FASTX-Toolkit

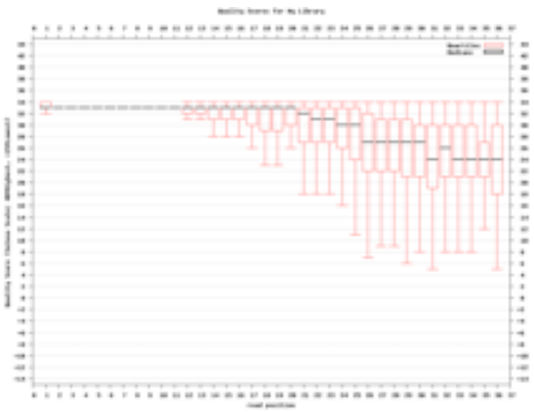
Short-Reads pre-processing Tools

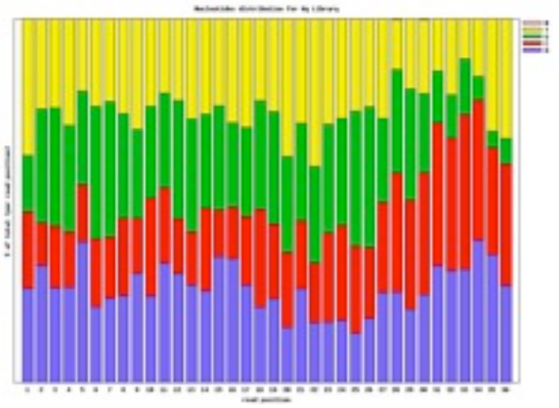
file formats, transformation

http://hannonlab.cshl.edu/fastx_toolkit/

FASTQ

@HWI-EAS82_3_FC204V1AAXX:2:1:900:453
CGAGGGCCTGGGTTCGAACCCCAGAGTTCGTATGC
+HWI-EAS82_3_FC204V1AAXX:2:1:900:453
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
@HWI-EAS82_3_FC204V1AAXX:2:1:752:642
CGAGTTCTACAGTCCGACGATTCGTATGTCGTCTT
+HWI-EAS82_3_FC204V1AAXX:2:1:752:642
hhhhhhhhhhhhhhAhLhhUhhhhghhVChhhdhh





FASTA

>HWI-EAS82_3_FC204V1AAXX:2:1:900:453
CGAGGGCCTGGGTTCGAACCCCAGAGTTCGTATGC
>HWI-EAS82_3_FC204V1AAXX:2:1:752:642
CGAGTTCTACAGTCCGACGATTCGTATGTCGTCTT

Tags

171628	GGGGTGAATGGTATAATGGTTA
109672	GGGGTGAACGGTATAATGGTTA
104596	GCCTGTGTAGCTCAACCGGATAGAGCATA
85933	GGGGTGAATGGTATAATGGTTAG
84602	GGGGTGAAAGGTATAATGGTTA
62107	CGGGTTCGATTCCCGGTGACCGAGCCA
60763	GCCTGTGTAGCTCAACCGGATAGAGCAT
46559	CGGGGTTCGACTCCCCGTGTCGGAACCA



Xmanager Enterprise 3

Microsoft ▸ Windows ▸ Start Menu ▸ Programs ▸ Xmanager Enterprise 3

搜尋 Xmanager Enterprise 3

組合管理 ▾ 開啟 新增資料夾

我的最愛
下載
桌面
最近的位置

名稱	修改日期	類型	大小
Xbrowser	2010/12/13 下午 ...	捷徑	2 KB
Xconfig	2010/12/13 下午 ...	捷徑	2 KB
Xftp	2010/12/13 下午 ...	捷徑	2 KB
Xlpd	2010/12/13 下午 ...	捷徑	2 KB
Xmanager - Broadcast	2010/12/13 下午 ...	捷徑	2 KB
Xmanager - Passive			
Xshell			
Xstart			

Xstart 修改
捷徑

Xstart

Session: MAC

Host: 163.25.92.42

Protocol: SSH

User Name: root

Password: ••••••••••

Save

New...
Save
Save As...
Delete
Shortcut...
Run
Close
Help
Advanced...

