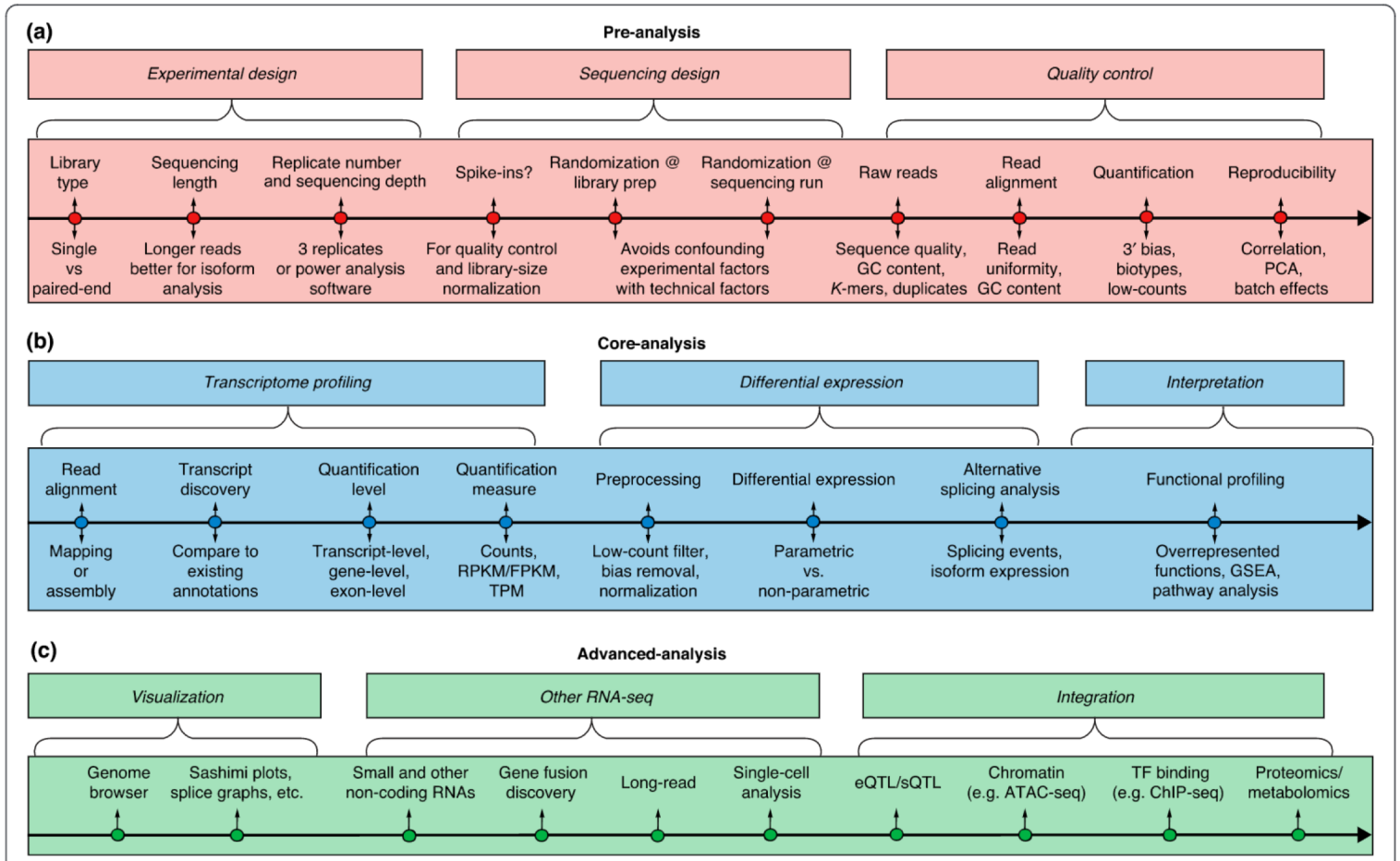


Fig. 1 A generic roadmap for RNA-seq computational analyses. The major analysis steps are listed above the lines for pre-analysis, core analysis and advanced analysis. The key analysis issues for each step that are listed below the lines are discussed in the text. **a** Preprocessing includes experimental design, sequencing design, and quality control steps. **b** Core analyses include transcriptome profiling, differential gene expression, and functional profiling. **c** Advanced analysis includes visualization, other RNA-seq technologies, and data integration. Abbreviations: *ChIP-seq* Chromatin immunoprecipitation sequencing, *eQTL* Expression quantitative loci, *FPKM* Fragments per kilobase of exon model per million mapped reads, *GSEA* Gene set enrichment analysis, *PCA* Principal component analysis, *RPKM* Reads per kilobase of exon model per million reads, *sQTL* Splicing quantitative trait loci, *TF* Transcription factor, *TPM* Transcripts per million

Library preparation

-**Type of library**:
polyA selected
Exon-capturing
ribosomal depletion miRNAs

-**Sequencing length**:
50, 75, 100 nts

-**SE versus PE**:
Gene-level SE 50, 75 nts
"If only a list of DEGs is desired, 50 bp single-end reads would be sufficient for most studies" Chhangawala et al. Genome Biology (2015) 16:131
Isoform-level PE 50, 75 nts
"for splicing detection, the longest reads possible should be used, including using paired-end reads" Chhangawala et al. Genome Biology (2015) 16:131
exon-capturing, PE 50, 75 nts
miRNAs SE 50

-**Sequencing dept**:
Gene-level 20-30 millions
Isoform-level 80-100 millions
rRNA depletion 80-100 millions miRNAs 1-3 millions

Coverage

- On a sequence basis **coverage** is the average number of reads representing a given nucleotide in the reconstructed sequence.
