AIBT Associazione Italiana di Immunogenetica e Biologia dei Trapianti
XXX Congresso Nazionale
Napoli, 10/12 ottobre 2024
"The programme of this event has been approved by the EFI Education Committee"

Analisi bioinformatica nell'era NGS: biotools a supporto della tipizzazione HLA.

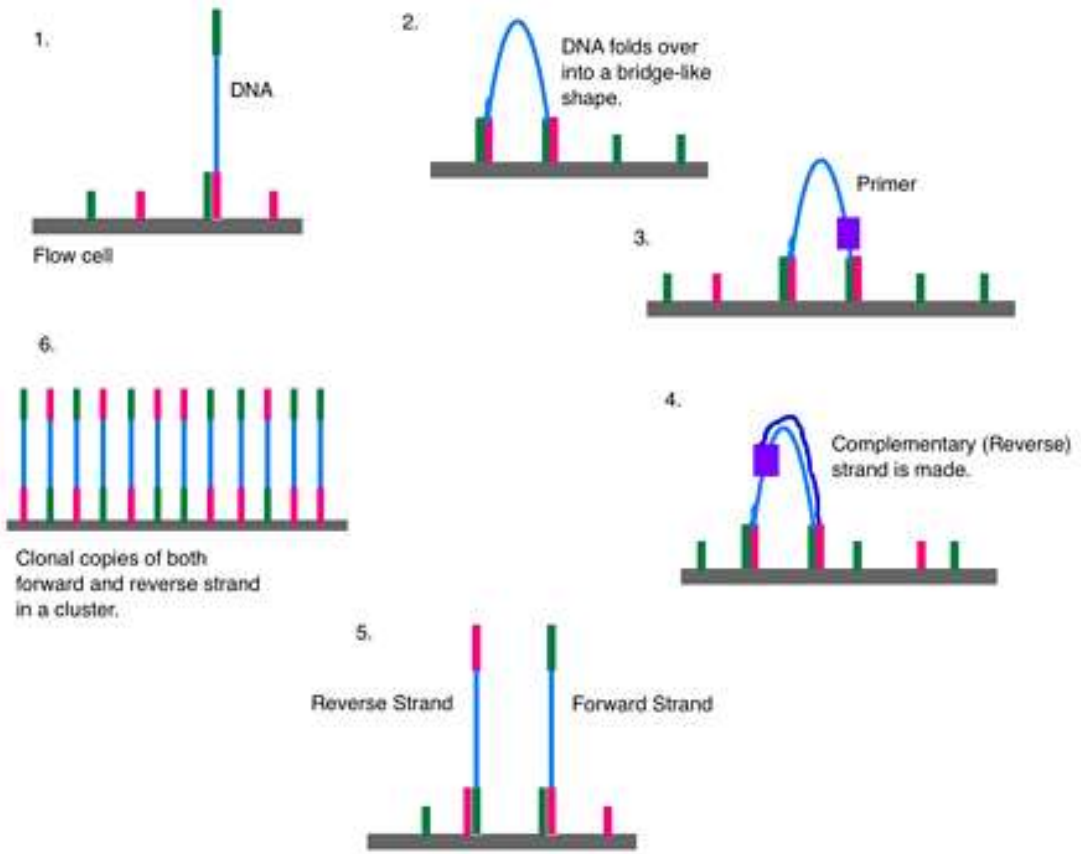
Veronica Scaglia

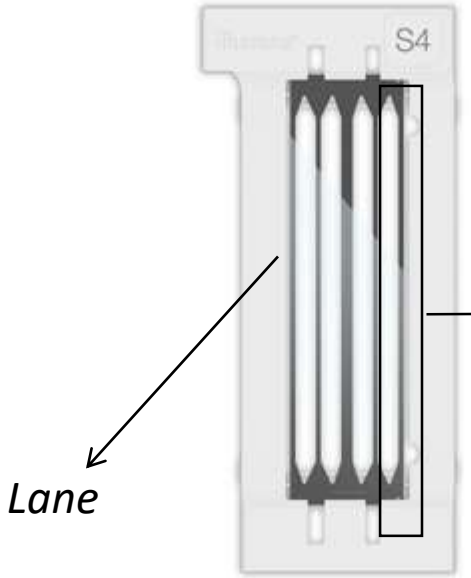
UO Biologia dei Trapianti, Diagnostica Molecolare e Manipolazione CSE

Dipartimento di Patologia Clinica, AUSL Piacenza

SERVIZIO SANITARIO REGIONALE
EMILIA ROMAGNA
Azienda Unità Sanitaria Locale di Piacenza

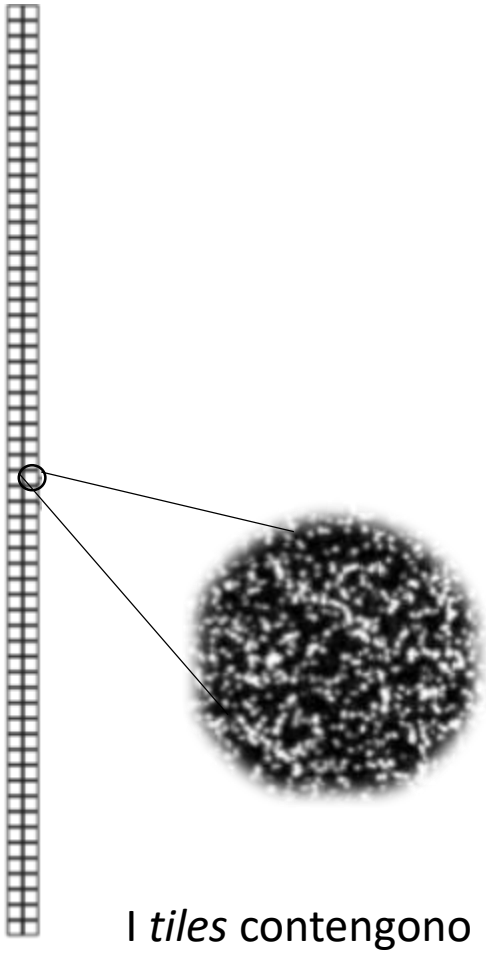
Sequenziamento NGS piattaforma Illumina





Lane

1 *lane* contiene 2 file di *tiles*



I *tiles* contengono poi i *clusters*

Analisi dei dati NGS: *workflow*

Analisi primaria

- Racchiude tutti gli step utili a determinare la sequenza delle basi, è tipicamente eseguita dal sequenziatore dopo ogni ciclo.
- Produce letture (*reads*) in formato FastQ

Analisi secondaria

Racchiude tutti i processi necessari ad allineare i frammenti sequenziati (*contig*), escludendo i dati di bassa qualità.

Analisi terziaria

- Visualizzazione dei dati prodotti grazie a software dedicati
- Interpretazione dei dati

A. RTA Logs folder— Contains log files that describe each step performed by RTA for each Read.

B. InterOp folder— Contains binary files used by Sequencing Analysis Viewer (SAV) to summarize various primary analysis metrics.

