ARSDA: a Quick Start

Xuhua Xia xxia@uottawa.ca
University of Ottawa

ARSDA stands for Analyzing RNA-Seq Data. It is a 64-bit Windows program, but can run on Linux or Mac computers that have Mono (http://www.mono-project.com) installed. I wish to highlight its two strengths at the beginning. First, it can dramatically reduce RNA-Seq file size without losing any sequence information. This is possible because many sequence reads from a transcriptomic study are identical. Take for example the characterized transcriptomic data for *Escherichia coli* K12 in the file SRR1536586.sra (where SRR1536586 is the SRA sequence file ID in NCBI/DDBJ/EBI). The file contains 6,503,557 sequences of 50 nt each, but 195310 sequences are all identical (TGTTATCACGGGAGACACACGGCGGGTGCTAACGTCCGTCGTGAAGAGGG), all mapping to sites 929-978 in *E. coli* 23S rRNA genes. A more dramatic example is the file SRR922264.sra (from another *E. coli* transcriptomic study) in which one forward read has 1,606,515 identical copies stored in the file as separate entries (The file contains 9,690,570 forward reads and same number of reverse reads). The current approach at NCBI/DDBJ/EBI stores individual reads in SRA or FASTQ files separate entries. There is no sequence information lost if all these identical sequences are stored by a single sequence with a sequence ID such as UniqueSeqX_1606515 (i.e., SequenceID_CopyNumber). Such storage scheme also leads to dramatic saving in analysis time. At present, all software packages for RNA-Seq analysis will take these identical reads and search them individually against the *E. coli* genome (or coding sequences). The SequenceID_CopyNumber storage scheme reduces all these separate searches of identical sequences to a single one. The new FASTQ+ and FASTA+ formats generated and used by ARSDA differ from the corresponding FASTQ and FASTA file formats only in the use of SequenceID_CopyNumber as sequence ID.

The conversion of original RNA-Seq files (which typically come in .fastq or .sra format) to FASTA+ format is time consuming, but it needs to be done only once. Ideally this should be done in major data centers. At the moment, I am in the process of converting RNA-Seq files from representative transcriptomic studies for model organism to FASTA+ files and deposit them at coevol.rdc.uottawa.ca, where you will find 28 converted files for *E. coli*, 24 for *Bacillus subtilis,* and 44 for the yeast *Saccharomyces cerevisiae.* If you want to use ARSDA to convert your own .sra or .fasta files, you should ideally have 16GB or even 32GB of RAM (not disk space). This memory requirement is essential because ARSDA builds a large sequence dictionary to count the copy number of unique reads in RNA-Seq files. All other functions of ARSDA works well with ordinary 64-bit computers with 8GB or even 4GB of RAM.

The second strength in ARSDA is in its explicit and rational allocation of reads to paralogous genes leading to more accurate quantification of gene expression. This method is missing in other software packages for RNA-Seq data analysis (Deng, *et al.*, 2014; Dobin, *et al.*, 2013; Langmead, Hansen and Leek, 2010; Langmead and Salzberg, 2012; Langmead, *et al.*, 2009; Roberts, Schaeffer and Pachter, 2013; Roberts, *et al.*, 2011; Trapnell, Pachter and Salzberg, 2009; Trapnell, *et al.*, 2012). The rationale for the allocation will be detailed in a publication, but you may request a draft paper from me any time.

This quick-start guide has three parts. First, it guides you through the conversion of FASTQ to FASTQ+ or FASTA+ format. Second, it demonstrates a variety of data quality visualization functions. Third, it takes you through the process of generating gene expression.

Some of ARSDA's functions make use of several NCBI programs for sequence matching and for processing SRA files (sratoolkit). These programs are included in the

ARSDA distribution for your convenience. Some of ARSDA functions such as gene expression quantification involves reading genomic data in GenBank format and extracting coding sequences, exons, introns, rRNAs and tRNAs, and are better done jointly with DAMBE (Xia, 2013; Xia and Xie, 2001) which features extensive data analysis. Both ARSDA and DAMBE are freely available at dambe.bio.uottawa.ca/Include/software.aspx, and can be installed with just a few mouse clicks.

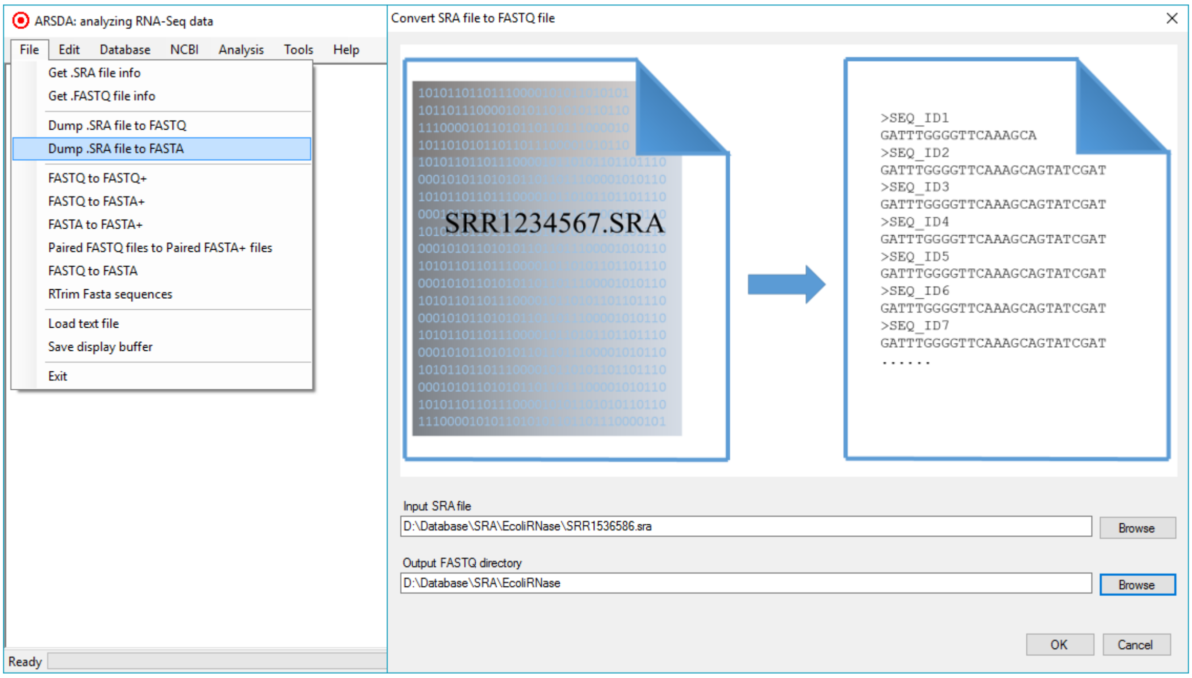
I will use the transcriptomic data in the file SRR1536586.sra downloaded from GenBank to demonstrate the characterization of gene expression in wild-type *E. coli* K-12 by ARSDA. The file is small by RNA-Seq standard, with only 198 MB. It is one of the four data sets with three others being from three *E. coli* K-12 mutants (Pobre and Arraiano, 2015). You can download SRR1536586.sra directly from NCBI Entrez or, alternatively, download within ARSDA by clicking ‘NCBI|Download .SRA files’. All functions in this quick start guide can be performed on a 64-bit computer with 16 GB of RAM when no other memory-hungry programs are running.

