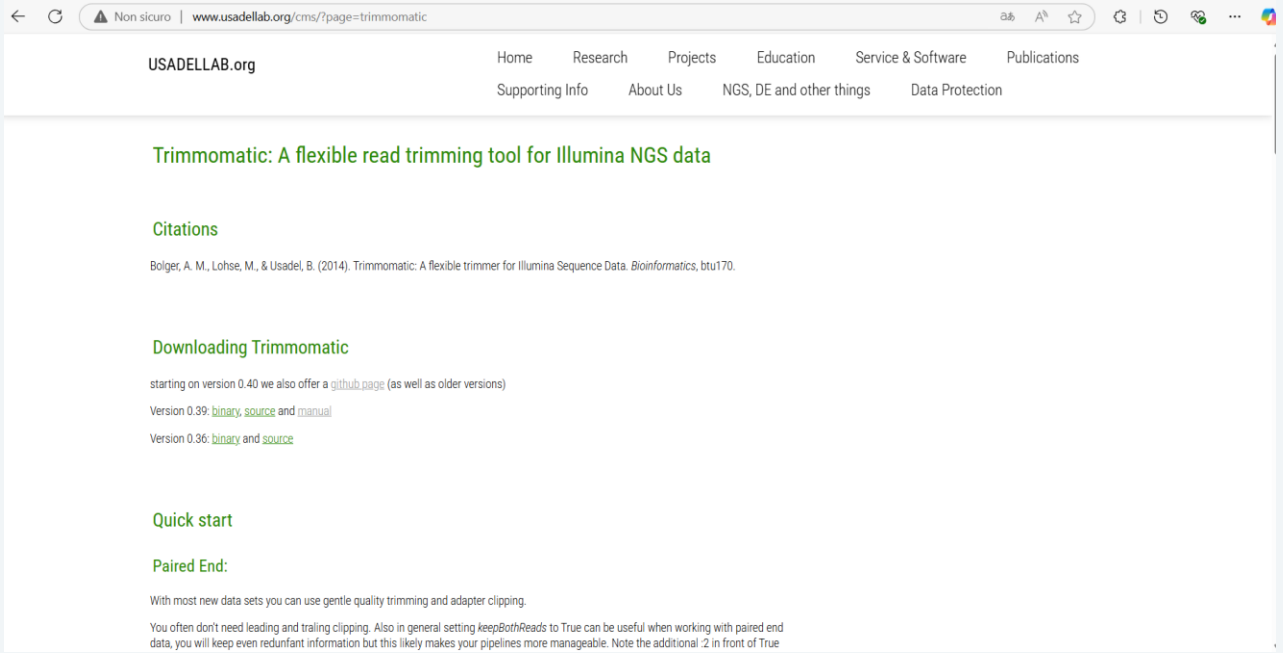
Trimmomatic è uno software di bioinformatica utilizzato per il trimming delle sequenze nei dati di sequenziamento.

È progettato per rimuovere basi di bassa qualità, adattatori e sequenze di scarsa qualità dalle estremità delle read nei file FASTQ.

L'uso di Trimmomatic è una fase comune nella pipeline di analisi dei dati di sequenziamento, poiché contribuisce a migliorare la qualità delle sequenze e a ridurre il rumore nei dati. La sua applicazione è particolarmente importante per garantire risultati accurati nelle successive fasi di assemblaggio, allineamento e analisi dei dati di sequenziamento.

TRIMMOMATIC: Sito

http://www.usadellab.org/cms/?page=trimmomatic



TRIMMOMATIC: Installazione

È sufficiente il download e la decompressione del file

http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/Trimmomatic-0.39.zip

Dipendenze:
Richiede la Java Virtual Machine

TRIMMOMATIC: Comando (Single-End)

Single End (SE):

java -jar trimmomatic-0.39.jar SE -*phredtype file.fastq file_trimmed.fastq* ILLUMINACLIP:*adapters.fa*:*val1*:*val2*:*val3* LEADING:*val* TRAILING:*val* SLIDINGWINDOW:*val1*:*val2* MINLEN:*val1*

TRIMMOMATIC: Comando (esempio)

java -jar trimmomatic-0.39.jar SE -phred33 input.fastq output_trimmed.fastq ILLUMINACLIP:adapters.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36

OPZIONE	Descrizione
SE	modalità single-end (per dati derivanti da un solo file FASTQ)
-phred33	specifica il formato della qualità (es. -phred33 o -phred64)
input.fastq	file di input delle sequenze non elaborate
output_trimmed.fastq	file di output delle sequenze pulite
ILLUMINACLIP adapters.fa:2:30:10	adapters.fa: file contenente le sequenze degli adattatori. 2: numero massimo di mismatch tra adattatore e sequenza. 30: soglia di punteggio per decidere il taglio. 10: lunghezza minima della sequenza di adattatore da considerare.
LEADING:3	rimuove le basi con qualità inferiore a 3 dall'inizio
TRAILING:3	rimuove le basi con qualità inferiore a 3 dalla fine
SLIDINGWINDOW:4:15	analizza le finestre mobili di 4 basi e rimuove se la qualità media è inferiore a 15
MINLEN:36	rimuove sequenze più corte di 36 basi

TRIMMOMATIC: Comando (Paired-End)