C. Logs folder— Contains log files that describe every step performed by the instrument for each cycle.

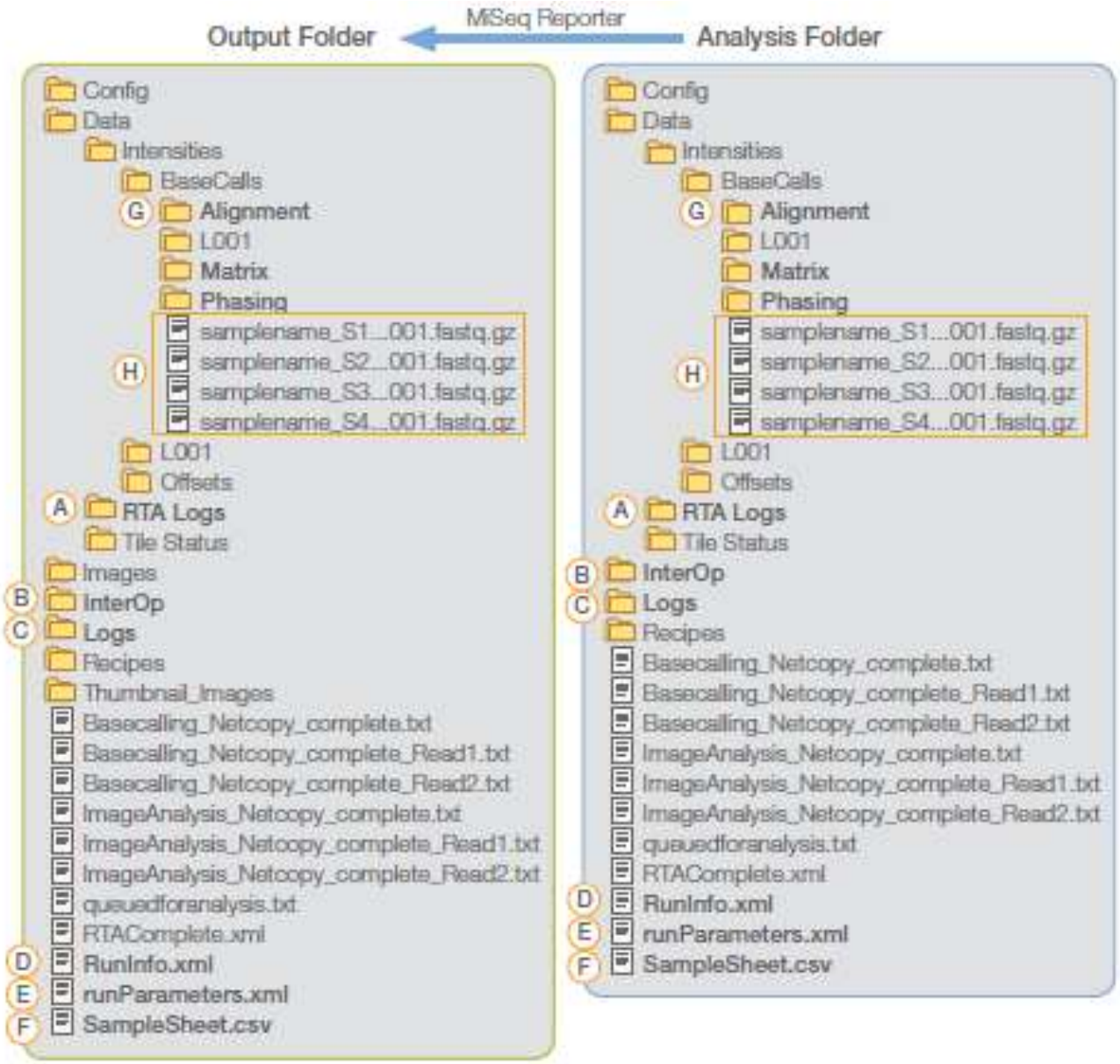
D. RunInfo.xml— Contains high-level run information, such as the number of Reads and cycles in the sequencing run.

E. runParameters.xml— Contains a summary of run parameters and information about run components.

F. SampleSheet.csv— Provides parameters for the run and subsequent analysis.

G. Alignment folder— Contains alignment (*.bam) and variant call (*.vcf) files. This folder is not created if you set up your run with the GenerateFASTQ workflow.

H. FASTQ files— Provide intermediate analysis data for downstream third-party analysis. Files are located in Data\Intensities\BaseCalls.



Sequencing Analysis Viewer

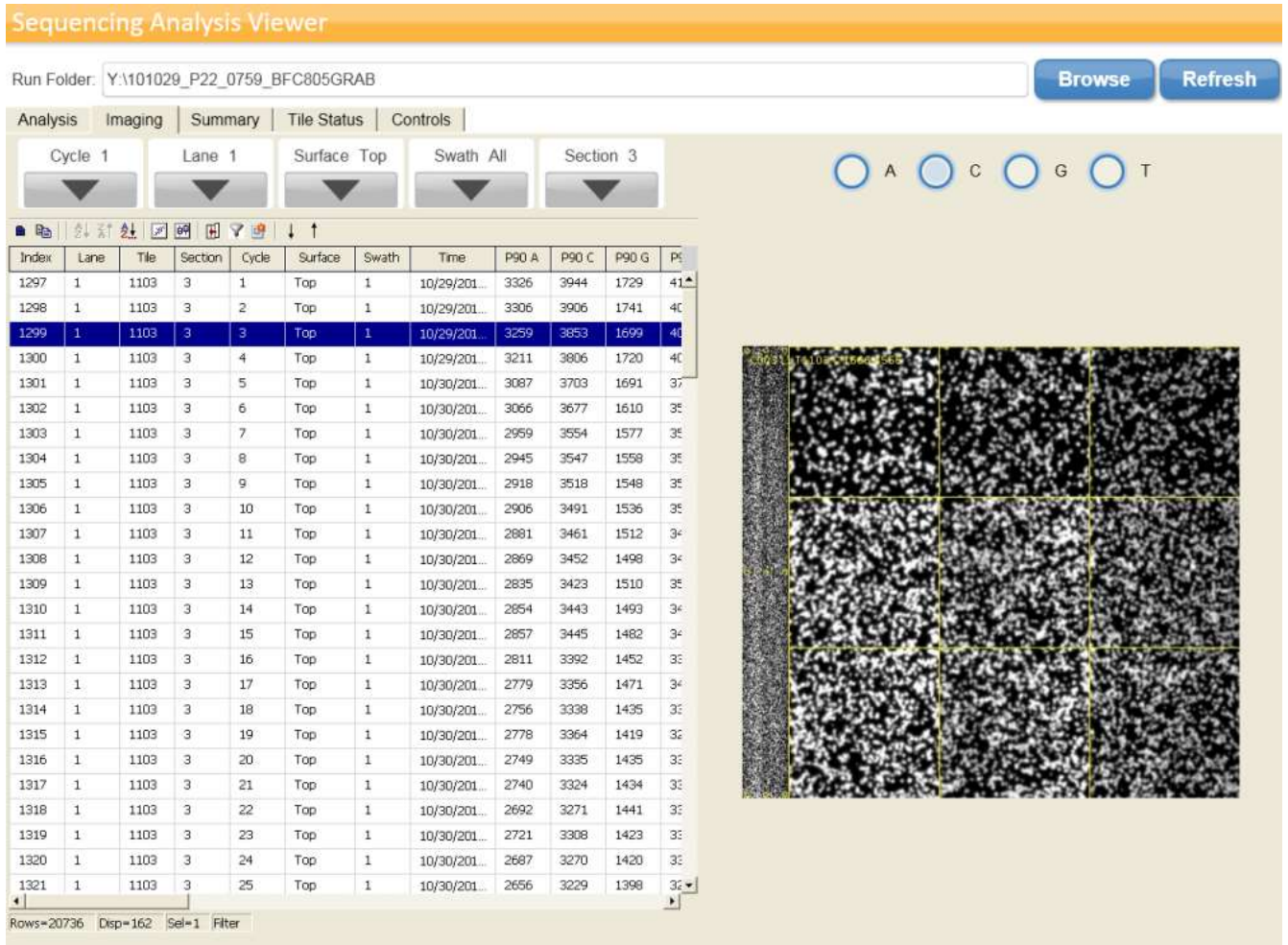
Analysis Tab

- **Flow Cell Chart** mostra i parametri di qualità delle sezioni sulla flowcell in codice colore
- **Data by Cicle** mostra un grafico che permette di seguire la progressione dei parametri di qualità durante tutta la corsa
- **Data by Lane** mostra i parametri di qualità per ogni lane
- **Q Score Distribution Plot** mostra in un grafico il numero di reads che superano il filtro di qualità
- **Q score Heatmap** mostra il Q score per ogni ciclo di sequenziamento.

Sequencing Analysis Viewer

Imaging Tab

- Mostra una lista dettagliata dei parametri e dei dati della corsa.
- Ogni lane, ciclo, sezione può essere selezionata per valutarne le caratteristiche specifiche.