The original publication (Pobre and Arraiano, 2015) scratches only the surface of RNA-Seq analysis of this data set. If you wish to compare the gene expression between the wild-type and the mutant, or between mutants, then you should also download the other three SRA files (SRR1536587.sra, SRR1536588.sra, SRR1536589.sra) and repeat what is detailed below for the other three files.

CONVERT .SRA OR .FASTQ FILES TO FASTA+ FILES

This is the only function in ARSDA that has high-memory requirement. Your computer should have 16GB or ideally 32 GB of RAM. Please check coevol.rdc.uottawa.ca site to see if the data you need have already been converted. I am in the process of converting RNA-Seq files from transcriptomic studies to FASTA+ format and deposit them at the web site.

Click ‘File|Dump .SRA file to FASTA’ (Fig. 1) or ‘File|Dump .SRA file to FASTQ’, fill in the two entries by browsing to the input file (e.g., SRR1536586.sra) and output directory, and click the ‘OK’ button. This will generate a either a SRR1536586.fasta or a SRR1536586.fastq file (which is 1.49 GB in size but small by RNA-Seq standard). If the input SRA file contains paired-end reads, then two FASTA files will be generated, one for the forward reads and one for the reverse reads.



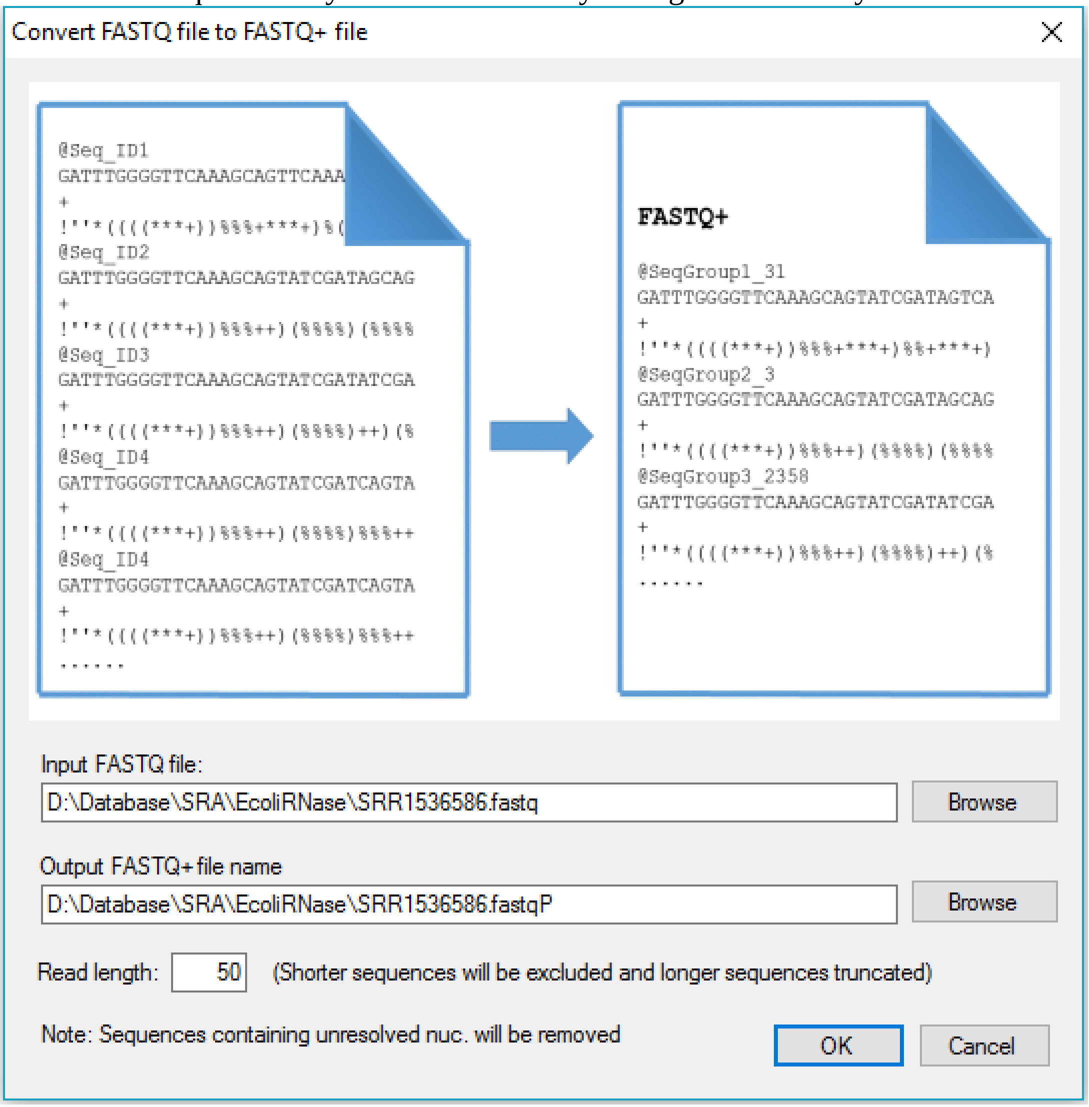
```
>SEQ_ID1
GATTTGGGGTTCAAAGCA
>SEQ_ID2
GATTTGGGGTTCAAAGCAGTATCGAT
>SEQ_ID3
GATTTGGGGTTCAAAGCAGTATCGAT
>SEQ_ID4
GATTTGGGGTTCAAAGCAGTATCGAT
>SEQ_ID5
GATTTGGGGTTCAAAGCAGTATCGAT
>SEQ_ID6
GATTTGGGGTTCAAAGCAGTATCGAT
>SEQ_ID7
GATTTGGGGTTCAAAGCAGTATCGAT
......
```