root@CLC_WB:~

[root@CLC_WB ~]#

FASTQ-to-FASTA

```
$ fastq_to_fasta -h
usage: fastq_to_fasta [-h] [-r] [-n] [-v] [-z] [-i INFILE] [-o OUTFILE]

version 0.0.6
   [-h]         = This helpful help screen.
   [-r]         = Rename sequence identifiers to numbers.
   [-n]         = keep sequences with unknown (N) nucleotides.
         Default is to discard such sequences.
   [-v]         = Verbose - report number of sequences.
         If [-o] is specified,  report will be printed to STDOUT.
         If [-o] is not specified (and output goes to STDOUT),
         report will be printed to STDERR.
         [-z]         = Compress output with GZIP.
         [-i INFILE]  = FASTA/Q input file. default is STDIN.
         [-o OUTFILE] = FASTA output file. default is STDOUT.
```

```
$ fastq_to_fasta -i test.fq -o test.fasta -v

$ fastq_to_fasta -i test.fq -o test.fasta.gz -v -r -n -z

$ zcat test.fasta.gz | more
```

Hands-on

```
fastq_to_fasta -i test.fq -o test.fasta.gz -v -r -n
-z
```

FASTX Statistics

```
$ fastx_quality_stats -h
usage: fastx_quality_stats [-h] [-i INFILE] [-o OUTFILE]

version 0.0.6 (C) 2008 by Assaf Gordon (gordon@cshl.edu)
   [-h] = This helpful help screen.
   [-i INFILE]  = FASTA/Q input file. default is STDIN.
                  If FASTA file is given, only nucleotides
          distribution is calculated (there's no quality info).
   [-o OUTFILE] = TEXT output file. default is STDOUT.

The output TEXT file will have the following fields (one row per column):
     column   = column number (1 to 36 for a 36-cycles read solexa file)
     count    = number of bases found in this column.
     min      = Lowest quality score value found in this column.
     max      = Highest quality score value found in this column.
     sum      = Sum of quality score values for this column.
     mean     = Mean quality score value for this column.
     Q1   = 1st quartile quality score.
     med  = Median quality score.
     Q3   = 3rd quartile quality score.
     IQR  = Inter-Quartile range (Q3-Q1).
     lW   = 'Left-Whisker' value (for boxplotting).
     rW   = 'Right-Whisker' value (for boxplotting).
     A_Count  = Count of 'A' nucleotides found in this column.
     C_Count  = Count of 'C' nucleotides found in this column.

     G_Count  = Count of 'G' nucleotides found in this column.
     T_Count  = Count of 'T' nucleotides found in this column.
     N_Count = Count of 'N' nucleotides found in this column.
     max-count = max. number of bases (in all cycles)
```

fastx_quality_stats

$ fastx_quality_stats ????

column	count	min	max	sum	mean	Q1	med	Q3	IQR	MV	rMV	A_Count	C_Count	G_Count	T_Count	N_Count	Max_count
1	4659813	-4	40	183053022392.8		40	40	40	0	40	40	431907	619813	2866727	741366	0	4659813
2	4659813	-4	40	183369781393.5		40	40	40	0	40	40	451496	1234727	2554555	419035	0	4659813
3	4659813	-4	40	181541467389.6		40	40	40	0	40	40	478287	747076	2854304	580146	0	4659813
4	4659813	-4	40	182737780392.2		40	40	40	0	40	40	659161	492123	2443872	1064657	0	4659813
5	4659813	-5	40	181784004390.1		40	40	40	0	40	40	409425	1280849	896894	2072645	0	4659813
6	4659813	-5	40	183193007393.1		40	40	40	0	40	40	871349	851611	1157630	1779223	0	4659813
7	4659813	-4	40	181504436389.5		40	40	40	0	40	40	1676856	725446	1460894	796617	0	4659813
8	4659813	-4	40	176129504378.0		40	40	40	0	40	40	1731485	1035433	495379	1397516	0	4659813
9	4659813	-5	40	168153481360.9		40	40	40	0	40	40	2372313	924023	626553	736924	0	4659813
10	4659813	-4	40	178493805383.0		40	40	40	0	40	40	548872	617066	3056855	437020	0	4659813
11	4659813	-5	40	178877896383.9		40	40	40	0	40	40	424669	1834593	1369146	1031405	0	4659813
12	4659813	-4	40	178828012383.8		40	40	40	0	40	40	747522	866581	768617	2277093	0	4659813
13	4659813	-5	40	173315028 37.19		40	40	40	0	40	40	1299532	1846594	413958	1097954	1775	4659813
14	4659813	-5	40	168961327362.6		40	40	40	0	40	40	2119311	797100	508790	1232122	2490	4659813
15	4659813	-5	40	162462960348.6		35	40	40	5	28	40	2226721	1212239	899328	320742	783	4659813