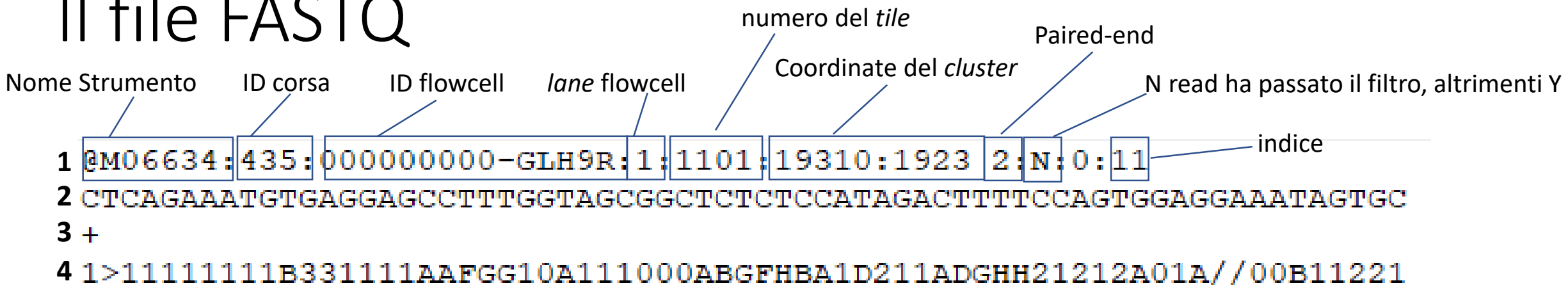


Sequencing Analysis Viewer

Summary Tab

- Mostra i dati di qualità grezzi in una tabella riassunti per lane e per read. Le statistiche mostrano medie e deviazioni standard delle varie lane.
- Cluster Density: 600-1200 K/mm^2
- Cluster Passing Filter: >80%
- % reads $\geq$ Q30: >90%

Il file FASTQ



1. Si riferisce all'identificativo della sequenza (inizia con @)
2. La seconda riga mostra la sequenza grezza
3. Contiene il simbolo ''+'' seguito da spazio
4. Contiene i parametri di qualità della sequenza codificati come conversione decimale del codice ASCII (**A**merican **S**tandard **C**ode for **I**nformation **I**nterchange)

I file *.SAM e *.BAM

- Il SAM (Sequence Alignment Map) è un formato di archiviazione di file di allineamento con le relative coordinate di interpretazione.
- Viene utilizzato in diversi *workflow* di sequenziamento per allineare le reads ad una sequenza di riferimento.
- BAM è un file SAM compresso, scritto in codice binario. Non decifrabile da un operatore, ma solo da software dedicati.
- Alcuni tools permettono di convertire i file *.SAM in *.BAM (es. Samtools)

Controllo qualità dei FASTQ nella tipizzazione HLA

- Permette di affiancare i software commerciali per migliorare la qualità del dato;
- Permette di identificare eventuali problemi sorti durante il sequenziamento;
- Utile in fase di validazione / monitoraggio continuo per identificare le caratteristiche intrinseche della metodica;
- Identificazione e miglioramento dell'eventuale errore introdotto durante le fasi della metodica.

FastQC

- Tool gratuito di controllo qualità dei dati NGS
- File input *.fastq, *.bam, *.sam
- Generazione di un report *.html con i parametri di qualità del file inserito (visualizzabile tramite browser)
- In caso di sequenziamenti *paired-end* vanno inseriti entrambi i *files* *.fastq generati e verranno analizzati separatamente.

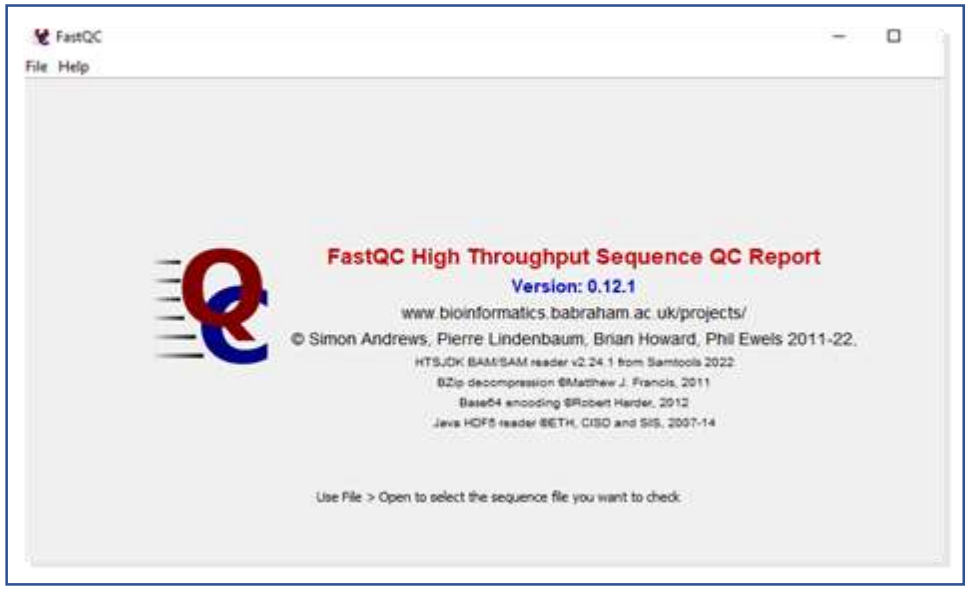
Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data

- Configuration
- Help
- net
- org
- Templates
- uk
- cisd-jhdf5.jar
- fastqc
- fastqc_icon.ico
- htsjdk.jar
- INSTALL.txt
- jbzip2-0.9.jar
- LICENSE
- LICENSE.txt
- LICENSE_JHDF5.txt
- README.md
- README.txt
- RELEASE_NOTES.txt
- run_fastqc.bat

- adapter_list.txt
- contaminant_list.txt
- limits.txt

N.B. Questi file *.txt possono essere modificati manualmente dall'operatore per adattarli alle esigenze specifiche di analisi.

- 1 Introduction
- 2 Basic Operations
- 3 Analysis Modules



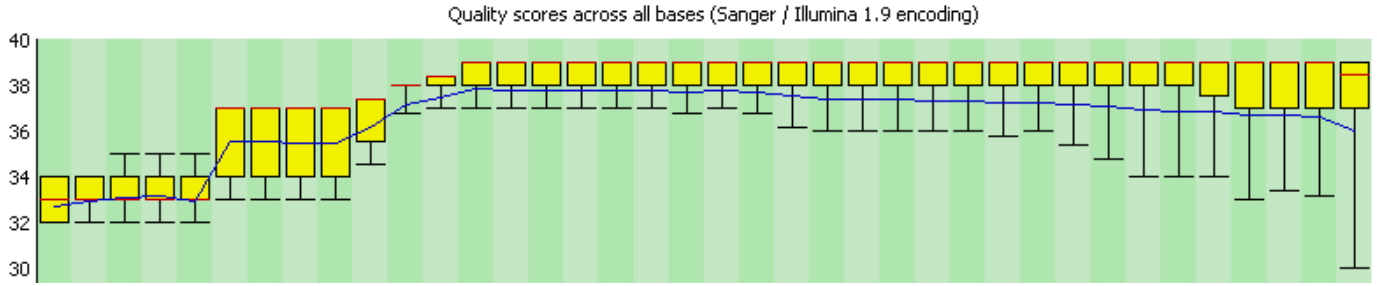
FastQC Report

Summary

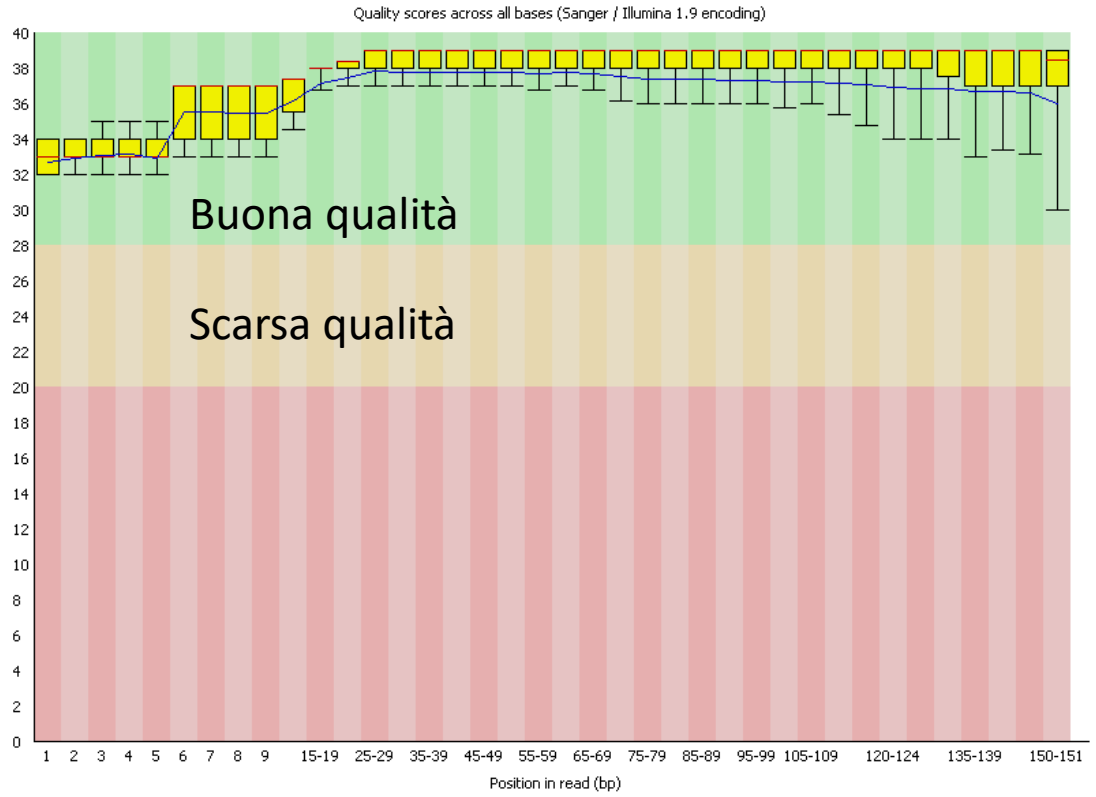
Basic Statistics

Measure	Value
Filename	HA-241163-514-706_S8_L001_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	313531
Total Bases	46.6 Mbp
Sequences flagged as poor quality	0
Sequence length	35-151
%GC	50