Paired End (SE):

```
java -jar trimmomatic-0.39.jar PE -phredType input_forward.fastq
input_reverse.fastq
output_forward_paired.fastq output_forward_unpaired.fastq
output_reverse_paired.fastq output_reverse_unpaired.fastq
ILLUMINACLIP:adapters.fa:val1:val2:val3
LEADING:val1
TRAILING:val1
SLIDINGWINDOW:val1:val2
MINLEN:val1
```

TRIMMOMATIC: Comando (esempio)

java -jar trimmomatic-0.39.jar PE -phred33 input_forward.fastq input_reverse.fastq output_forward_paired.fastq output_forward_unpaired.fastq output_reverse_paired.fastq output_reverse_unpaired.fastq ILLUMINACLIP:adapters.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36

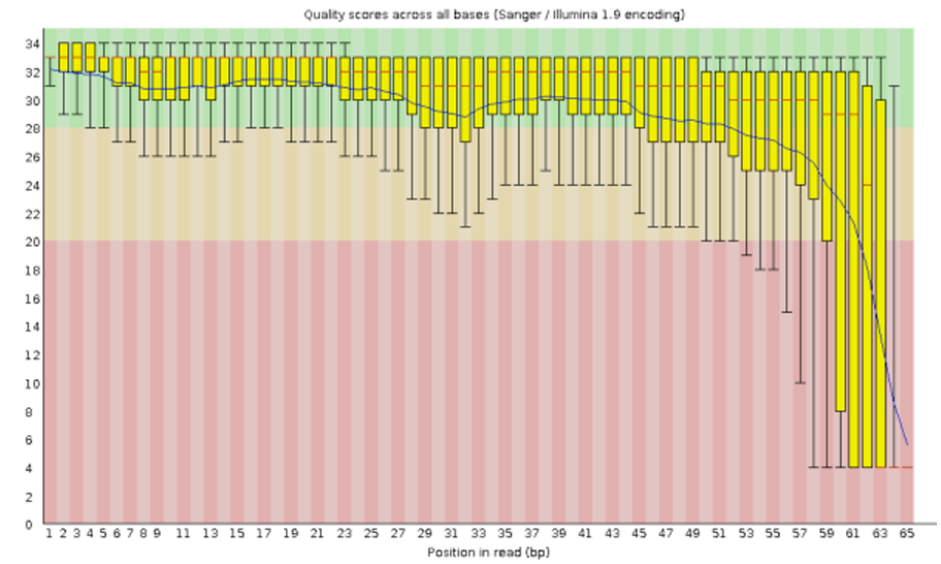
Parametro	Descrizione
SE	modalità paired-end (dati provenienti da due file FASTQ, uno per ciascun lato della lettura)
-phred33	specifica il formato della qualità (es. -phred33 o -phred64)
input_forward.fastq input_reverse.fastq	file di input per le letture forward e reverse
output_forward_paired.fastq, output_reverse_paired.fastq	file di output per le letture forward e reverse che rimangono appaiate dopo il trimming
output_forward_unpaired.fastq, output_reverse_unpaired.fastq	file di output per le letture forward e reverse che sono state scartate perché non appaiate
ILLUMINACLIP adapters.fa:2:30:10	adapters.fa: file contenente le sequenze degli adattatori. 2: numero massimo di mismatch tra adattatore e sequenza. 30: soglia di punteggio per decidere il taglio. 10: lunghezza minima della sequenza di adattatore da considerare.
LEADING:3	rimuove le basi con qualità inferiore a 3 dall'inizio
TRAILING:3	rimuove le basi con qualità inferiore a 3 dalla fine
SLIDINGWINDOW:4:15	analizza le finestre mobili di 4 basi e rimuove se la qualità media è inferiore a 15
MINLEN:36	rimuove sequenze più corte di 36 basi

TRIMMOMATIC: Altri parametri

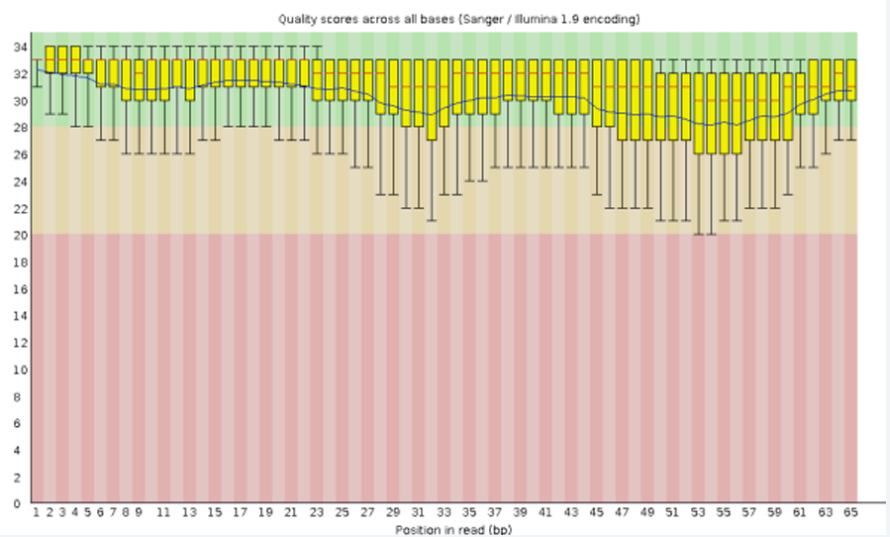
Parametro	Significato
HEADCROP	Rimuove un numero specificato di basi dall'inizio di ogni lettura
CROP	Taglia la lettura a una lunghezza specificata rimuovendo le basi in eccesso dalla fine
TOPHRED33/TOPHRED64	Converte la qualità dei punteggi dal formato Phred+64 al formato Phred+33 e viceversa
MAXINFO	Opzione di trimming adattivo che bilancia la lunghezza della lettura e la qualità per determinare la migliore lunghezza di trimming.
AVGQUAL	Rimuove la lettura se la qualità media è al di sotto di una soglia specificata

TRIMMOMATIC

PRIMA DEL TRIMMING



DOPO IL TRIMMING



FASTQ

FASTQ è un software di bioinformatica ampiamente utilizzato per il trimming delle sequenze nei dati di sequenziamento

Opera effettuando una profilazione completa della qualità sia prima sia dopo il filtraggio dei dati (curve di qualità, contenuti di base, KMER, Q20/Q30, GC Ratio, duplicazione, contenuti dell'adattatore...)