- On a genome basis, it means that, on average, each base has been sequenced a certain number of times (10X, 20X...)

The Lander/Waterman equation is a method for computing coverage .

The general equation is:

$$C = LN / G$$

- C stands for coverage
- G is the haploid genome length
- L is the read length
- N is the number of reads

So, if we take one lane of single read human sequence with v3 chemistry, we get

C = (100 bp)*(189×10^6)/(3×10^9 bp) = 6.3 This tells us that each base in the genome will be sequenced between six and seven times on average.

Coverage

Estimating the number of times a base is expected to be sequenced. Lander and Waterman made two assumptions about the sequencing:

- Reads will be distributed randomly across the genome

- Overlap detection doesn't vary between reads.

Based upon these two assumptions, they reached the conclusion that the number of times a base is sequenced follows a **Poisson distribution**.
The Poisson distribution can be used to model any discrete occurrence given an average number of occurrences.
The probability function is the following:

$$P(Y=y) = (C^y * e^{-C})/y!$$

- y is the number of times a base is read
- C stands for coverage

We can use the Poisson distribution to compute the probability of a **base being sequenced a certain number** of times

Coverage

$$P(Y=y) = (C^y * e^{-C})/y!$$

We can use the **coverage** **as the average number of occurrences** and **y** **as the exact number of times a base is sequenced**, and then compute the probability that would happen:

$$P(Y=3) = (6.3^3 * e^{-6.3})/3! = 0.077$$

Of course, this is the value for exactly 3 times.

It probably is more interesting to see the probability the base is sequenced 3 times or less, as most SNP callers require at least four calls at a base position to call SNPs.
We can determine this probability simply by summing up the probabilities for