Per base sequence quality

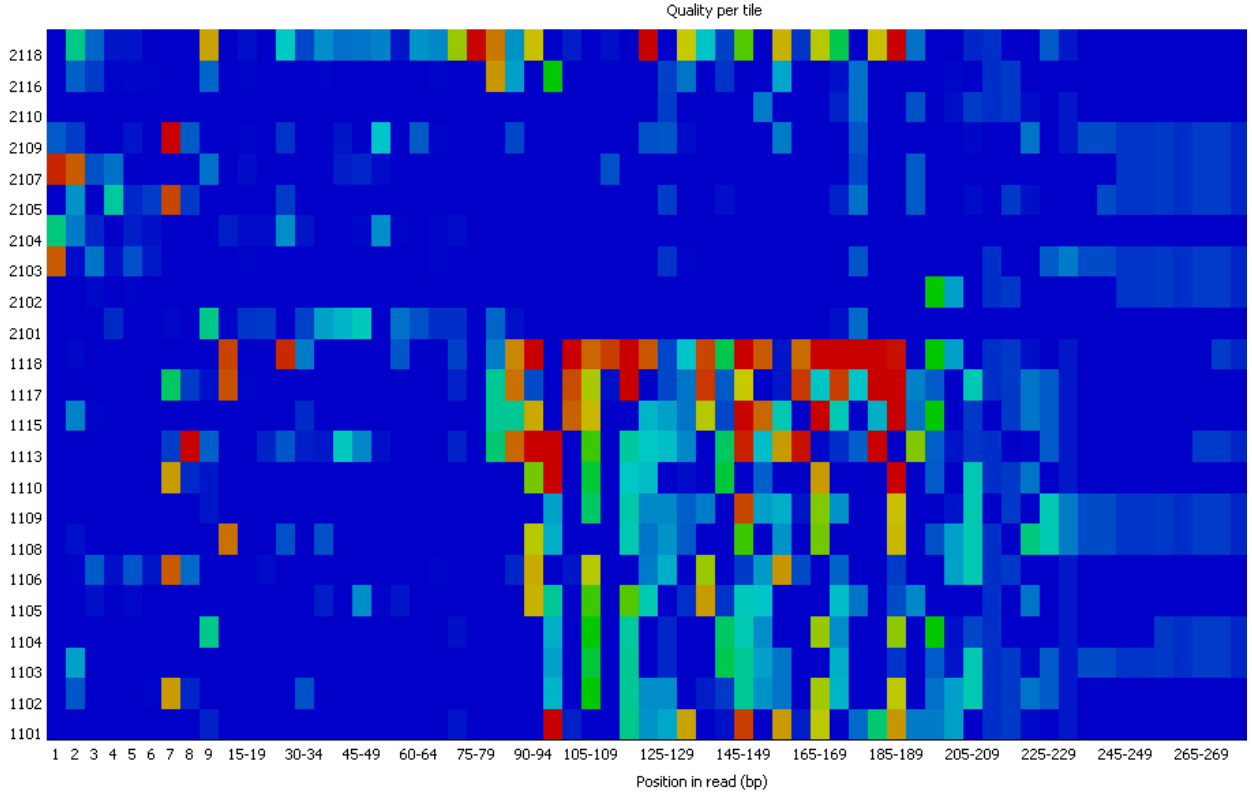


Per base sequence quality



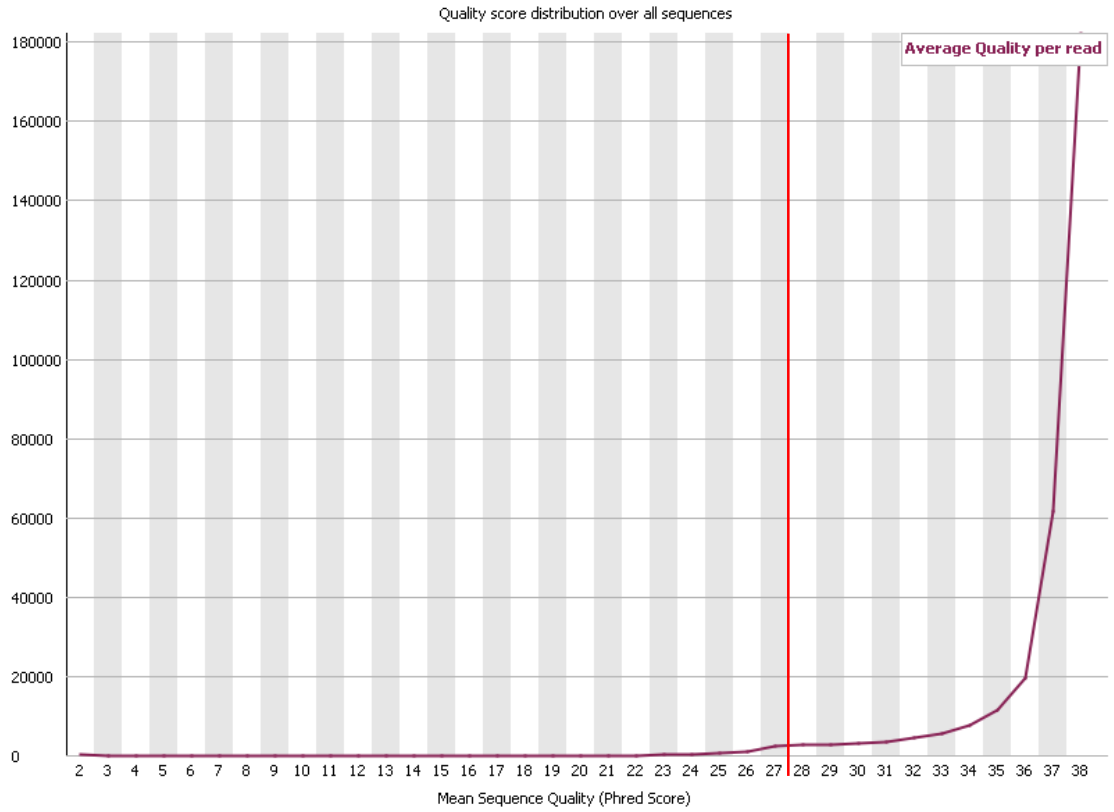
- Il box giallo rappresenta la qualità del 25-75% delle reads in quella posizione, i whiskers sotto il box giallo rappresentano il limite del 10% e 90%. Tendenzialmente la qualità tende ad abbassarsi per corse molto lunghe.