FASTQ Quality Chart

$ fastq_quality_boxplot_graph.sh -h

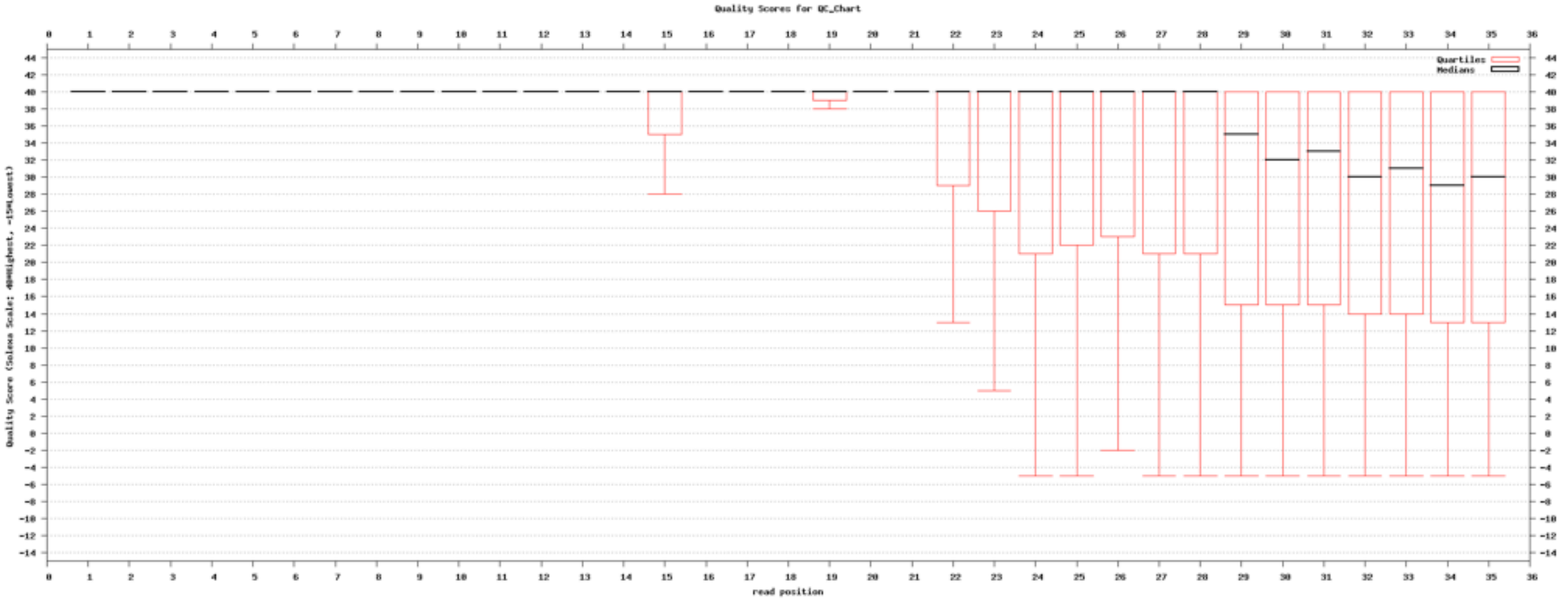
```
Solexa-Quality BoxPlot plotter
Generates a solexa quality score box-plot graph

Usage: /usr/local/bin/fastq_quality_boxplot_graph.sh [-i INPUT.TXT] [-t TITLE] [-p] [-o OUTPUT]

  [-p]           - Generate PostScript (.PS) file. Default is PNG image.
  [-i INPUT.TXT] - Input file. Should be the output of "solexa_quality_statistics" program.
  [-o OUTPUT]    - Output file name. default is STDOUT.
  [-t TITLE]     - Title (usually the solexa file name) - will be plotted on the graph.
```