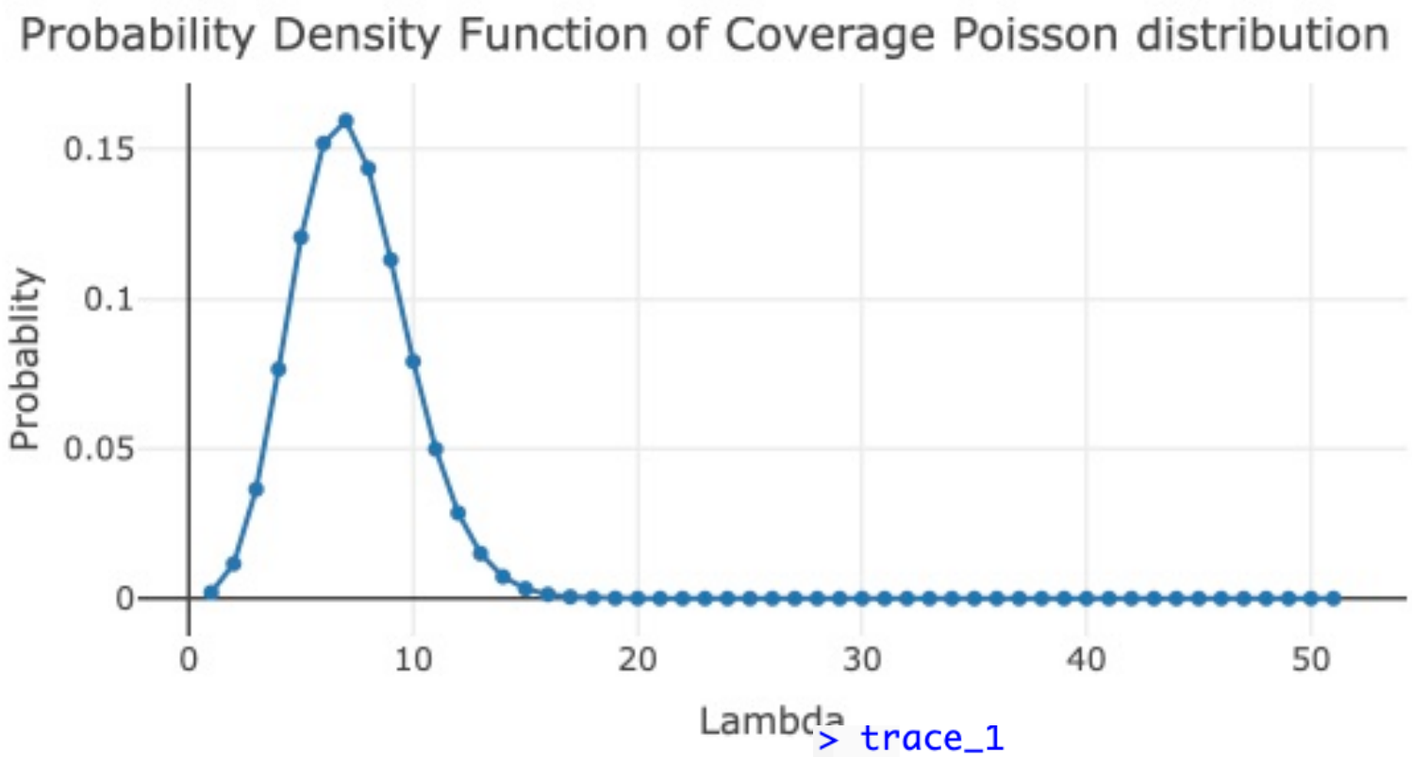
Y=2, Y=1, and Y=0:
P(Y<=3) = P(Y=3) + P(Y=2) + P(Y=1) + P(Y=0) = 0.077 + 0.036 + 0.012 + 0.002 = 0.127

So **we see that about 12.7% of the bases in the genome will be covered by three or fewer reads**

Coverage

$P(Y<=8) = P(Y=8) + \ldots + P(Y=0) = 0.113+0.143+0.159+0.152+0.120 +0.077 + 0.036 + 0.012 + 0.002 =$ **0.66**