Input SRA file
D:\Database\SRA\EcoliRNase\SRR1536586.sra

Output FASTQ directory
D:\Database\SRA\EcoliRNase

Fig. 1. ARSDA's main menu system displaying the function for dumping SRA file to FASTQ. Note that the output entry is a directory.

Now to convert the FASTA (or FASTQ) file to FASTQ+ file, click ‘File|FASTA to FASTA+’ or ‘File|FASTQ to FASTQ+’, enter the input and output file names (Fig. 2), and click ‘OK’. The conversion is a rather lengthy process. ARSDA will create a dictionary of unique sequences as well as a count for each unique sequences. This dictionary is necessarily large and is the only function in ARSDA that requires 16GB or more RAM. However, the conversion needs to be done only once for data storage, and the resulting saving in storage space, internet traffic and computation time in downstream data analysis is tremendous. For example, one can use this file to obtain gene expression for coding sequences or tRNAs, and it reduces the computation time from many hours to a few minutes. This conversion will make RNA-Seq data analysis feasible in every biological laboratory.



```
@Seq_ID1
GATTTGGGGTTCAAAGCAGTTCAAA
+
!''*((((***+))%%%+***+)%(
@Seq_ID2
GATTTGGGGTTCAAAGCAGTATCGATAGCAG
+
!''*((((***+))%%%++)(%%%%)(%%%%
@Seq_ID3
GATTTGGGGTTCAAAGCAGTATCGATATCGA
+
!''*((((***+))%%%++)(%%%%)++)(%
@Seq_ID4
GATTTGGGGTTCAAAGCAGTATCGATCAGTA
+
!''*((((***+))%%%++)(%%%%)%%%++
@Seq_ID4
GATTTGGGGTTCAAAGCAGTATCGATCAGTA
+
!''*((((***+))%%%++)(%%%%)%%%++
......
```

```
FASTQ+

@SeqGroup1_31
GATTTGGGGTTCAAAGCAGTATCGATAGTCA
+
!''*((((***+))%%%+***+)%%+***+)
@SeqGroup2_3
GATTTGGGGTTCAAAGCAGTATCGATAGCAG
+
!''*((((***+))%%%++)(%%%%)(%%%%
@SeqGroup3_2358
GATTTGGGGTTCAAAGCAGTATCGATATCGA
+
!''*((((***+))%%%++)(%%%%)++)(%
......
```

Input FASTQ file: D:\Database\SRA\EcoliRNase\SRR1536586.fastq

Output FASTQ+ file name: D:\Database\SRA\EcoliRNase\SRR1536586.fastqP

Read length: 50 (Shorter sequences will be excluded and longer sequences truncated)

Note: Sequences containing unresolved nuc. will be removed

Fig. 2. User interface in ARSDA for converting a FASTQ file to a FASTQ+ file. For a set of N sequences represented as SequenceID_N, the quality score for each site is the average of N quality values. The interface for converting FASTA file to FASTA+ file is the same except that FASTQ will be replaced by FASTA.

The output also include a table showing how many reads are represented only once, twice, etc., and part of the table is replicated in Table 1. Some sequences are represented many times. As I mentioned before, one 50mer mapped to sites 929-978 in *E. coli* 23S rRNA gene is represented 195310 times in the SRR1536586.sra file. The SRA file (and the FASTA or FASTQ file derived from it) lists these 195310 sequences individually. The resulting FASTQ+ file lists them by a single entry (either > S17_195310 in FASTA+ format or @S17_195310 in FASTQ format) where ‘S17’ means that the read is the 17th unique sequence in the read dictionary and it has 195310 identical copies in the FASTQ file. This condensed representation of UniqueSeqID_N leads to dramatic reduction in file size, from the original FASTQ file of 1.49 GB to the new FASTQ+ file of only 114 MB, and the FASTA+ file will be only about 60 MB.

Table 1. Part of read-matching output from ARSDA, with 195310 identical reads matching a segment of large subunit (LSU) rRNA, 86308 identical reads matching another segment of LSU rRNA, and so on. Results generated from ARSDA analysis of the SRR1536586.sra file from GenBank.