$fastq_quality_boxplot_graph.sh -i test.stat -o qc.png -t QC_Chart

FASTQ Quality Chart



FASTA/Q Nucleotide Distribution

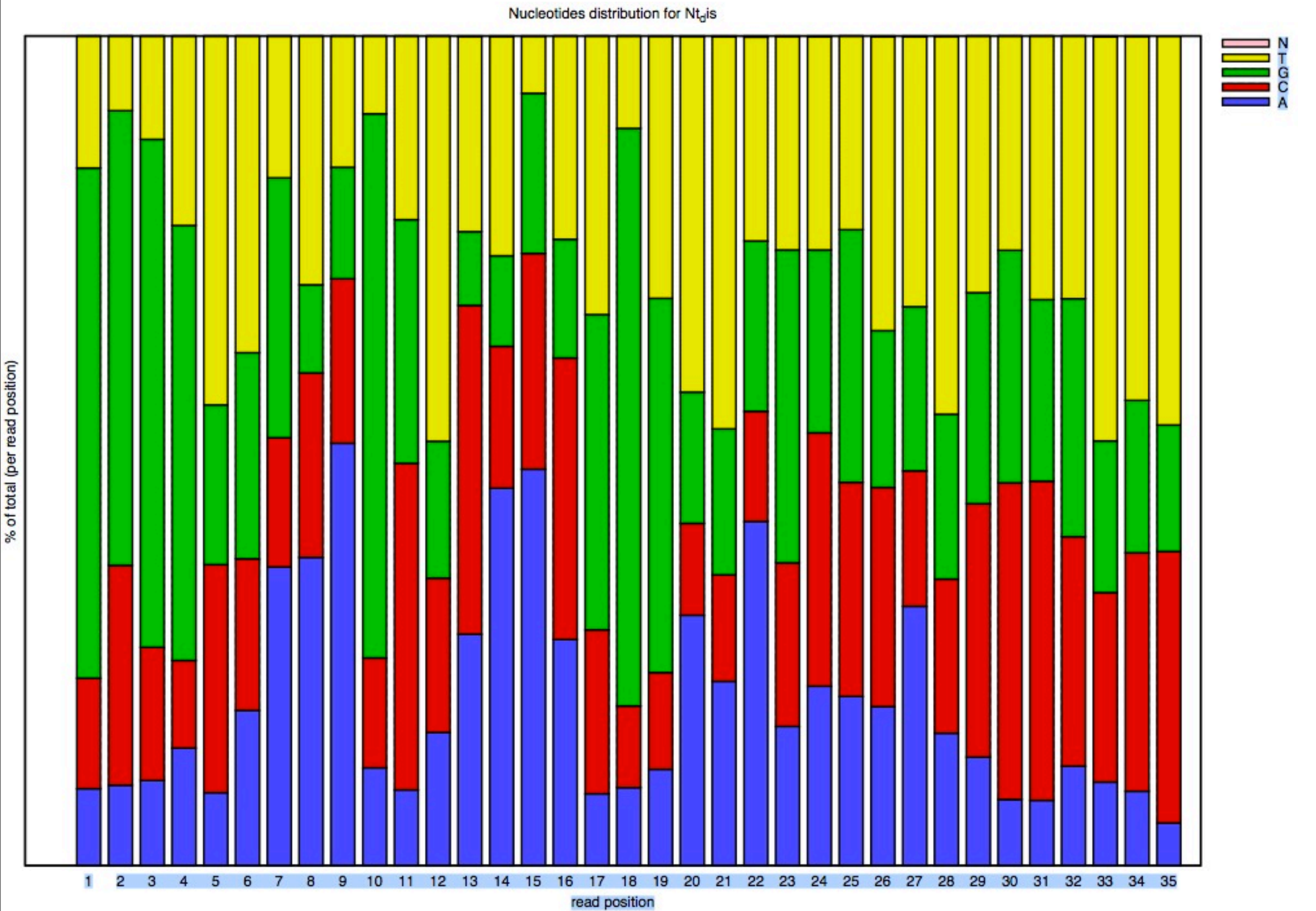
$fastx_nucleotide_distribution_graph.sh -h

```
FASTA/Q Nucleotide Distribution Plotter

Usage: /usr/local/bin/fastx_nucleotide_distribution_graph.sh [-i INPUT.TXT] [-t TITLE] [-p] [-o
OUTPUT]

  [-p]           - Generate PostScript  (.PS) file. Default is PNG image.
  [-i INPUT.TXT] - Input file. Should be the output of "fastx_quality_statistics" program.
  [-o OUTPUT]    - Output file name. default is STDOUT.
  [-t TITLE]     - Title - will be plotted on the graph.
```