Per tile sequence quality



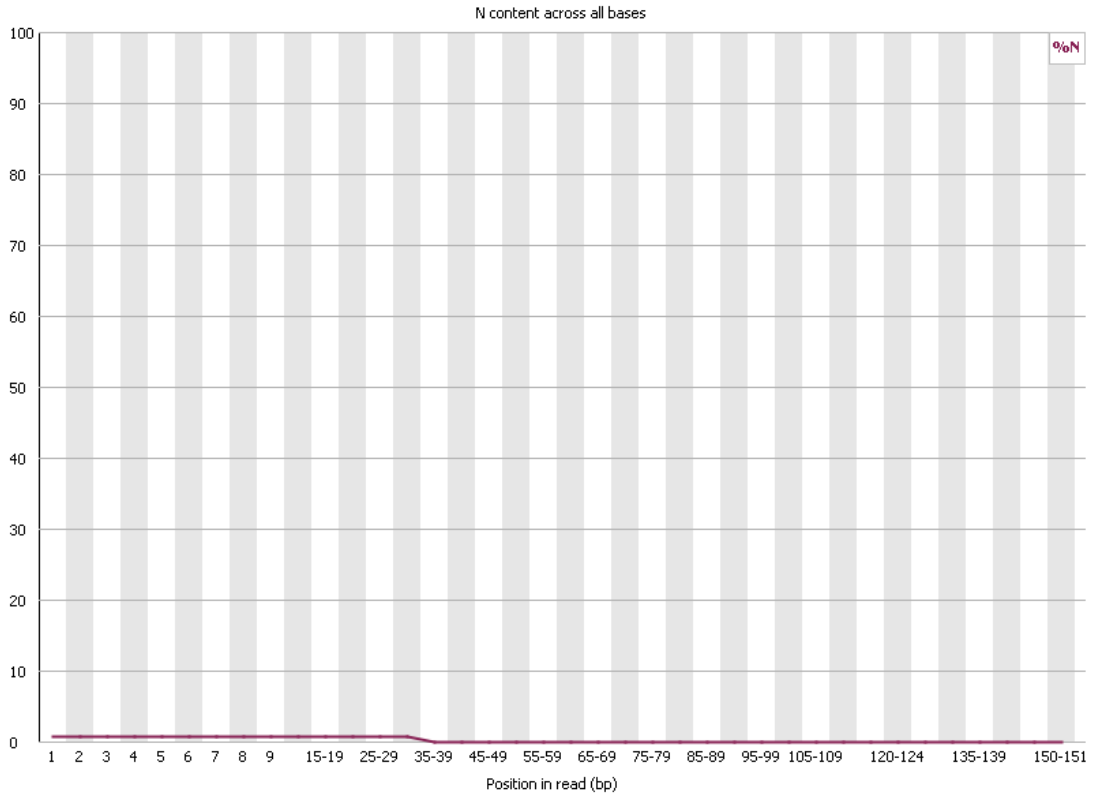
- Nelle librerie di Illumina sono presenti informazioni relative all'identificativo della sequenza, per cui è possibile ricostruire la qualità associata ai *tile* per identificare eventuali problemi in una regione della flowcell. Le sfumature di blu indicano buona qualità.

Per sequence quality scores



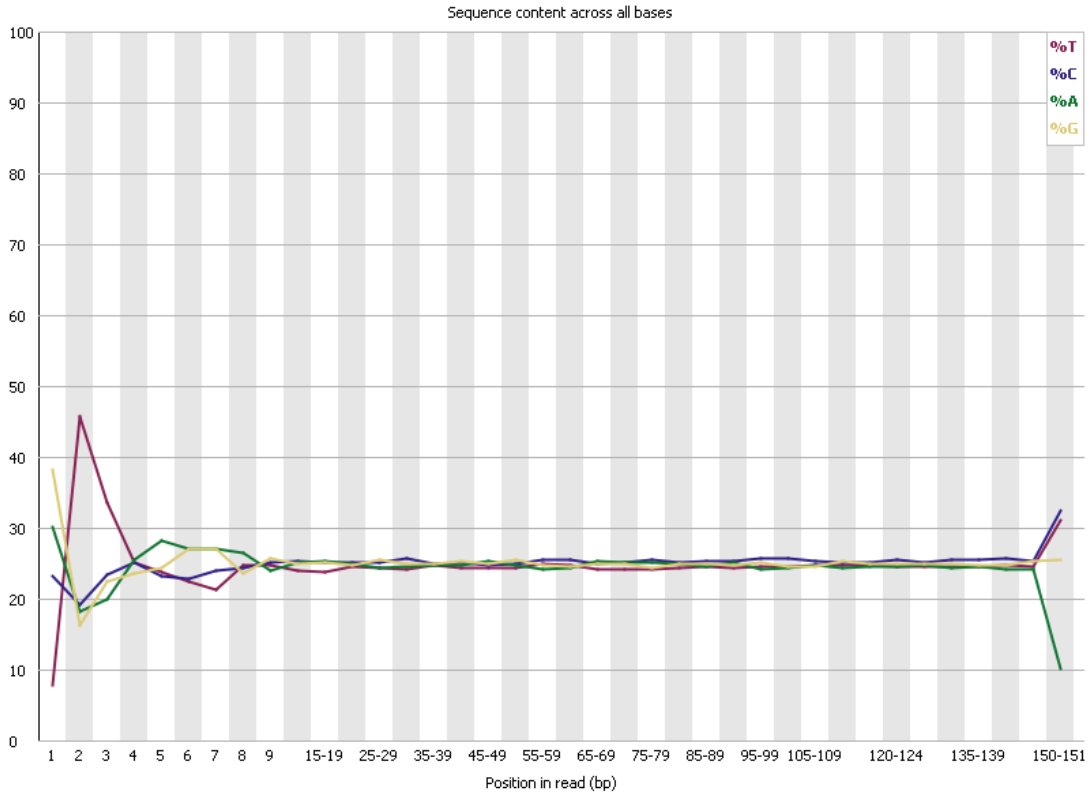
- Parametri di qualità medi per read, valori sotto 20 indicano probabilità di errore dell' 1% nell'assegnazione delle basi, il valore ottimale è sopra 27.

Per base N content



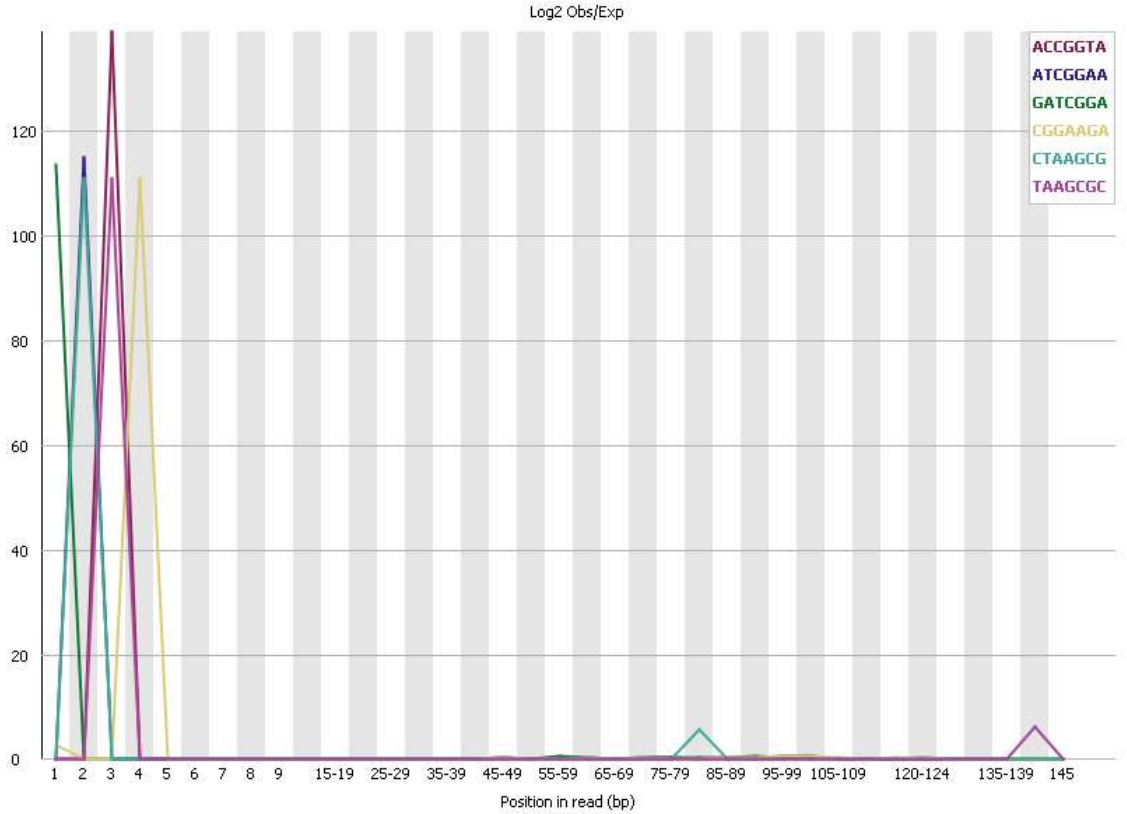
- Il valore N viene assegnato in una posizione quando il sequenziamento non è in grado di assegnare la base. Questo grafico mostra la % di basi chiamate in ogni posizione, per le quali è stata assegnata una N. Un aumento di %N può indicare che un bias della reazione rende il *basecaller* instabile.