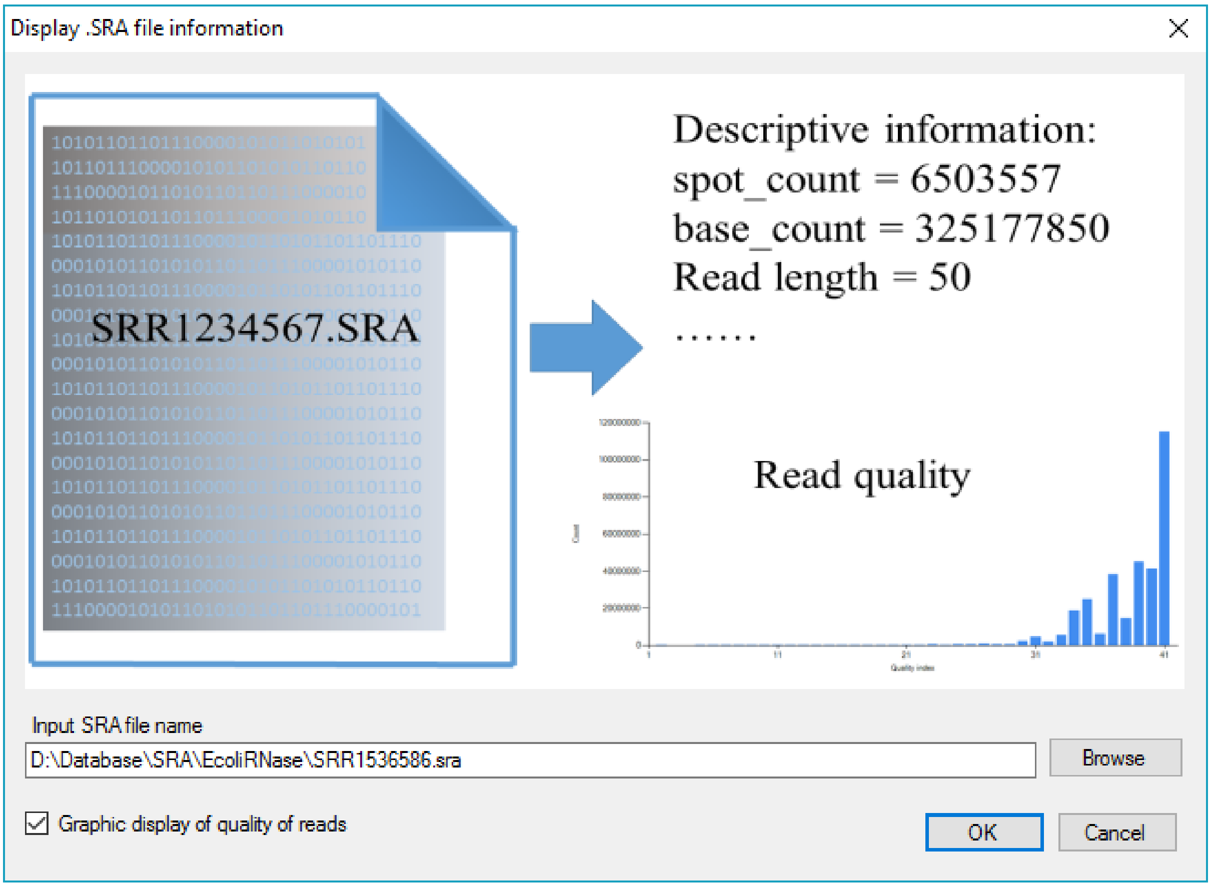
Gene	N_{copy}	Gene	N_{copy}
LSU rRNA	195310	SSU rRNA	30417
LSU rRNA	86308	LSU rRNA	29508
LSU rRNA	58400	5S rRNA	28187
SSU rRNA	47323	LSU rRNA	24982
LSU rRNA	45695	SSU rRNA	23286
LSU rRNA	36258	LSU rRNA	19991
5S rRNA	33674	SSU rRNA	19268

The interface for converting FASTA file to FASTA+ file is similar to that in Fig. 2, except that I have added a ‘Multiple files’ option to the dialog. A transcriptomic study typically generates a number of files (e.g., one for wild type and several for various treatments). It is better to select all of them and let the computer run over night.

THREE WAYS TO VISUALIZE SEQUENCE QUALITY

Global base-calling quality

Data for global base-calling quality is already present in the downloaded SRA file and needs little computation. To visualize the quality report within an SRA file, click ‘File|Get .SRA file info’ and browse to the input SRA file (Fig. 3). Leave the default option of ‘Graphic display of quality of reads’ and click ‘OK’.



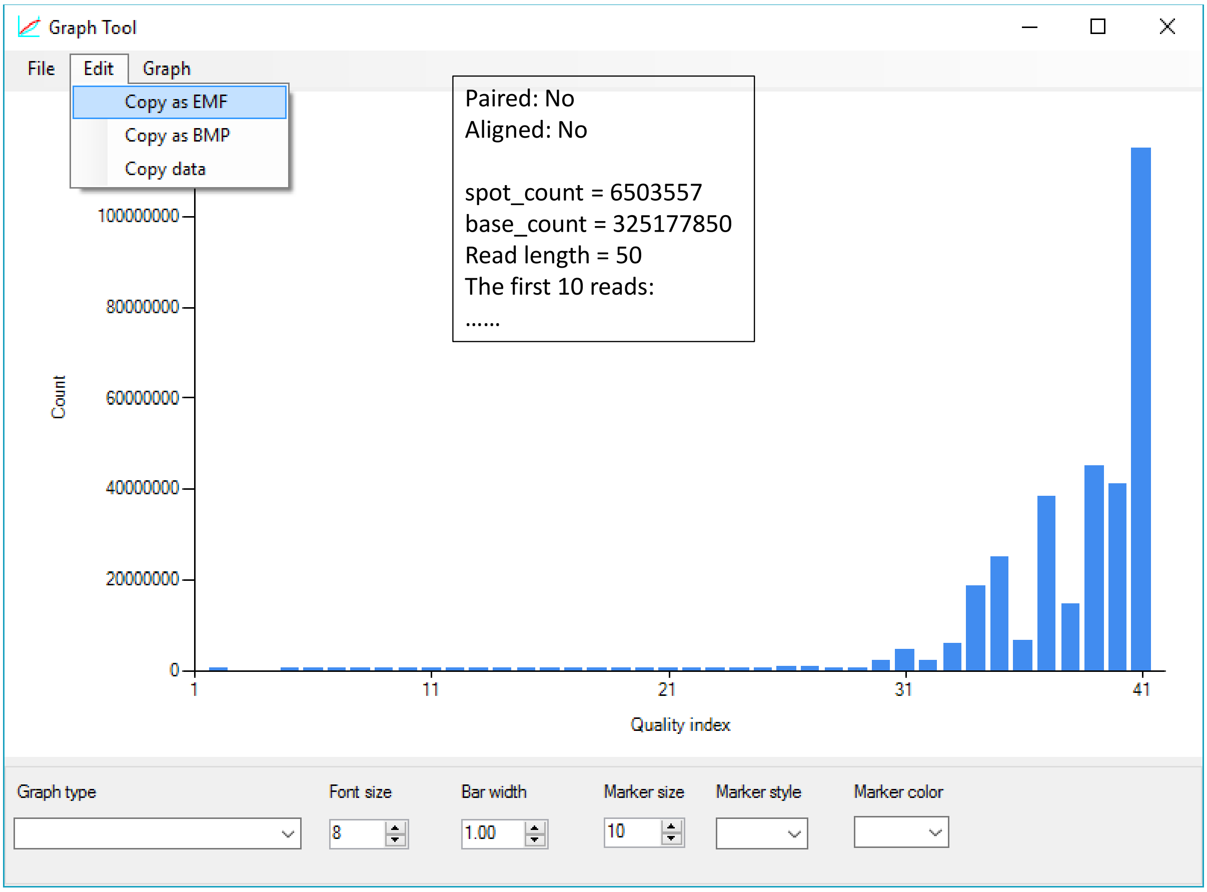
Descriptive information:
spot_count = 6503557
base_count = 325177850
Read length = 50
……