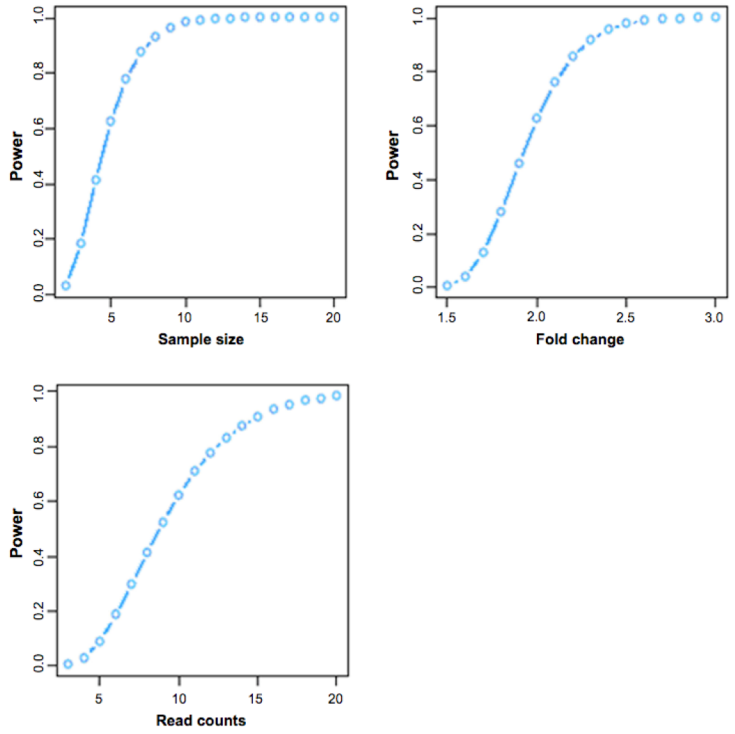
So **we see that about 66% of the bases in the genome will be covered by eight or fewer reads**




```
> trace_1
 [1] 2.000000e-03 1.156872e-02 3.644147e-02 7.652708e-02 1.205302e-01 1.518680e-01
 [7] 1.594614e-01 1.435153e-01 1.130183e-01 7.911279e-02 4.984105e-02 2.854533e-02
[13] 1.498630e-02 7.262591e-03 3.268166e-03 1.372630e-03 5.404729e-04 2.002929e-04
[19] 7.010252e-05 2.324452e-05 7.322024e-06 2.196607e-06 6.290284e-07 1.722991e-07
[25] 4.522851e-08 1.139758e-08 2.761722e-09 6.444019e-10 1.449904e-10 3.149792e-11
[31] 6.614563e-12 1.344250e-12 2.646492e-13 5.052394e-14 9.361789e-15 1.685122e-15
[37] 2.948963e-16 5.021208e-17 8.324634e-18 1.344749e-18 2.117979e-19 3.254456e-20
[43] 4.881684e-21 7.152234e-22 1.024070e-22 1.433698e-23 1.963543e-24 2.631983e-25
[49] 3.454477e-26 4.441471e-27 5.596253e-28
```

Library preparation

-**Replicates**:
Cell lines: starting from 3
Inbreed animals: starting from 6
Humans: pilot experiment starting from 6
To evaluate the statistical power of a RNAseq experiment we need to count RNA features, i.e. genes, transcripts, miRNAs, etc.



Examples of the power curves produced by RNAseqPS.

Table 1 Statistical power to detect differential expression varies with effect size, sequencing depth and number of replicates

	Replicates per group		
	3	5	10
Effect size (fold change)			
1.25	17 %	25 %	44 %
1.5	43 %	64 %	91 %
2	87 %	98 %	100 %
Sequencing depth (millions of reads)			
3	19 %	29 %	52 %
10	33 %	51 %	80 %
15	38 %	57 %	85 %

Example of calculations for the probability of detecting differential expression in a single test at a significance level of 5 %, for a two-group comparison using a Negative Binomial model, as computed by the RNASeqPower package of Hart et al. [190]. For a fixed within-group variance (package default value), the statistical power increases with the difference between the two groups (effect size), the sequencing depth, and the number of replicates per group. This table shows the statistical power for a gene with 70 aligned reads, which was the median coverage for a protein-coding gene for one whole-blood RNA-seq sample with 30 million aligned reads from the GTEx Project [214]

FASTA file

A sequence in FASTA format begins with a single-line description, followed by lines of sequence data.

The description line (defline) is distinguished from the sequence data by a greater-than (">") symbol at the beginning. It is recommended that all lines of text be shorter than 80 characters in length.