Per base sequence content



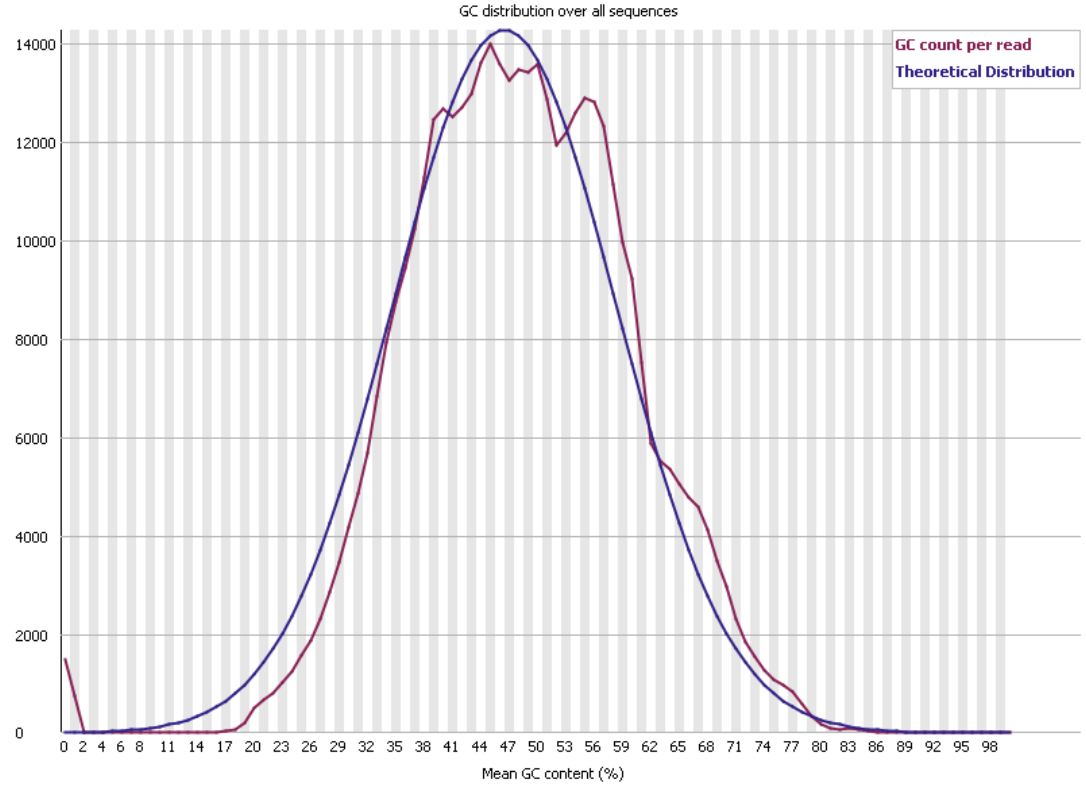
- Questo parametro mostra la proporzione della presenza di ogni base nel sequenziamento. In una libreria ottimale le proporzioni dovrebbero mantenersi costanti. All'inizio della sequenza si può verificare sbilanciamento per bias nella frammentazione o *random priming*; in questo caso non interferisce con la qualità del dato a valle.

Kmer Content



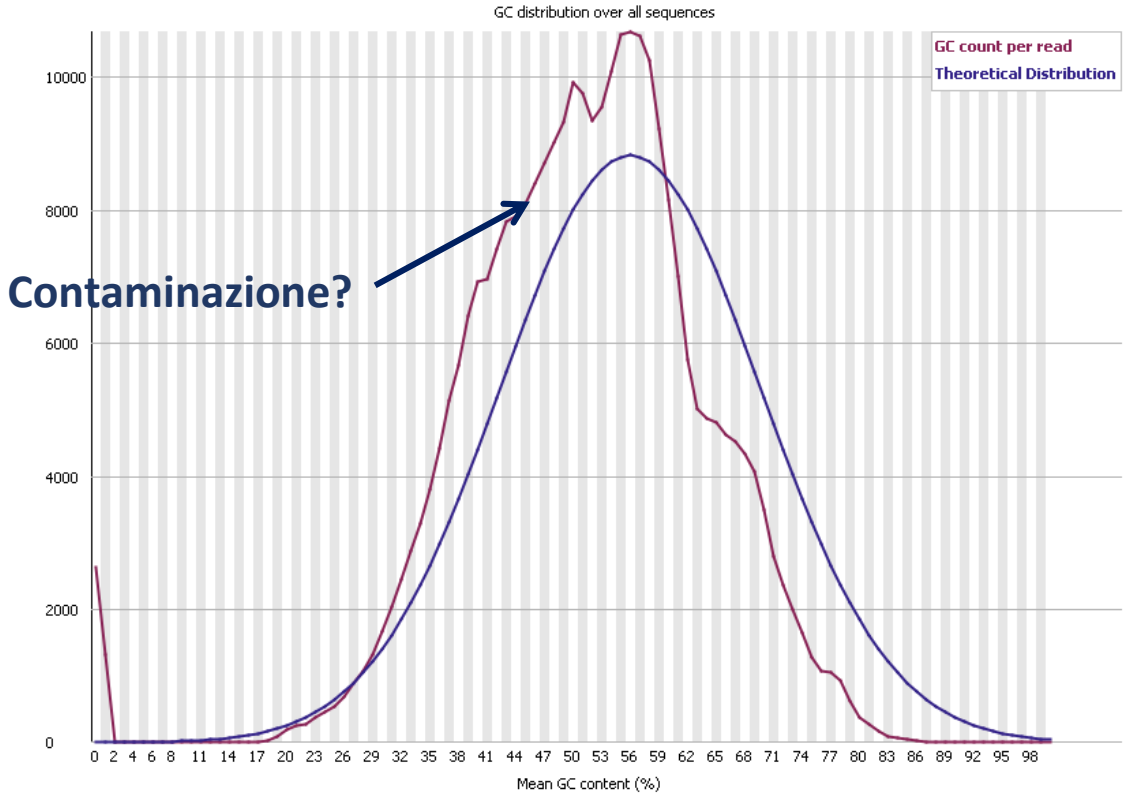
- Il Kmer content parte dal presupposto che piccoli segmenti di sequenza (*Kmeri*) dovrebbero essere equamente rappresentati all'interno di una *library*. All'inizio della sequenza questo dato può aumentare in caso di *random priming*, ed è correlato al ''*Per base sequence content*''.

Per sequence GC content



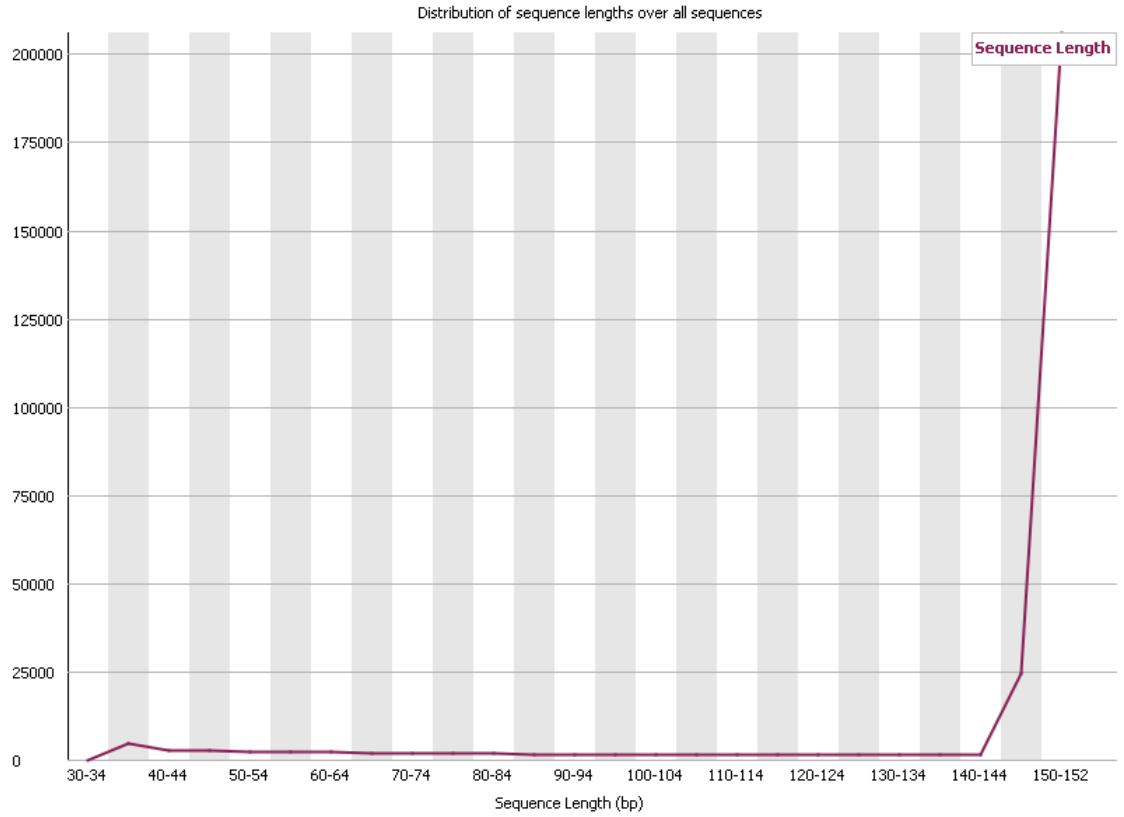
- Questa sezione mostra il *GC content* sull'intera distribuzione delle sequenze, viene comparata ad una distribuzione normale teorica del *GC content*. La curva relativa al sequenziamento deve essere il più possibile sovrapponibile alla distribuzione normale.