Input SRA file name: D:\Database\SRA\EcoliRNase\SRR1536586.sra

☑ Graphic display of quality of reads

Fig. 3. Input for visualizing global base-calling quality based on information stored within individual SRA files.

The output (Fig. 4) shows the frequency distribution (Y-axis) of base-calling quality (‘Quality index in X-axis). Good quality corresponds to large ‘Quality index’. A ‘Quality index’ of 41 in an .sra file is equivalent to an error probability of base-calling (P) of 0.000079433, i.e., it is equal to -10*log$_{10}$(P). In contrast, base quality in a FASTQ file is represented by symbols from ‘!’ to ‘~’ corresponding to ASCII codes from 33 to 126, so a ‘Quality index’ of 41 in Fig. 4 would be represented by character ‘I’ corresponding to the ASCII value of 73 [=-10log$_{10}$(P) + 32].



```
Paired: No
Aligned: No

spot_count = 6503557
base_count = 325177850
Read length = 50
The first 10 reads:
......
```

Graph type
Font size 8
Bar width 1.00
Marker size 10
Marker style
Marker color

Fig. 4. Output for visualizing global read quality based on information stored within individual SRA files. One can copy and paste high-resolution image to graphic programs such as Microsoft PowerPoint by clicking 'Edit|Copy as EMF'. Alternatively, one may copy and paste the graphic data to EXCEL and re-generate graphs in EXCEL. The inset shows part of the text output.

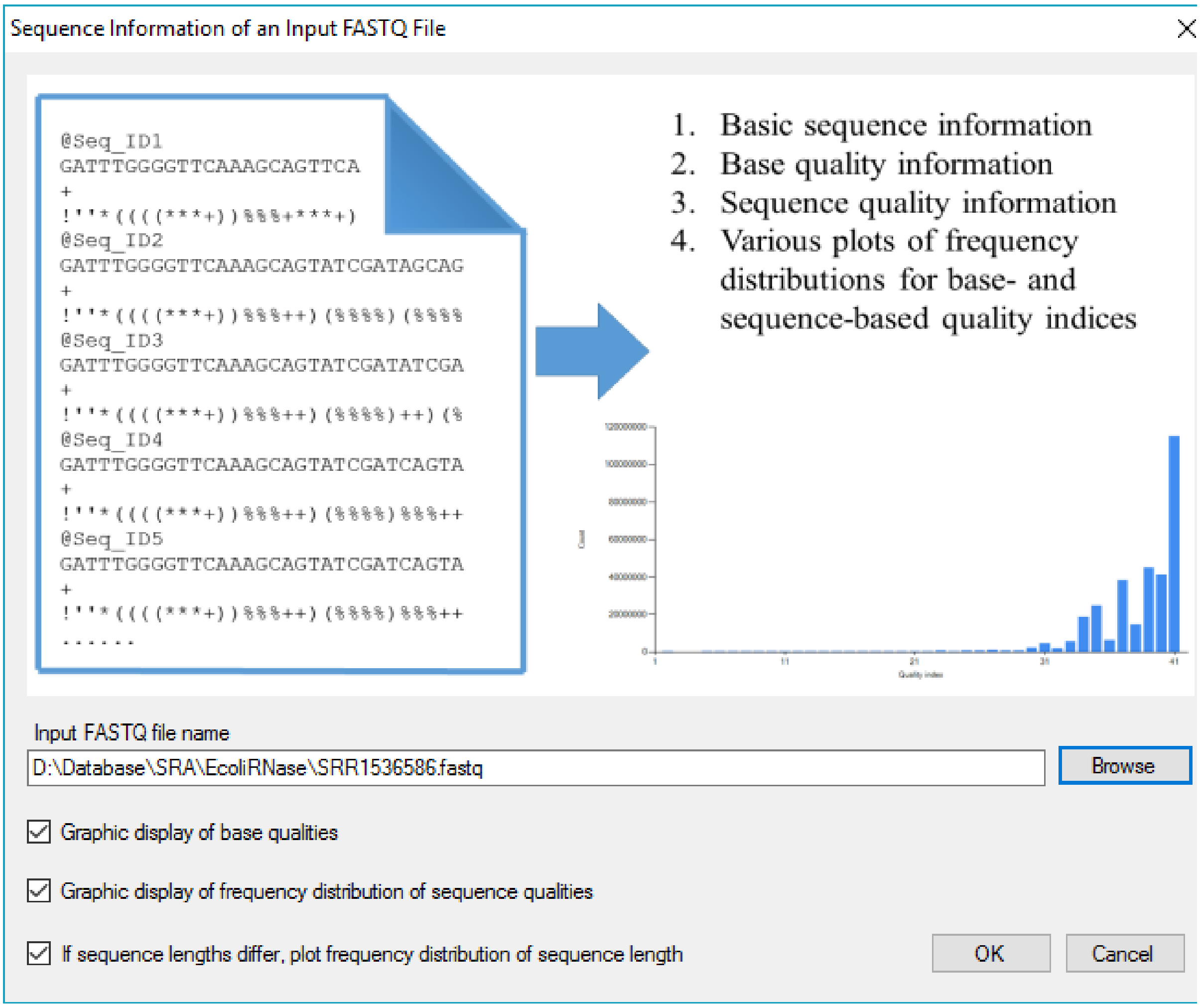
Among a random selection of 10 SRA files, the global base-calling quality in SRR1536586 is the second best. Some files, especially those with long reads of 250 bases, are often quite poor.