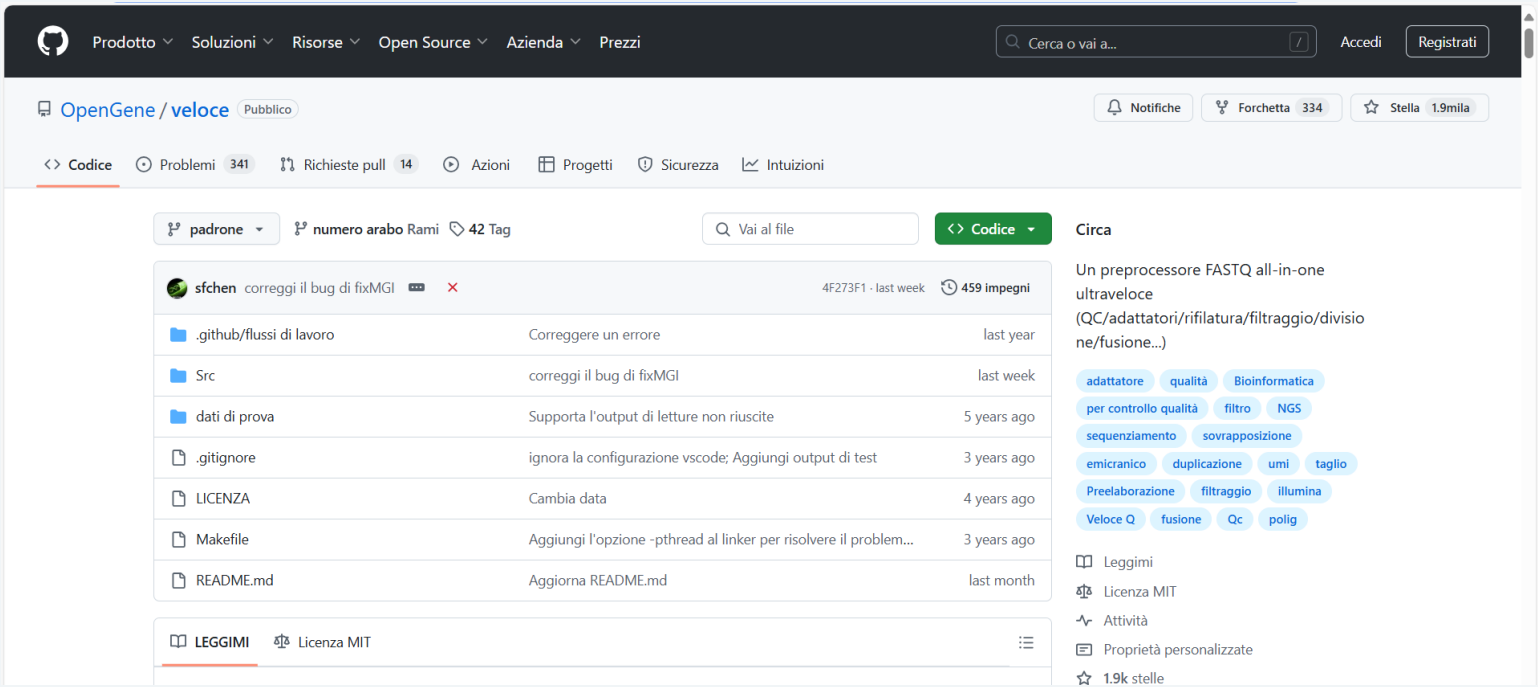
FASTQ

FASTQ opera nella seguente maniera:

- filtra le reads errate: qualità troppo bassa, troppo breve o troppi N...
- tagliare basi di bassa qualità valutando la qualità media considerando una finestra scorrevole (come Trimmomatic).
- Taglia tutte le reads davanti e in coda e taglia gli adattatori.
- Corregge le coppie di basi non corrispondenti nelle regioni sovrapposte delle letture finali accoppiate, se una base è di alta qualità mentre l'altra è di qualità ultra bassa
- Toglie le trim polyG in 3' ends, (spesso nei dati ricavati da NovaSeq/NextSeq)
- Taglio le polyX in estremità 3'
- riporta il risultato in formato JSON per svolgere analisi di approfondimento
- visualizza i risultati del controllo qualità e del filtraggio su una singola pagina HTML (come FASTQC).
- Consente l'elaborazione parallela.

