

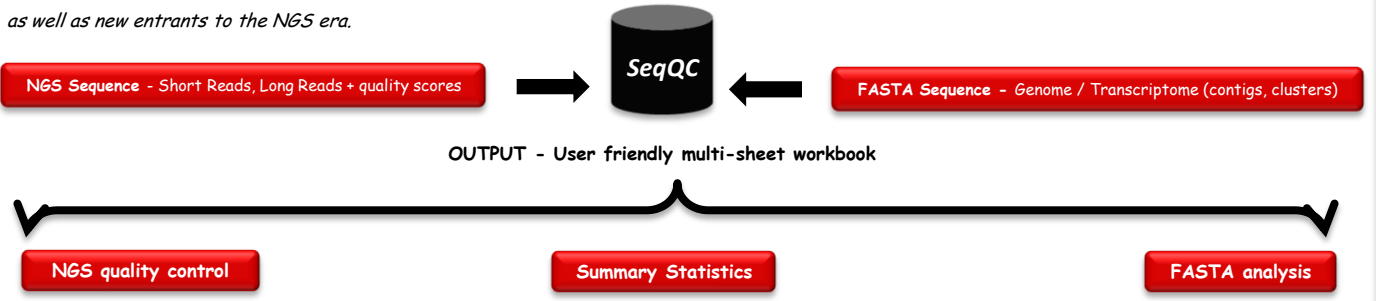
Raja Mugasimangalam*; Veer Marwah; Vasanthan Jayakumar; Shruti Sinha; Rohit Shukla
Genotypic Technology, 2-13, 80 Feet Road, RMV Second Stage, Bangalore- 560094, INDIA
X Gen Congress (Now Generation Sequencing) Year 2010 * Email - raja@genotypic.co.in

Abstract

NGS technologies have moved large scale sequencing from Genome centers to Labs. They have empowered biologists to not only undertake large scale sequencing, but use them for various other applications like DNA-Protein interactions, epigenetic changes, gene expression profiling and genotyping. Next generation Bioinformatics tools have been and are being developed to quickly align, assemble and analyze the large volumes of NGS sequence data. We describe here a tool which can enable scientists to perform a quick quality control of short read sequencing data.

Quality Control and summarization of the NGS data and the processed sequence data is essential to plan downstream bioinformatics processes. We report a simple, fully automated and easy to use desktop application SeqQC for Windows and Linux platforms that can process NGS data and any form of sequence data to generate user friendly graphs and summaries. Windows and Linux executables of the tool can be downloaded (free) from www.genotypic.co.in/SeqQC.html

Advanced bioinformatics skills or high performance computers are not required to run SeqQC and we believe the tool would be useful for advanced users as well as new entrants to the NGS era.



QC REPORT

Fastq file name	1.txt
Fastq file size	1020.92 MB
Time taken for Analysis	11 minutes
Maximum Read Length	50 b
Minimum Read Length	50 b
Median Read Length	50 b
Total Number of Reads	6016805
Total Number of HQ Reads 1*	5829568
Percentage of HQ Reads	96.888 %
Total Number of Bases	300.840 Mb
Total Number of HQ Bases 2*	290.992 Mb
Percentage of HQ Bases	96.726 %
Total Number of Non-ATGC Bases	3531 b
Percentage of Non-ATGC Bases	0.001 %
Number of Reads with Non-ATGC Bases	3271
Percentage of Reads with Non-ATGC Bases	0.054 %

QUALITY GRAPH

SNAPSHOT

```
@HWUSI-EAS232:7:8:3:1213#0/1
TTTTTCATTCTCTCATNGCTTCNCGCAGAGAATCTCAGCTTGATTTGTAT
+HWUSI-EAS232:7:8:3:1213#0/1
`aaabaabbaabba`ab`DT`aa`JDD`j`a` `aU` ]X`_aa`Z`Xa`Q`aa[ ]ZS`V
@HWUSI-EAS232:7:8:3:843#0/1
GCTTTAGGTGAAATGGACATTTATTTATGTCGTTTATGATATTACTAAOGACAT
+HWUSI-EAS232:7:8:3:843#0/1
TTVVZ ]Y`QSK`FX[VHK`TWPQN[Y]RP[ ]USXKXS ] [NR[Y]UOV`OXU`FP
```

Quick view of top and bottom lines of the sequence file

MOTIF SEARCH

Mismatch / Gaps	GACTGTGACGTACAT
0	9023
1	5987
2	2365

LOW COMPLEXITY SEQUENCE SEARCH

PolyA	19455
PolyT	6579
PolyG	35
PolyC	109
DiNucleotide Repeats	11255
TriNucleotide Repeats	5278
TetraNucleotide Repeats	17674

PRIMER / ADAPTER SEQUENCE SEARCH

PCR Primer 1	AATGATACGGCGACCACCGAGATCTCACTCT	2
Adapter 2	ACACAATTGGTGTGTGTCGGGCTCTTCCGATC	11
Sequencing Primer	ACACTCTTTCCCTACACGACGCTCTTCCGATC	7
PCR Primer 2	CAAGCAGAAAGACGGCATAAGCAGCTCTTCCGAT	35
Adapter 1	GATCGGAAGAGCTCGTATGCGCTCTCTCGCTT	37

NUCLEOTIDE COMPOSITION / BASE

QC REPORT

Sequence File Name	human.rna.fna
Sequence File Size	116.64 MB
Time taken for analysis	2.17 minutes
Maximum Sequence Length	101519
Minimum Sequence Length	33
Average Sequence Length	2564.912
No. of Sequences	45173
Total Sequences Length	115864776
Total Number of Non-ATGC Characters	22
Percentage of Non-ATGC Characters	0.00002%

FASTA TO TABLE

SequenceID	Description	Length	GC %
NM_001200	BMP2	2587	46.3
NM_007294	BRCA1	3597	61.9
NM_001145306	CDK6	3445	41
NM_021101	CLDN1	2182	58.1
NM_152727	CPNE2	1595	54.6
NM_130808	CPNE4	4624	60.2
NM_201283	E6FR	3150	48.5
NM_032809	FAM73B	11733	39.8
NM_004448	HER2	7224	41.9

NUCLEOTIDE COMPOSITION

SeqQC is an Ideal tool to bridge the gap between sequencing and analysis
For Windows/Linux/Mac download FREE from www.genotypic.co.in/SeqQC.html

SeqQC can Process up to 8 NGS sequence files or 8 FASTA files simultaneously and outputs a single Report file.