Read-specific quality and site-specific quality

Read-specific quality: A read with many low-quality bases is better excluded from the analysis. For this reason, it is important to know how many poor-quality reads there are in the RNA-Seq data and what threshold one should use to exclude them. A read with 50 bases has 50 individual base quality values, and a read quality is simply the average of these 50 values.

Site-specific quality: The sequencing by synthesis step in RNA-Seq by Illumina and the like is particularly error prone so that the base quality decreases rapidly with read length. A researcher with long reads of 250 bases would wish to know whether all bases are good or only the first 150 bases are good. A sequencer manufacturer would want to know the optimal read length to extract so that the sequencer will not waste time to generate long but poor reads (which would be embarrassing to the sequencer manufacturer). Site-specific base quality helps to address this problem.

Both read-specific quality and the site-specific quality can be obtained by clicking ‘File|Get FASTQ file info’, which displays the dialog in Fig. 5. Browse to the input FASTQ file, click ‘OK’, and ARSDA will start a lengthy process of reading and processing the input file. For large FASTQ files, ARSDA read them in chunks so there is no high-memory requirement for this function. For the SRR1536586.fastq, ARSDA may take half an hour before generating the output.



Sequence Information of an Input FASTQ File

```
@Seq_ID1
GATTTGGGGTTCAAAGCAGTTCA
+
!''*((((***+))%%%+****+)
@Seq_ID2
GATTTGGGGTTCAAAGCAGTATCGATAGCAG
+
!''*((((***+))%%%++)(%%%%)(%%%%
@Seq_ID3
GATTTGGGGTTCAAAGCAGTATCGATATCGA
+
!''*((((***+))%%%++)(%%%%)++)(%
@Seq_ID4
GATTTGGGGTTCAAAGCAGTATCGATCAGTA
+
!''*((((***+))%%%++)(%%%%)%%%++
@Seq_ID5
GATTTGGGGTTCAAAGCAGTATCGATCAGTA
+
!''*((((***+))%%%++)(%%%%)%%%++
......
```

1. Basic sequence information
2. Base quality information
3. Sequence quality information
4. Various plots of frequency distributions for base- and sequence-based quality indices

Input FASTQ file name: D:\Database\SRA\EcoliRNase\SRR1536586.fastq

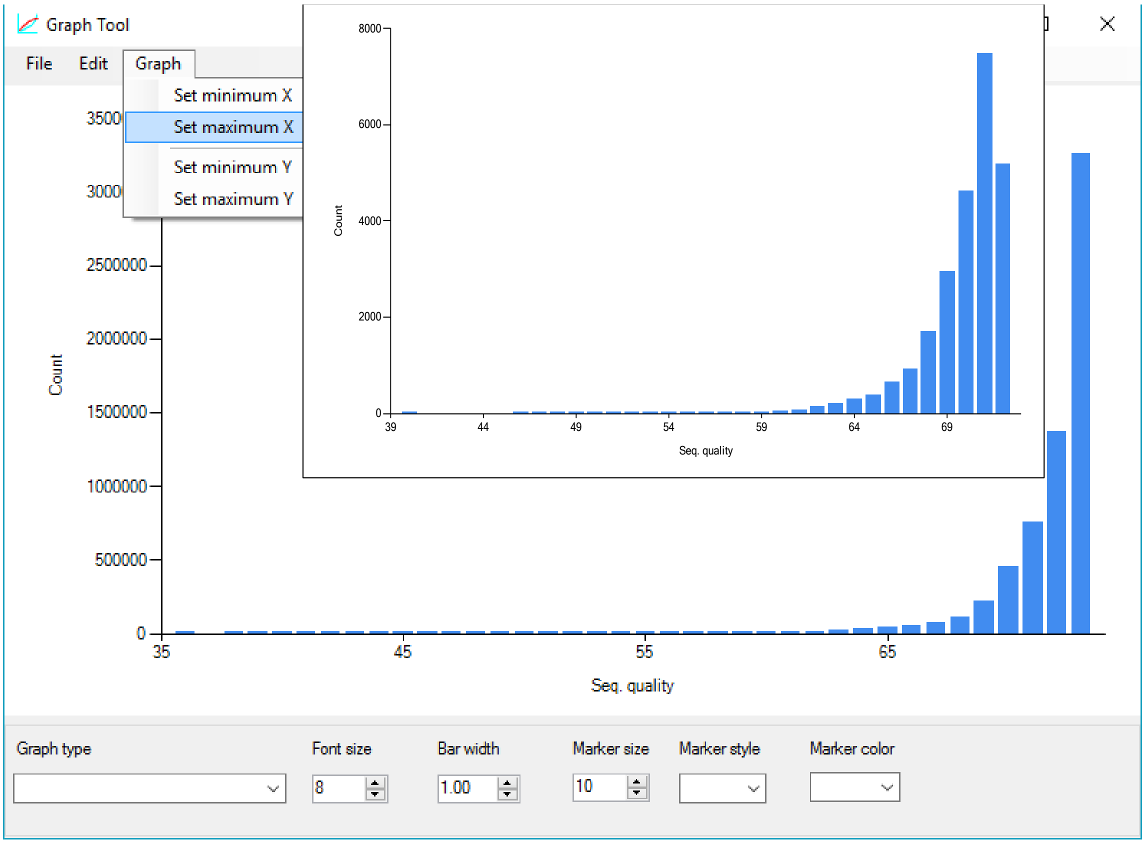
☑ Graphic display of base qualities

☑ Graphic display of frequency distribution of sequence qualities

☑ If sequence lengths differ, plot frequency distribution of sequence length

Fig. 5. Dialog box for accessing read-specific and site-specific quality characterization.