FASTQ: Sito

https://github.com/OpenGene/fastp



FASTQ: Installazione

È sufficiente il download del file eseguibile:

```
wget http://opengene.org/fastp/fastp
chmod a+x ./fastp
```

Se si vuole una versione specifica:

```
wget http://opengene.org/fastp/fastp.0.23.4
mv fastp.0.23.4 fastp
chmod a+x ./fastp
```

Richiede:

libdeflate e *libisal* presenti nel compilatore GCC

FASTQ: Comando

Per le reads Single-end

fastp -i in.fq -o out.fq

Per le reads Paired-end

fastp -i in.R1.fq -I in.R2.fq -o out.R1.fq -O out.R2.fq

Opzione	Significato
-i	File FASTQ da trimmare
-o	Nome di uscita del file trimmato

FASTQ: Esempi di report

fastp report

Summary

General

fastp version:	0.17.0 (https://github.com/OpenGene/fastp)
sequencing:	paired end (151 cycles + 151 cycles)
mean length before filtering:	108bp, 108bp
mean length after filtering:	107bp, 107bp
duplication rate:	30.641418%
Insert size peak:	95

Before filtering

total reads:	16.763944 M
total bases:	1.818801 G
Q20 bases:	1.716550 G (94.378124%)
Q30 bases:	1.672955 G (91.981195%)
GC content:	47.006320%

After filtering

total reads:	16.034314 M
total bases:	1.722358 G
Q20 bases:	1.659759 G (96.365462%)
Q30 bases:	1.622287 G (94.189832%)
GC content:	46.794079%

Filtering result

reads passed filters:	16.034314 M (95.647623%)
reads with low quality:	695.022000 K (4.145934%)
reads with too many N:	14.740000 K (0.087927%)
reads too short:	19.868000 K (0.118516%)

FASTQ: Esempi di report

Adapters

Adapter or bad ligation of read1

The input has little adapter percentage (~0.039878%), probably it's trimmed before.

Sequence	Occurrences
N	8579
AN	7450
AGN	6492
AGAN	4593
AGNN	374
other adapter sequences	5774

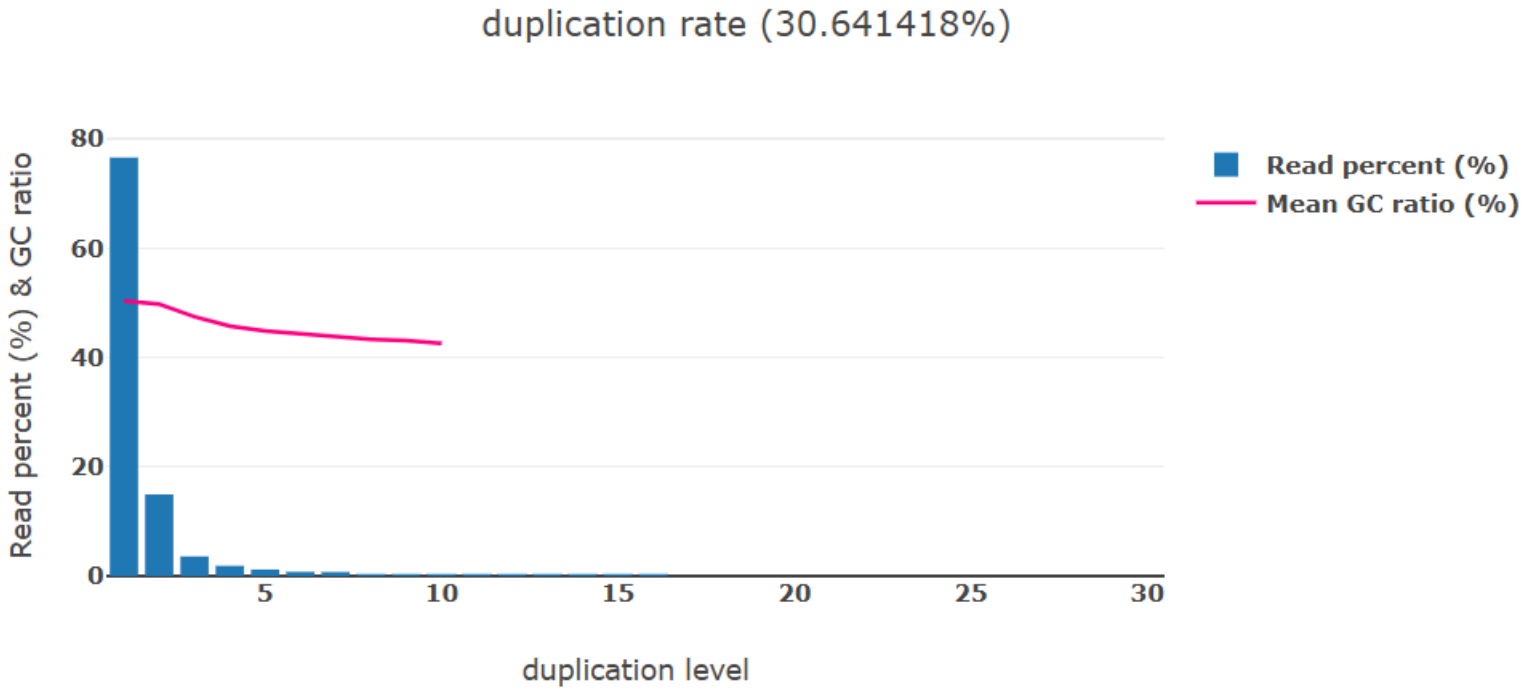
Adapter or bad ligation of read2

The input has little adapter percentage (~0.226118%), probably it's trimmed before.