The sequence (aminoacid or nucleotide) is report by lines which length is equal to 80

An example sequence in FASTA format is:

```
>P01013 GENE X PROTEIN (OVALBUMIN-RELATED)
QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAEDTREMPFHVTKQESKPVQMMCMNNSFNVATLPAE
KMKILELPFASGDLSMLVLLPDEVSDLERIEKTINFEKLTEWTNPNTMEKRRVKVYLPQMKIEEKYNLTS
VLMALGMTDLFIPSANLTGISSAESLKISQAVHGAFMELSEDGIEMAGSTGVIEDIKHSPESEQFRADHP
FLFLIKHNPTNTIVYFGRYWSP
```

Input Data Structure

Fastq File

```
@D44TDFP1_1:1:1101:1320:1948/1
NGGAGGCAGAGGCAGGTGGATTTCTGAGTTCAAGGCCAGCCTGGTCTACAAAGTGAGTNCCAGGACGGCCAGGGCTATACAGAGAAACAGAGAAACCCTGT
+
#1=DDDDDHFHHHIIIAEHGHIIGIIGHGHHIIIIGIIGHIIIIFHIIIIIIIFHIIG#-5@EHHHECCBBBBBBBBCECECCCCCCCCCCCCCCCABBCC
@D44TDFP1_1:1:1101:1817:1955/1
NGGGTTGGGGAGGAGAAGATGACGACATTTTTAACAGATTAGTTCATAAAGGCATGTCNATATCACGTCCAAATGCTGTAGTAGGGAGGTGTCGAATGATC
+
#1=DBDDFHHHHHGIIJJJJJIJHGIJJJJJJJJJJJIJJJJIIJJJJJJJJJJJHHH#-;BFAEEDEDDDDDDDCDDDDDDEEDDCBD<BCDDDDDDDDD
@D44TDFP1_1:1:1101:1790:1968/1
GAGGCCAGGTTGAGGATTTTGGAGGACAGAGGGATAAGAAAAATAAGTGGAACAGGAANGGCATTAGCAAAAGCAGAAAAGTATGAACACAAAAGTGAAGT
+
CCCFFFFFHFFHHJJIJHJJJJGHIIJJIGHJJHIJJJJJJJJJJJJGIJJJJJJJJJ#-;EHHHHFFFFFFEEECEDDDDACDEDEDDDDDDDDDDEDC
@D44TDFP1_1:1:1101:1870:1994/1
AGGGGCTGAGTGACTCGGGGCCACATAGGCAGCAAGGAGCAAGGGGCCTGAGCAAGAGNTACCATATTTACCTCAGTGTGTGAAGATCATTTGCCCAGGCT
+
CCCFFFFFHHHHHJJJJJJJJJIJJIJJJIIIIJIIJIGIIIIIJJFGHHHFFFFFEE#,5=BDDDDEFEEDDDDDDDEDDDDEDDDDDDEEEDDDDDDDD
@D44TDFP1_1:1:1101:2070:1923/1
NGCAGNCCNAGGTCTGAGTTCCAAGGACANGTATGTGAAAGGCCTGATTGAGGGCAAANCGGATCCCTACGCGCTCGTCCGTGTGGGCACCCAGACGTTCT
+
#0;@@#2@#2=?=@@@?@?@@@@@@@@@@#1:???>??????????????????????#-;?????????==<<<<<:<<:<<<<<:<<<<<<<<<:<<<<
```