Per sequence GC content



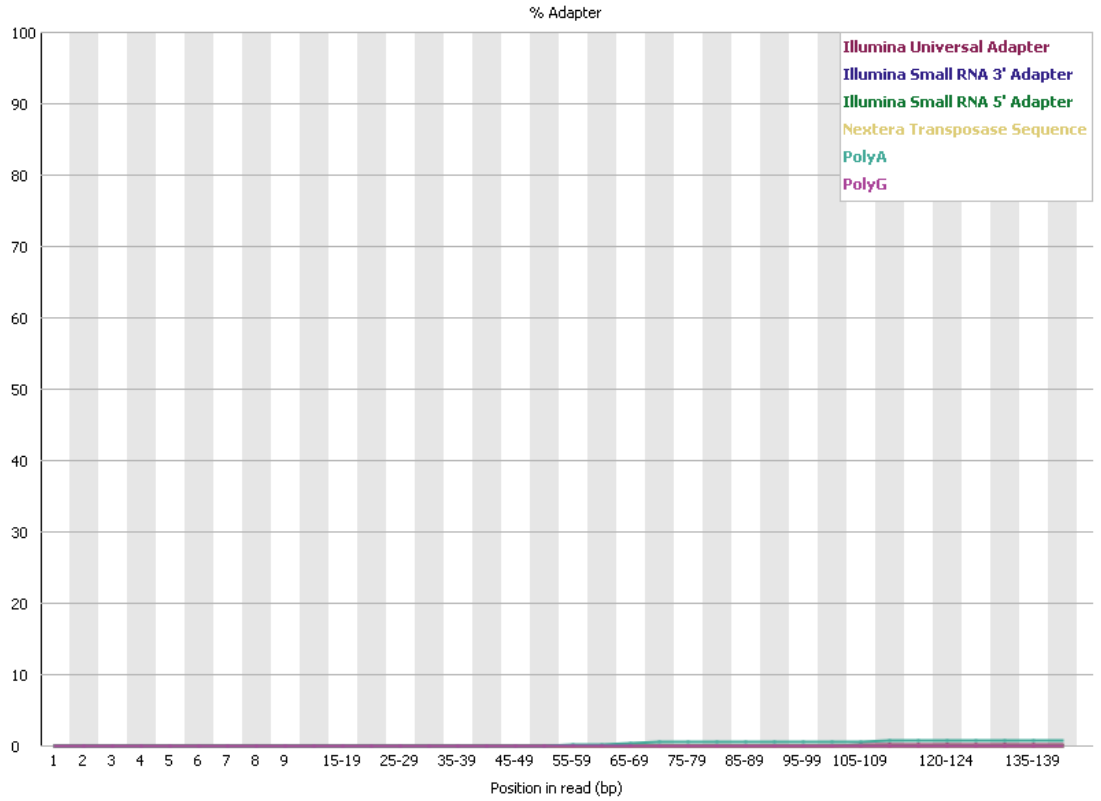
-Uno *shift* nella curva osservata indica un bias nella rappresentazione delle basi, che può essere dato da contaminazione o da problemi della *library.* Uno *shift* sistematico non viene indicato come errore, siccome può essere una caratteristica intrinseca della *library.*

Sequence Length Distribution



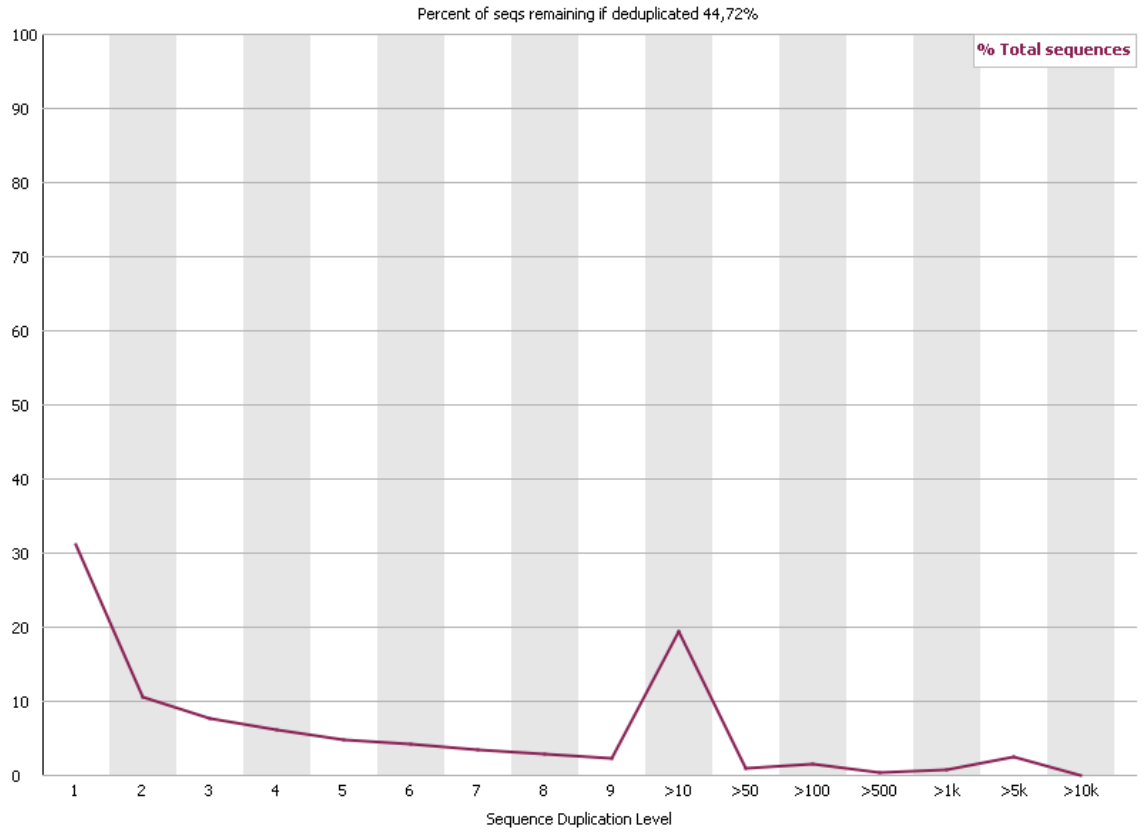
- Le sequenze dovrebbero avere tutte la stessa lunghezza, che dipende dalla metodica utilizzata.

Adapter Content



- Questo grafico mostra l'eventuale sovra rappresentazione di adapter nel sequenziamento. Nel software è presente una lista di adapter di default che può essere modificata manualmente dall'operatore nella sezione Configuration > adapter_list.txt

Sequence Duplication Levels



-In una *library* tendenzialmente le sequenze sono diversificate; quando alcune sequenze sono presenti in duplicato ciò può indicare over-amplificazione nella PCR, contaminati a bassa complessità, sequenze sovra rappresentate.