Sequence	Occurrences
N	8529
AN	7402
AGA	1480
AGN	6429
AGAN	4663
other adapter sequences	45383

FASTQ: Esempi di report

Duplication

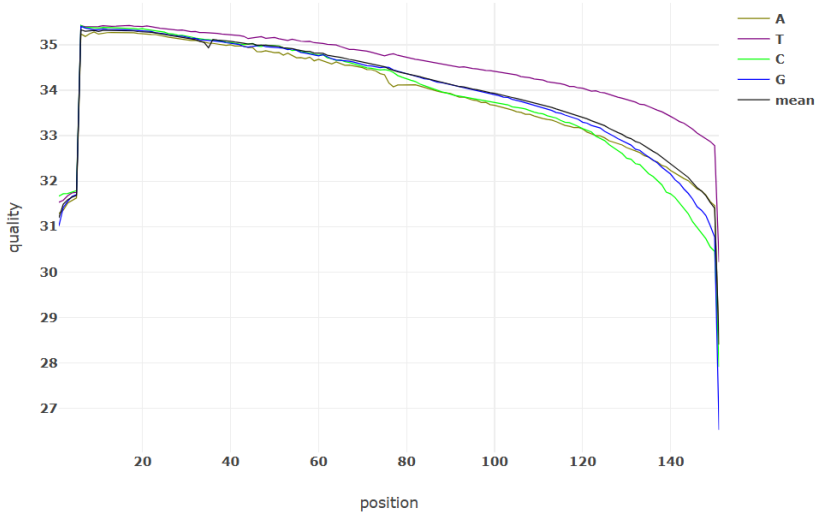


FASTQ: Esempi di report

Before filtering

Before filtering: read1: quality

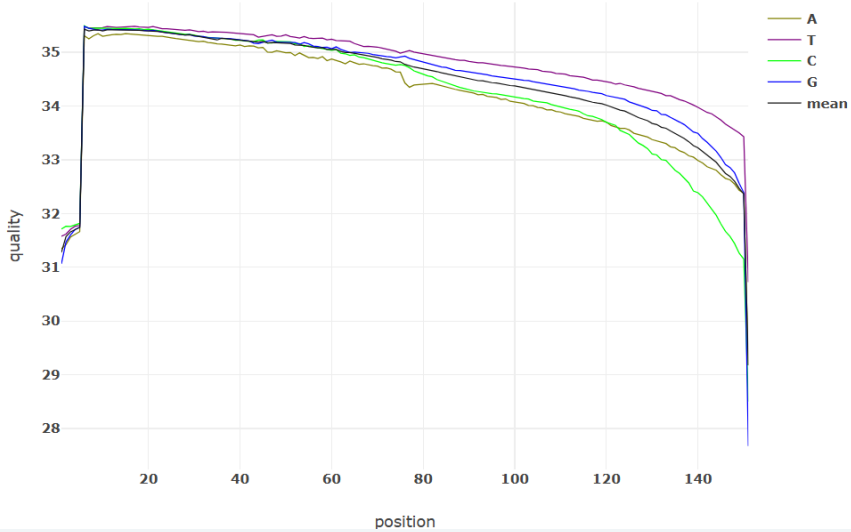
Value of each position will be shown on mouse over.



After filtering

After filtering: read1: quality

Value of each position will be shown on mouse over.



FASTQ: Esempi di report

Before filtering: read2: KMER counting

Darker background means larger counts. The count will be shown on mouse over.