The output is of three parts. The first is the same as in Fig. 4, and will not be repeated here. The second is read-specific quality distribution, which plots reads with and without ambiguous codes separately for sequences (Fig. 6 for SRR1536586.fastq). As I mentioned before, this data set is of high quality, and it is helpful to contrast it with another data set that is of lower quality (Fig. 7 for sequences in the file SRR2056426.sra, which is also for *E. coli*, but is of paired-end reads with read length of 250 nt). In general, quality decreases rapidly with sequence length. For paired-end reads, the reverse read is much worse than the forward read.



Font size 8

Bar width 1.00

Marker size 10

Fig. 6. Frequency distribution of the quality of individual reads for all reads with fully resolved bases in file SRR1536586.sra. The inset is an equivalent plot for sequence reads with at least one unresolved base.

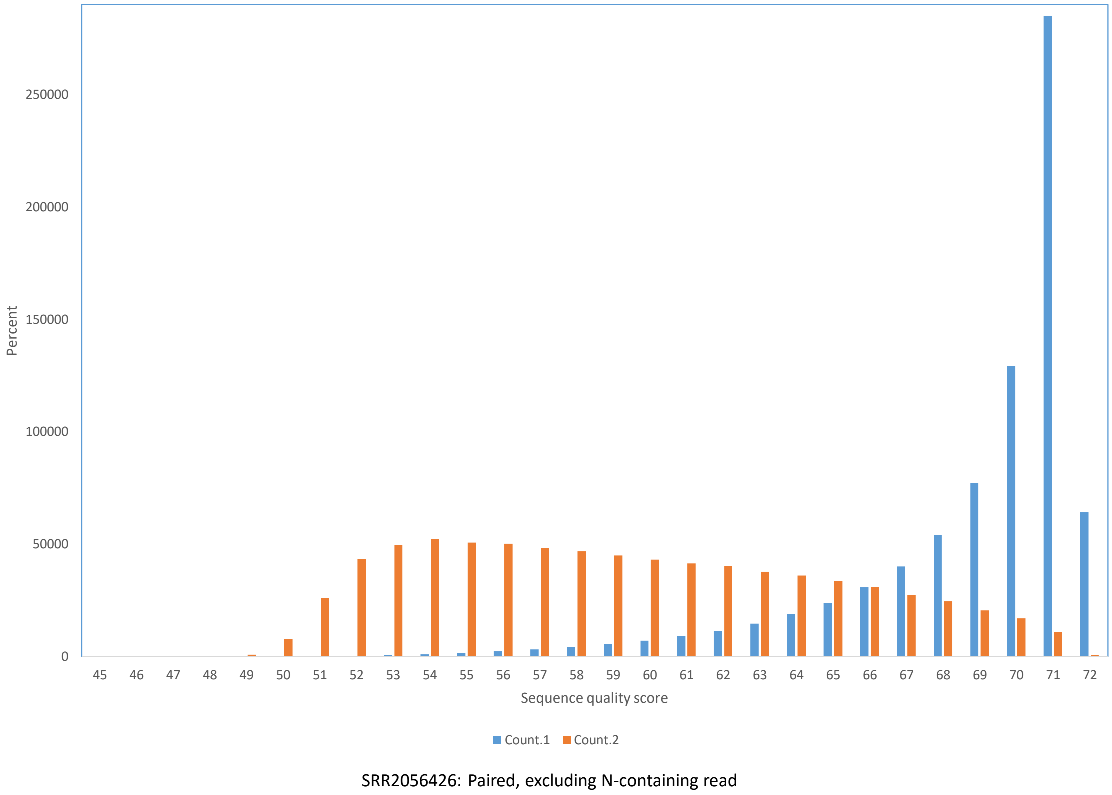


Fig. 7. Frequency distribution of the quality of individual reads for sequences in SRR2056426.sra with paired-end reads. The reverse read (Count2, in orange) is generally poor in quality. The graph includes only sequences without ambiguous codes, otherwise the quality would be even worse.

The site-specific quality distribution (Fig. 8 for SRR1536586.fastq) shows the change of base-calling quality with sites. The values for the first 15 or so sites can be ignored as the sequencing machine needs to have enough data to assign appropriate base-calling qualities. Fig. 8 shows the decreasing trend of base-calling quality with sites. However, because this data set have only short reads (50 nt) and is of high quality even among RNA-Seq data set with read length of 50, the decrease is not alarming. It might help to contrast this pattern with RNA-Seq data in a file with longer reads, such as SRRSRR2056426.sra with paired-end reads and read length of 250 (Fig. 9).