Overrepresented sequences

Sequence	Count	Percentage
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGACATAGTATCTCGTAT	7459	2.67885361298664
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	2629	0.9441890532969401
GGCAGACAGTGTGACAAAGAGGCTGGTGTAGGAGAAGAGGGATCAGGACG	677	0.24314035339750034
TGGATACTCACGACGCGGACCCAGTTCTCACTCCCATTGGGTGTCGGGTT	580	0.20830340468323516
CTCATCTTCTGCCTCTCCATTTCCTTGCTTGCTATAGTTCAGTCCCTCTA	426	0.15299525930182445
CTCATCTTCTGCCTCTCCATTTCCTTGCTTGCTATACTTCAGTCCCTCTA	406	0.14581238327826462
GCAGGTGCCTTTGCAGAAACAAAGTCAGGGTTCTTCAAGTCACAAAGGGA	400	0.14365752047119665
CTGGGGAGGAAACACAGGTCAGCATGGGAACAGGGGTCACAGTGGACACG	294	0.10558827754632956

-Nella maggioranza dei sequenziamenti HLA da noi analizzati, sono presenti sequenze over rappresentate. Il software riporta la percentuale delle sequenze più abbondanti sopra lo 0,1% del totale. Nel software è presente una lista di contaminati di default identificabili, che può essere modificata manualmente dall'operatore nella cartella Configuration > contaminat_list.txt

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGTCTATGAATCTCGTAT	5189	1.2673902848880378	TruSeq Adapter, Index 11 (97% over 38bp)
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	1483	0.36221618664269806	No Hit
CTCATCTTCTGCCTCTCCATTTCCTTGCTTGCTATAGTTCAGTCCCTCTA	1148	0.28039391926218293	No Hit
CTCATCTTCTGCCTCTCCATTTCCTTACTTTCTATACATCAGCCCCTCTA	466	0.11381843760991052	No Hit

- Nel sequenziamento HLA la sequenza più abbondante (sopra l'1%) fa spesso riferimento agli indici di Illumina
- Sono presenti altre sequenze di origine ignota che si è reso necessario indagare

Ricerca delle sequenze nel database NCBI BLAST

FastQC

File Help

225033UCLA1072-512-711_S4_L001_R1_001.fastq.gz | 225033UCLA1072-512-711_S4_L001_R2_001.fastq.gz

- 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

Overrepresented sequences

Sequence	Count	Percentage	Possible Sou...
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGTCTATGAATCTCGTAT	5189	1,267	TruSeq Adapt...
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	1483	0,362	No Hit
CTCATCTTCTGCCTCTCCATTTCCTTGCTTGCTATAGTTCAGTCCCTCTA	1148	0,28	No Hit
CTCATCTTCTGCCTCTCCATTTCCTTACTTTCTATACATCAGCCCCTCTA	466	0,114	No Hit

Homo sapiens, HLA-DRB1 gene for MHC class II antigen, cell line DKMS-LSL_ID15821130_DRB1_1, allele HLA-DRB1*14:new

GenBank: LR882672.1

GenBank FASTA

Link To This View | Feedback

FastQC

File Help

UCLAHLA1114ANNO2024-241036-513-702_S8_L001_R1_001.fastq.gz | UCLAHLA1114ANNO2024-241036-513-702_S8_L001_R2_001.fastq.gz

- 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

Overrepresented sequences

Sequence	Count	Percentage	Possible Sou...
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGACATAGTATCTCGTAT	7459	2,679	TruSeq Adapt...
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	2629	0,944	No Hit
GGCAGACAGTGTGACAAAGAGGCTGGTGTAGGAGAAGAGGGATCAGGACG	677	0,243	No Hit
TGGATACTCACGACGCGGACCCAGTTCTCACTCCCATTGGGTGTCGGGTT	580	0,208	No Hit
CTCATCTTCTGCCTCTCCATTTCCTTGCTTGCTATAGTTCAGTCCCTCTA	426	0,153	No Hit
CTCATCTTCTGCCTCTCCATTTCCTTGCTTGCTATACTTCAGTCCCTCTA	406	0,146	No Hit
GCAGGTGCCTTTGCAGAAACAAAGTCAGGGTTCTTCAAGTCACAAAGGGA	400	0,144	No Hit
CTGGGGAGGAAACACAGGTCAGCATGGGAACAGGGGTCACAGTGGACACG	294	0,106	No Hit

Homo sapiens HLA-B gene for MHC class I antigen, allele HLA-B*13:02:01:new

GenBank: OX464614.1

GenBank FASTA

Link To This View | Feedback

```
HLAB5'UTR          GGCAGACAGTGTGACAAAGAGGCTGGTGTAGGAGAAGAGGGATCAGGACG
HLAA5'UTR          TGGATACTCACGACGCGGACCCAGTTCTCACTCCCATTGGGTGTCGGGTT
HLADRB1INTRON1     CTCATCTTCTGCCTCTCCATTTCCTTGCTTGCTATAGTTCAGTCCCTCTA
HLADRB1INTRON1     CGCCTCAGGAAGACAGAGGAGGAGCCCCTGGGCTGCAGGTGGTGGGCGTT
HLAA3'UTR          GCAGGTGCCTTTGCAGAAACAAAGTCAGGGTTCTTCAAGTCACAAAGGGA
HLAB3'UTR          CTGGGGAGGAAACACAGGTCAGCATGGGAACAGGGGTCACAGTGGACACG
HLADQB15'UTR       GCCAGGTACATCAGATCCATCAGGTCCGAGCTGTGTTGACTACCACTACT
HLADQB15'UTR       CGCGGGCTGTTCCACAGCTCCGGGCCGGGTCAGGGCGGCGGCTGCGGGGT
HLADQB13'UTR       GCGTGACAGCCACTGTAGGACTTTGATCTCAGGGGGACAAGCTGACACAG
HLADRB15'UTR       TGTCAAGTGGGAAATGAATGCTCTTACAAGGCTCAAACTTGTGAACACAT
HLADPB13'UTR       GGGGTTCATGCCTGAGATTCCCTTCCTGGAGGAGCCTCAGTGTTACAATG
HLADPB13'UTR       GGCCTGGAGAGGCTCTGCGACCCGCTTAGGACCACAGAACTCGGTACTAG
HLADPA13'UTR       GTCCAAGTAGAGACAGTAAGTAGTTACAGGCACCCACCACACTGTGTTGT
HLADQA15'UTR       AGACATGCACACACCAGAGAAGATTCCGATTTCGTGTCCTCCCTCTATTC
HLADRB53'UTR       GAACCGGTAGCAACCAGGGCCTGAGAAAGTCCTCTCTTGTGGAAGAATAA
HLADRB5INTRON1     CGGATCCTCCTCCAGCTCCTGCTTGGAGGTCTCCAGAATAGGCTGGAGGC
HLADRB5INTRON1     GCATTTCCTTGCTTGCTCACTTCAGTCCCTCTAGCCTTCTTTCTGTCCCT
HLADRB4 CTGAAGTAGATGAACAGCCCTGTCCCAAGGAAGAGCAGGCCCAGCACAAA
```

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGACATAGTATCTCGTAT	7459	2.67885361298664	TruSeq Adapter, Index 6 (97% over 37bp)
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	2629	0.9441890532969401	No Hit
GGCAGACAGTGTGACAAAGAGGCTGGTGTAGGAGAAGAGGGATCAGGACG	677	0.24314035339750034	HLAB5'UTR (100% over 50bp)
TGGATACTCACGACGCGGACCCAGTTCTCACTCCCATTGGGTGTCGGGTT	580	0.20830340468323516	HLAA5'UTR (100% over 50bp)
CTCATCTTCTGCCTCTCCATTTCCTTGCTTGCTATAGTTCAGTCCCTCTA	426	0.15299525930182445	HLADRB1INTRON1 (100% over 50bp)
CTCATCTTCTGCCTCTCCATTTCCTTGCTTGCTATACTTCAGTCCCTCTA	406	0.14581238327826462	HLADRB1INTRON1 (98% over 50bp)
GCAGGTGCCTTTGCAGAAACAAAGTCAGGGTTCTTCAAGTCACAAAGGGA	400	0.14365752047119665	HLAA3'UTR (100% over 50bp)
CTGGGGAGGAAACACAGGTCAGCATGGGAACAGGGGTCACAGTGGACACG	294	0.10558827754632956	HLAB3'UTR (100% over 50bp)

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGTCTATGAATCTCGTAT	5189	1.2673902848880378	TruSeq Adapter, Index 11 (97% over 38bp)
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	1483	0.36221618664269806	No Hit
CTCATCTTCTGCCTCTCCATTTCCTTGCTTGCTATAGTTCAGTCCCTCTA	1148	0.28039391926218293	HLADRB1INTRON1 (100% over 50bp)
CTCATCTTCTGCCTCTCCATTTCCTTACTTTCTATACATCAGCCCCTCTA	466	0.11381843760991052	HLADRB1INTRON1 (96% over 30bp)

- E’ stato verificato che nella maggioranza delle sedute analizzate i contaminati facevano riferimento sempre alle stesse sequenze;
- È stata creata una lista di sequenze contaminanti da inserire nelle impostazioni del software per verificare la presenza di questi contaminati nelle successive sedute.

E’ emerso come nella metodica in uso siano le regioni 5’ e 3’ UTR della prima classe e l’introne 1 dei loci DRB3/4/5 a creare contaminazione. E’ possibile utilizzare questo strumento nella determinazione di un eventuale mix-up.

Conclusioni

- La metodica descritta si propone come supporto nella valutazione della qualità dei file FASTQ di sequenza, parametro che riveste particolare importanza nella routine laboratoristica.

- L' analisi in oggetto permette inoltre di caratterizzare i propri dati di sequenziamento in fase di validazione e affiancare i software commerciali nell'interpretazione del dato nella pratica lavorativa.

Grazie per l'attenzione