	AA	AT	AC	AG	TA	TT	TC	TG	CA	CT	CC	CG	GA	GT	GC	GG
AAA	AAAAA	AAAAT	AAAAC	AAAAG	AAATA	AAATT	AAATC	AAATG	AAACA	AAACT	AAACC	AAACG	AAAGA	AAAGT	AAAGC	AAAGG
AAT	AATAA	AATAT	AATAC	AATAG	AATTA	AATTT	AATTC	AATTG	AATCA	AATCT	AATCC	AATCG	AATGA	AATGT	AATGC	AATGG
AAC	AACAA	AACAT	AACAC	AACAG	AACTA	AACTT	AACTC	AACTG	AACCA	AACCT	AACCC	AACCG	AACGA	AACGT	AACGC	AACGG
AAG	AAGAA	AAGAT	AAGAC	AAGAG	AAGTA	AAGTT	AAGTC	AAGTG	AAGCA	AAGCT	AAGCC	AAGCG	AAGGA	AAGGT	AAGGC	AAGGG
ATA	ATAAA	ATAAT	ATAAC	ATAAG	ATATA	ATATT	ATATC	ATATG	ATACA	ATACT	ATACC	ATACG	ATAGA	ATAGT	ATAGC	ATAGG
ATT	ATTAA	ATTAT	ATTAC	ATTAG	ATTTA	ATTTT	ATTTC	ATTTG	ATTCA	ATTCT	ATTCC	ATTCG	ATTGA	ATTGT	ATTGC	ATTGG
ATC	ATCAA	ATCAT	ATCAC	ATCAG	ATCTA	ATCTT	ATCTC	ATCTG	ATCCA	ATCCT	ATCCC	ATCCG	ATCGA	ATCGT	ATCGC	ATCGG
ATG	ATGAA	ATGAT	ATGAC	ATGAG	ATGTA	ATGTT	ATGTC	ATGTG	ATGCA	ATGCT	ATGCC	ATGCG	ATGGA	ATGGT	ATGGC	ATGGG
ACA	ACAAA	ACAAT	ACAAC	ACAAG	ACATA	ACATT	ACATC	ACATG	ACACA	ACACT	ACACC	ACACG	ACAGA	ACAGT	ACAGC	ACAGG
ACT	ACTAA	ACTAT	ACTAC	ACTAG	ACTTA	ACTTT	ACTTC	ACTTG	ACTCA	ACTCT	ACTCC	ACTCG	ACTGA	ACTGT	ACTGC	ACTGG
ACC	ACCAA	ACCAT	ACCAC	ACCAG	ACCTA	ACCTT	ACCTC	ACCTG	ACCCA	ACCCT	ACCCC	ACCCG	ACCGA	ACCGT	ACCGC	ACCGG
ACG	ACGAA	ACGAT	ACGAC	ACGAG	ACGTA	ACGTT	ACGTC	ACGTG	ACGCA	ACGCT	ACGCC	ACGCG	ACGGA	ACGGT	ACGGC	ACGGG
AGA	AGAAA	AGAAT	AGAAC	AGAAG	AGATA	AGATT	AGATC	AGATG	AGACA	AGACT	AGACC	AGACG	AGAGA	AGAGT	AGAGC	AGAGG
AGT	AGTAA	AGTAT	AGTAC	AGTAG	AGTTA	AGTTT	AGTTC	AGTTG	AGTCA	AGTCT	AGTCC	AGTCG	AGTGA	AGTGT	AGTGC	AGTGG
AGC	AGCAA	AGCAT	AGCAC	AGCAG	AGCTA	AGCTT	AGCTC	AGCTG	AGCCA	AGCCT	AGCCC	AGCCG	AGCGA	AGCGT	AGCGC	AGCGG
AGG	AGGAA	AGGAT	AGGAC	AGGAG	AGGTA	AGGTT	AGGTC	AGGTG	AGGCA	AGGCT	AGGCC	AGGCG	AGGGA	AGGGT	AGGGC	AGGGG
TAA	TAAAA	TAAAT	TAAAC	TAAAG	TAATA	TAATT	TAATC	TAATG	TAACA	TAACT	TAACC	TAACG	TAAGA	TAAGT	TAAGC	TAAGG
TAT	TATAA	TATAT	TATAC	TATAG	TATTA	TATTT	TATTC	TATTG	TATCA	TATCT	TATCC	TATCG	TATGA	TATGT	TATGC	TATGG
TAC	TACAA	TACAT	TACAC	TACAG	TACTA	TACTT	TACTC	TACTG	TACCA	TACCT	TACCC	TACCG	TACGA	TACGT	TACGC	TACGG
TAG	TAGAA	TAGAT	TAGAC	TAGAG	TAGTA	TAGTT	TAGTC	TAGTG	TAGCA	TAGCT	TAGCC	TAGCG	TAGGA	TAGGT	TAGGC	TAGGG
TTA	TTAAA	TTAAT	TTAAC	TTAAG	TTATA	TTATT	TTATC	TTATG	TTACA	TTACT	TTACC	TTACG	TTAGA	TTAGT	TTAGC	TTAGG
TTT	TTTAA	TTTAT	TTTAC	TTTAG	TTTTA	TTTTT	TTTTC	TTTTG	TTTCA	TTTCT	TTTCC	TTTCG	TTTGA	TTTGT	TTTGC	TTTGG
TTC	TTCAA	TTCAT	TTCAC	TTCAG	TTCTA	TTCTT	TTCTC	TTCTG	TTCCA	TTCCT	TTCCC	TTCCG	TTCGA	TTCGT	TTCGC	TTCGG
TTG	TTGAA	TTGAT	TTGAC	TTGAG	TTGTA	TTGTT	TTGTC	TTGTG	TTGCA	TTGCT	TTGCC	TTGCG	TTGGA	TTGGT	TTGGC	TTGGG
TCA	TCAAA	TCAAT	TCAAC	TCAAG	TCATA	TCATT	TCATC	TCATG	TCACA	TCACT	TCACC	TCACG	TCAGA	TCAGT	TCAGC	TCAGG
TCT	TCTAA	TCTAT	TCTAC	TCTAG	TCTTA	TCTTT	TCTTC	TCTTG	TCTCA	TCTCT	TCTCC	TCTCG	TCTGA	TCTGT	TCTGC	TCTGG
TCC	TCCAA	TCCAT	TCCAC	TCCAG	TCCTA	TCCTT	TCCTC	TCCTG	TCCCA	TCCCT	TCCCC	TCCCG	TCCGA	TCCGT	TCCGC	TCCGG
TCG	TCGAA	TCGAT	TCGAC	TCGAG	TCGTA	TCGTT	TCGTC	TCGTG	TCGCA	TCGCT	TCGCC	TCGCG	TCGGA	TCGGT	TCGGC	TCGGG
TGA	TGAAA	TGAAT	TGAAC	TGAAG	TGATA	TGATT	TGATC	TGATG	TGACA	TGACT	TGACC	TGACG	TGAGA	TGAGT	TGAGC	TGAGG
TGT	TGTAA	TGTAT	TGTAC	TGTAG	TGTTA	TGTTT	TGTTC	TGTTG	TGTCA	TGTCT	TGTCC	TGTCG	TGTGA	TGTGT	TGTGC	TGTGG
TGC	TGCAA	TGCAT	TGCAC	TGCAG	TGCTA	TGCTT	TGCTC	TGCTG	TGCCA	TGCCT	TGCCC	TGCCG	TGCGA	TGCGT	TGCGC	TGCGG
TGG	TGGAA	TGGAT	TGGAC	TGGAG	TGGTA	TGGTT	TGGTC	TGGTG	TGGCA	TGGCT	TGGCC	TGGCG	TGGGA	TGGGT	TGGGC	TGGGG