Input Data Structure

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
```

EAS139	the unique instrument name
136	the run id
FC706VJ	the flowcell id
2	flowcell lane
2104	tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 *(paired-end or mate-pair reads only)*
Y	Y if the read is filtered, N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	index sequence

Input Data Structure

FASTQ: Phred base-call qualities

quality score Q_{phred}	error prob. p	characters
0 .. 9	1 .. 0.13	!"#$%&'()*
10 .. 19	0.1 .. 0.013	+,-./01234
20 .. 29	0.01 .. 0.0013	56789:;<=>
30 .. 39	0.001 .. 0.00013	?@ABCDEFGH
40	0.0001	I

- If p is the probability that the base call is wrong, the Phred score is:
- $Q = -10 \log_{10} p$
- The score is written with the character whose ASCII code is Q +33 (Sanger Institute standard).

Input Data Structure

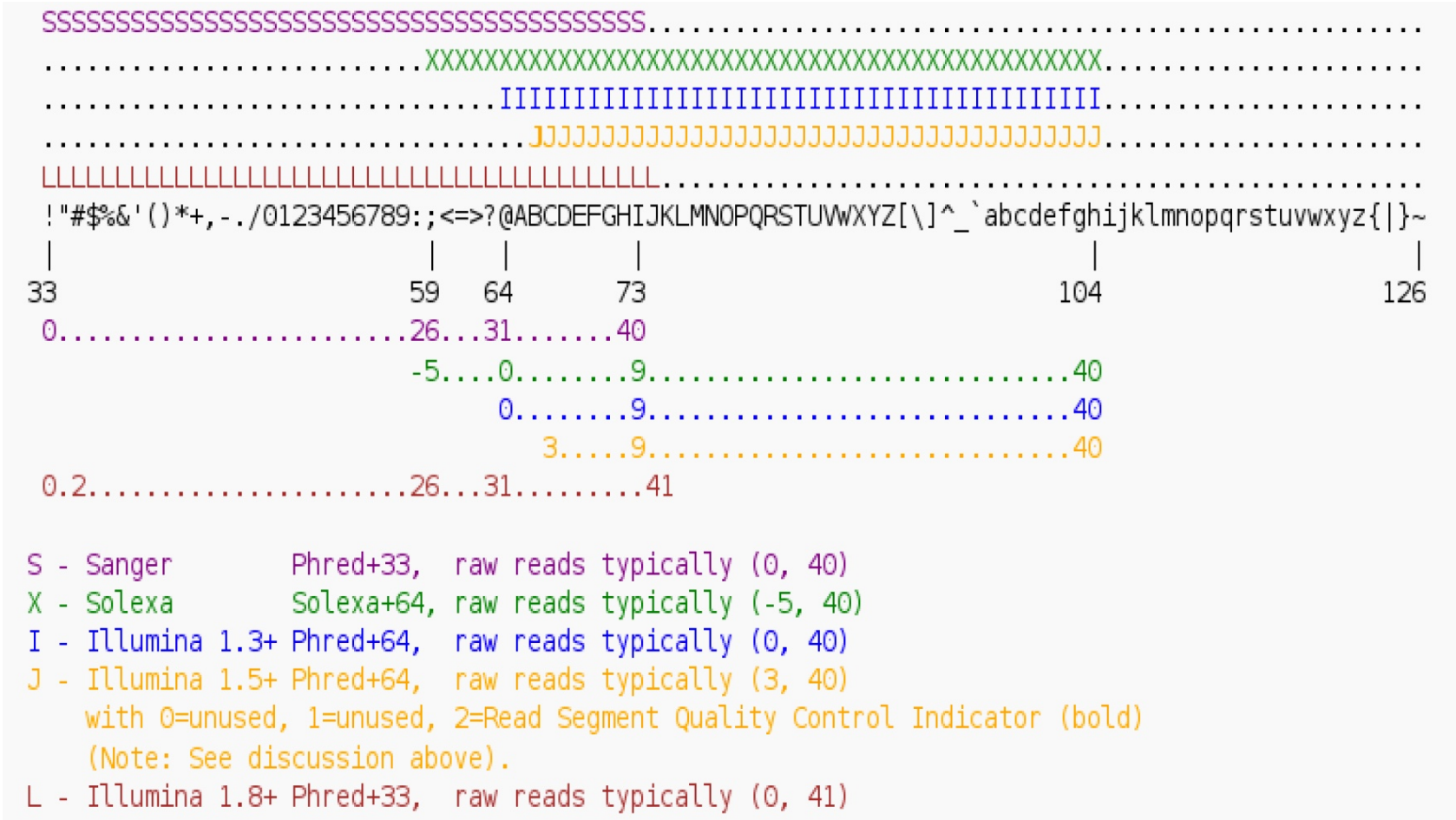


Figure 8: Phred score