$fastx_nucleotide_distribution_graph.sh -i test.stat -p -o -t


Nucleotides distribution for Ntgis
N
T
G
C
A
% of total (per read position)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35
read position

FASTA/Q Clipper

$ fastx_clipper -h

```
usage: fastx_clipper [-h] [-a ADAPTER] [-D] [-l N] [-n] [-d N] [-c] [-C] [-o] [-v] [-z] [-i
INFILE] [-o OUTFILE]

version 0.0.6
   [-h]         = This helpful help screen.
   [-a ADAPTER] = ADAPTER string. default is CCTTAAGG (dummy adapter).
   [-l N]       = discard sequences shorter than N nucleotides. default is 5.
   [-d N]       = Keep the adapter and N bases after it.
                  (using '-d 0' is the same as not using '-d' at all. which is the default).
          [-c]         = Discard non-clipped sequences (i.e. - keep only sequences which contained
the adapter).
          [-C]         = Discard clipped sequences (i.e. - keep only sequences which did not
contained the adapter).
          [-k]         = Report Adapter-Only sequences.
          [-n]         = keep sequences with unknown (N) nucleotides. default is to discard such
sequences.
          [-v]         = Verbose - report number of sequences.
                         If [-o] is specified,  report will be printed to STDOUT.
                         If [-o] is not specified (and output goes to STDOUT),
                         report will be printed to STDERR.
          [-z]         = Compress output with GZIP.
          [-D]      = DEBUG output.
          [-i INFILE]  = FASTA/Q input file. default is STDIN.
          [-o OUTFILE] = FASTA/Q output file. default is STDOUT.
```

FASTA/Q Clipper

$ fastx_clipper -a TCGTATGCCGTCTTCTGCTTG -l 15 -c -v -i test.fasta
-o test.adaptor_trimmed.fasta

Clipping Adapter: TCGTATGCCGTCTTCTGCTTG
Min. Length: 15
Non-Clipped reads - discarded.
Input: 4659813 reads.
Output: 4472584 reads.
discarded 0 too-short reads.
discarded 55200 adapter-only reads.
discarded 126710 non-clipped reads.
discarded 5319 N reads.

FASTA/Q Clipper

>1

CGAGGGCCTGGGTTCGAACCCCAGAGT

>2

CGAGTTCTACAGTCCGACGAT

>3

AGTTCTACAGTCCGACGGATC

>4

AGTTCTACAGTCCGACGATC

>5

GGGGCTGAAGCATAATGGCATTGCTG

>1

CGAGGGCCTGGGTTCGAACCCCAGAGT**TCGTATGC**

>2

CGAGTTCTACAGTCCGACGAT**TCGTATGTCGTCTT**

>3

AGTTCTACAGTCCGACGGATC**TCGTATGCCGTCTT**

>4

AGTTCTACAGTCCGACGATC**TCGTATGCCGTCTTC**

>5

GGGGCTGAAGCATAATGGCATTGCTG**TCGTATGCC**

FASTA Clipping Histogram