CAA	CAAAA	CAAAT	CAAAC	CAAAG	CAATA	CAATT	CAATC	CAATG	CAACA	CAACT	CAACC	CAACG	CAAGA	CAAGT	CAAGC	CAAGG
CAT	CATAA	CATAT	CATAC	CATAG	CATTA	CATTT	CATTC	CATTG	CATCA	CATCT	CATCC	CATCG	CATGA	CATGT	CATGC	CATGG
CAC	CACAA	CACAT	CACAC	CACAG	CACTA	CACTT	CACTC	CACTG	CACCA	CACCT	CACCC	CACCG	CACGA	CACGT	CACGC	CACGG
CAG	CAGAA	CAGAT	CAGAC	CAGAG	CAGTA	CAGTT	CAGTC	CAGTG	CAGCA	CAGCT	CAGCC	CAGCG	CAGGA	CAGGT	CAGGC	CAGGG
CTA	CTAAA	CTAAT	CTAAC	CTAAG	CTATA	CTATT	CTATC	CTATG	CTACA	CTACT	CTACC	CTACG	CTAGA	CTAGT	CTAGC	CTAGG
CTT	CTTAA	CTTAT	CTTAC	CTTAG	CTTTA	CTTTT	CTTTC	CTTTG	CTTCA	CTTCT	CTTCC	CTTCG	CTTGA	CTTGT	CTTGC	CTTGG
CTC	CTCAA	CTCAT	CTCAC	CTCAG	CTCTA	CTCTT	CTCTC	CTCTG	CTCCA	CTCCT	CTCCC	CTCCG	CTCGA	CTCGT	CTCGC	CTCGG
CTG	CTGAA	CTGAT	CTGAC	CTGAG	CTGTA	CTGTT	CTGTC	CTGTG	CTGCA	CTGCT	CTGCC	CTGCG	CTGGA	CTGGT	CTGGC	CTGGG
CCA	CCAAA	CCAAT	CCAAC	CCAAG	CCATA	CCATT	CCATC	CCATG	CCACA	CCACT	CCACC	CCACG	CCAGA	CCAGT	CCAGC	CCAGG
CCT	CCTAA	CCTAT	CCTAC	CCTAG	CCTTA	CCTTT	CCTTC	CCTTG	CCTCA	CCTCT	CCTCC	CCTCG	CCTGA	CCTGT	CCTGC	CCTGG
CCC	CCCAA	CCCAT	CCCAC	CCCAG	CCCTA	CCCTT	CCCTC	CCCTG	CCCCA	CCCCT	CCCCC	CCCCG	CCCGA	CCCGT	CCCGC	CCCGG
CCG	CCGAA	CCGAT	CCGAC	CCGAG	CCGTA	CCGTT	CCGTC	CCGTG	CCGCA	CCGCT	CCGCC	CCGCG	CCGGA	CCGGT	CCGGC	CCGGG
CGA	CGAAA	CGAAT	CGAAC	CGAAG	CGATA	CGATT	CGATC	CGATG	CGACA	CGACT	CGACC	CGACG	CGAGA	CGAGT	CGAGC	CGAGG
CGT	CGTAA	CGTAT	CGTAC	CGTAG	CGTTA	CGTTT	CGTTC	CGTTG	CGTCA	CGTCT	CGTCC	CGTCG	CGTGA	CGTGT	CGTGC	CGTGG
CGC	CGCAA	CGCAT	CGCAC	CGCAG	CGCTA	CGCTT	CGCTC	CGCTG	CGCCA	CGCCT	CGCCC	CGCCG	CGCGA	CGCGT	CGCGC	CGCGG
CGG	CGGAA	CGGAT	CGGAC	CGGAG	CGGTA	CGGTT	CGGTC	CGGTG	CGGCA	CGGCT	CGGCC	CGGCG	CGGGA	CGGGT	CGGGC	CGGGG
GAA	GAAAA	GAAAT	GAAAC	GAAAG	GAATA	GAATT	GAATC	GAATG	GAACA	GAACT	GAACC	GAACG	GAAGA	GAAGT	GAAGC	GAAGG
GAT	GATAA	GATAT	GATAC	GATAG	GATTA	GATTT	GATTC	GATTG	GATCA	GATCT	GATCC	GATCG	GATGA	GATGT	GATGC	GATGG
GAC	GACAA	GACAT	GACAC	GACAG	GACTA	GACTT	GACTC	GACTG	GACCA	GACCT	GACCC	GACCG	GACGA	GACGT	GACGC	GACGG
GAG	GAGAA	GAGAT	GAGAC	GAGAG	GAGTA	GAGTT	GAGTC	GAGTG	GAGCA	GAGCT	GAGCC	GAGCG	GAGGA	GAGGT	GAGGC	GAGGG
GTA	GTAAA	GTAAT	GTAAC	GTAAG	GTATA	GTATT	GTATC	GTATG	GTACA	GTACT	GTACC	GTACG	GTAGA	GTAGT	GTAGC	GTAGG
GTT	GTTAA	GTTAT	GTTAC	GTTAG	GTTTA	GTTTT	GTTTC	GTTTG	GTTCA	GTTCT	GTTCC	GTTCG	GTTGA	GTTGT	GTTGC	GTTGG
GTC	GTCAA	GTCAT	GTCAC	GTCAG	GTCTA	GTCTT	GTCTC	GTCTG	GTCCA	GTCCT	GTCCC	GTCCG	GTCGA	GTCGT	GTCGC	GTCGG
GTG	GTGAA	GTGAT	GTGAC	GTGAG	GTGTA	GTGTT	GTGTC	GTGTG	GTGCA	GTGCT	GTGCC	GTGCG	GTGGA	GTGGT	GTGGC	GTGGG
GCA	GCAAA	GCAAT	GCAAC	GCAAG	GCATA	GCATT	GCATC	GCATG	GCACA	GCACT	GCACC	GCACG	GCAGA	GCAGT	GCAGC	GCAGG
GCT	GCTAA	GCTAT	GCTAC	GCTAG	GCTTA	GCTTT	GCTTC	GCTTG	GCTCA	GCTCT	GCTCC	GCTCG	GCTGA	GCTGT	GCTGC	GCTGG
GCC	GCCAA	GCCAT	GCCAC	GCCAG	GCCTA	GCCTT	GCCTC	GCCTG	GCCCA	GCCCT	GCCCC	GCCCG	GCCGA	GCCGT	GCCGC	GCCGG
GCG	GCGAA	GCGAT	GCGAC	GCGAG	GCGTA	GCGTT	GCGTC	GCGTG	GCGCA	GCGCT	GCGCC	GCGCG	GCGGA	GCGGT	GCGGC	GCGGG
GGA	GGAAA	GGAAT	GGAAC	GGAAG	GGATA	GGATT	GGATC	GGATG	GGACA	GGACT	GGACC	GGACG	GGAGA	GGAGT	GGAGC	GGAGG
GGT	GGTAA	GGTAT	GGTAC	GGTAG	GGTTA	GGTTT	GGTTC	GGTTG	GGTCA	GGTCT	GGTCC	GGTCG	GGTGA	GGTGT	GGTGC	GGTGG
GGC	GGCAA	GGCAT	GGCAC	GGCAG	GGCTA	GGCTT	GGCTC	GGCTG	GGCCA	GGCCT	GGCCC	GGCCG	GGCGA	GGCGT	GGCGC	GGCGG
GGG	GGGAA	GGGAT	GGGAC	GGGAG	GGGTA	GGGTT	GGGTC	GGGTG	GGGCA	GGGCT	GGGCC	GGGCG	GGGGA	GGGGT	GGGGC	GGGGG

Before filtering: read2: overrepresented sequences

Sampling rate: 1 / 20

overrepresented sequence	count (% of bases)	distribution: cycle 1 ~ cycle 151
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	106 (0.009295%)	
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	19 (0.004165%)	
AAAGTGAGCTCAGAGCTTTCCCCCGAAAGCCAAAAGTTTCAACCCGAGTTAGGTCGGCAAACTCTTGAATCTTAGATGTGTGATGTTTCAGAGCTGAAAG	3 (0.000658%)	
ATTCTTCCTTTGAAAAGAGCTAGCCTCTCAGCCACTGGGTCTTGTCCATAGGCTCAGAACGACCTGATCTGGAACATCAAGGACGAGCTAAAGAAAGTGT	10 (0.002192%)	
CAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	228 (0.019993%)	
CTATAAAACAACCTAACCAGAAAATTCCTTGGATTTGTCATGTATTAAACTTTGGGTTTTTTTTCCAGATTGAAATGCACAGTTGGTCCTGCCATGAATA	4 (0.000877%)	
GAGAAGAATGCTTCTGAAGAGCTGCCATTCTGACGTCAAAAAACTACTAAAACGTTGAACCATTGTGTCTCACACTCAGGTGATGATTGGGACTGCTCTT	20 (0.004385%)	
GAGCTTTCCCCCGAAAGCCAAAAGTTTCAACCCGAGTTAGGTCGGCAAACTCTTGAATCTTAGATGTGTGATGTTTCAGAGCTGAAAGAGACTGTAAAGT	3 (0.000658%)	
GCACAAAGTGAGCTCAGAGCTTTCCCCCGAAAGCCAAAAGTTTCAACCCGAGTTAGGTCGGCAAACTCTTGAATCTTAGATGTGTGATGTTTCAGAGCTG	27 (0.005919%)	
GCTCAGAGCTTTCCCCCGAAAGCCAAAAGTTTCAACCCGAGTTAGGTCGGCAAACTCTTGAATCTTAGATGTGTGATGTTTCAGAGCTGAAAGAGACTGT	3 (0.000658%)	
GCTTTCCCCCGAAAGCCAAAAGTTTCAACCCGAGTTAGGTCGGCAAACTCTTGAATCTTAGATGTGTGATGTTTCAGAGCTGAAAGAGACTGTAAAGTAA	4 (0.000877%)	
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	393 (0.086156%)	
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	284 (0.092768%)	
GTCACTTCCTATAAAACAACCTAACCAGAAAATTCCTTGGATTTGTCATGTATTAAACTTTGGGTTTTTTTTCCAGATTGAAATGCACAGTTGGTCCTGC	14 (0.003069%)	
GTGAGCTCAGAGCTTTCCCCCGAAAGCCAAAAGTTTCAACCCGAGTTAGGTCGGCAAACTCTTGAATCTTAGATGTGTGATGTTTCAGAGCTGAAAGAGA	4 (0.000877%)	
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	101 (0.008857%)	
TAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	158 (0.013855%)	
TTTTTTTTTT	4504 (0.098739%)	

Grazie!

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