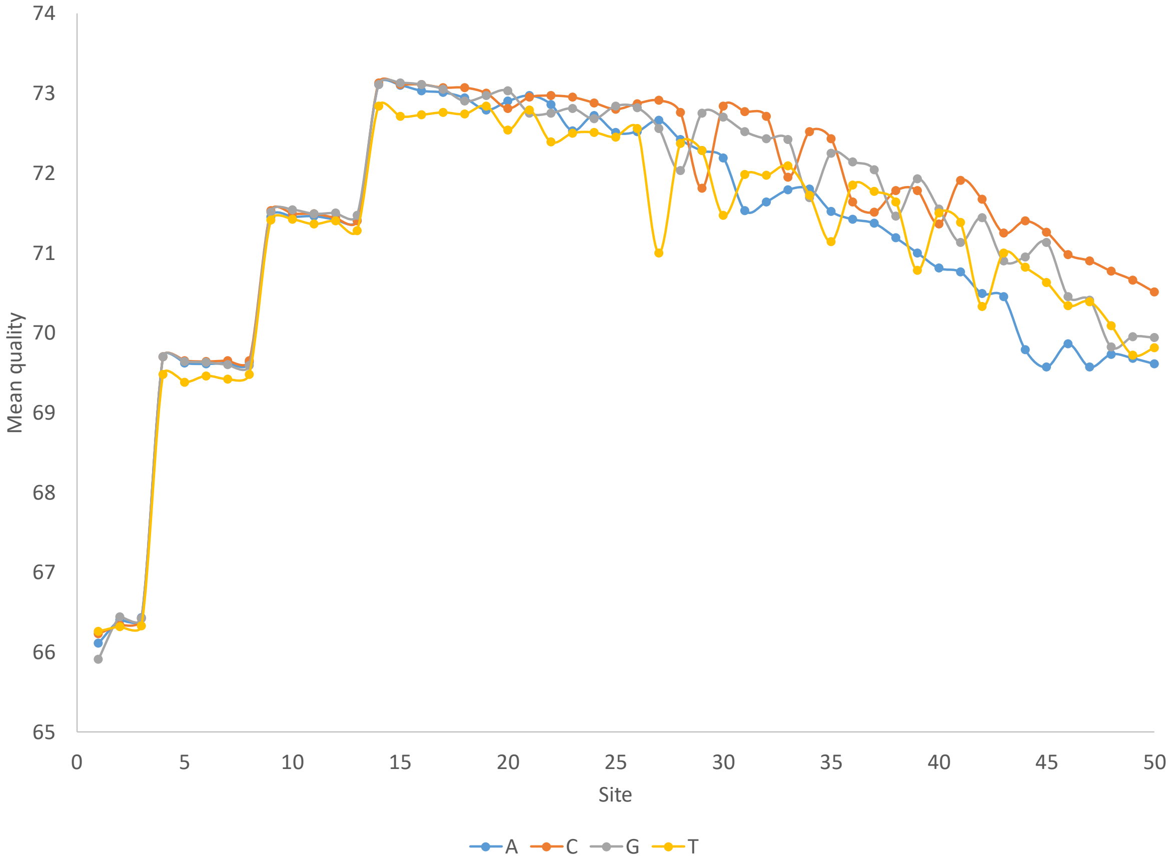


Fig. 8. Site-specific quality for sequences in SRR1536586.sra, including only sequences without ambiguous codes (otherwise the quality would be worse). The quality score of the first ~15 sites may be ignored because the sequencing machine needs to accumulate enough information to generate quality scores properly.

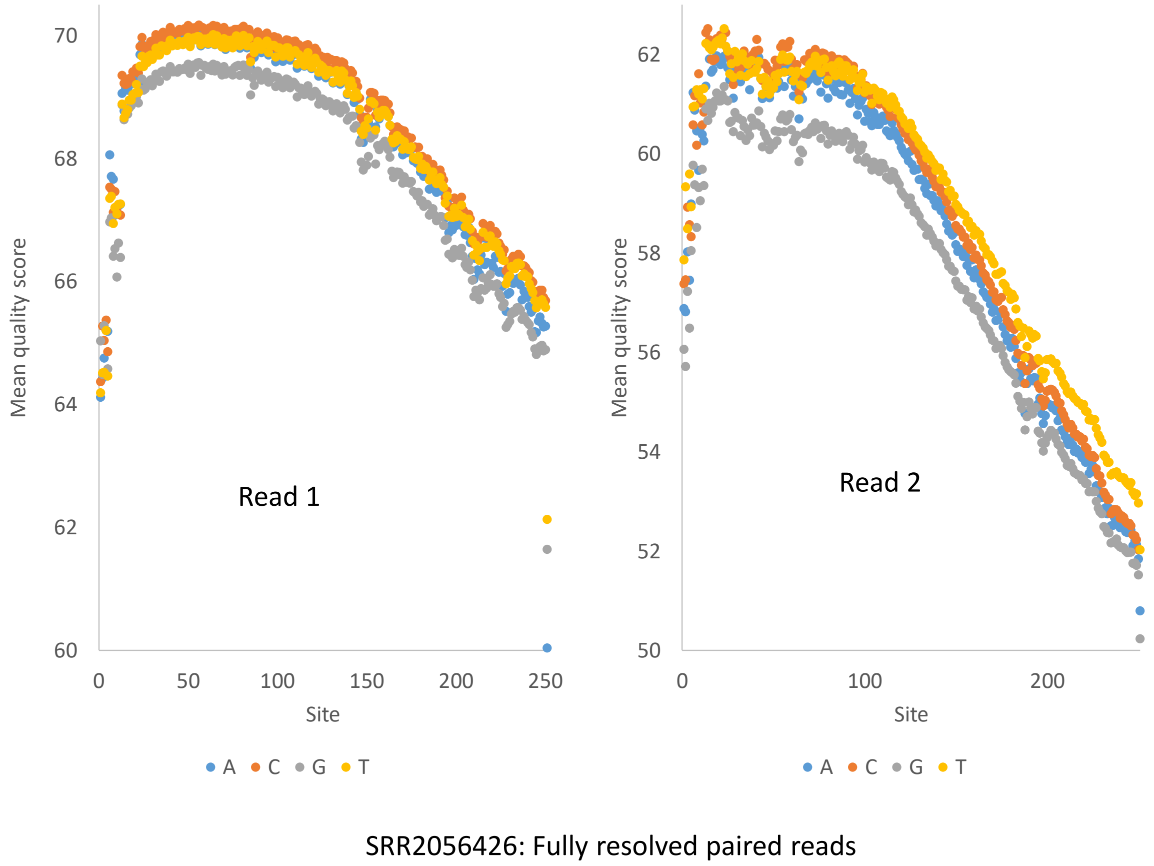


Fig. 9. Quality of individual reads for sequences in SRR2056426.sra with paired-end reads. The reverse read (Read 2) is generally poor in quality. The graph includes only sequences without ambiguous codes, otherwise the quality would be worse. The quality score of the first ~15 sites may be ignored because the sequencing machine needs to accumulate enough information to generate quality scores properly.

QUANTIFYING GENE EXPRESSION (FPKM)

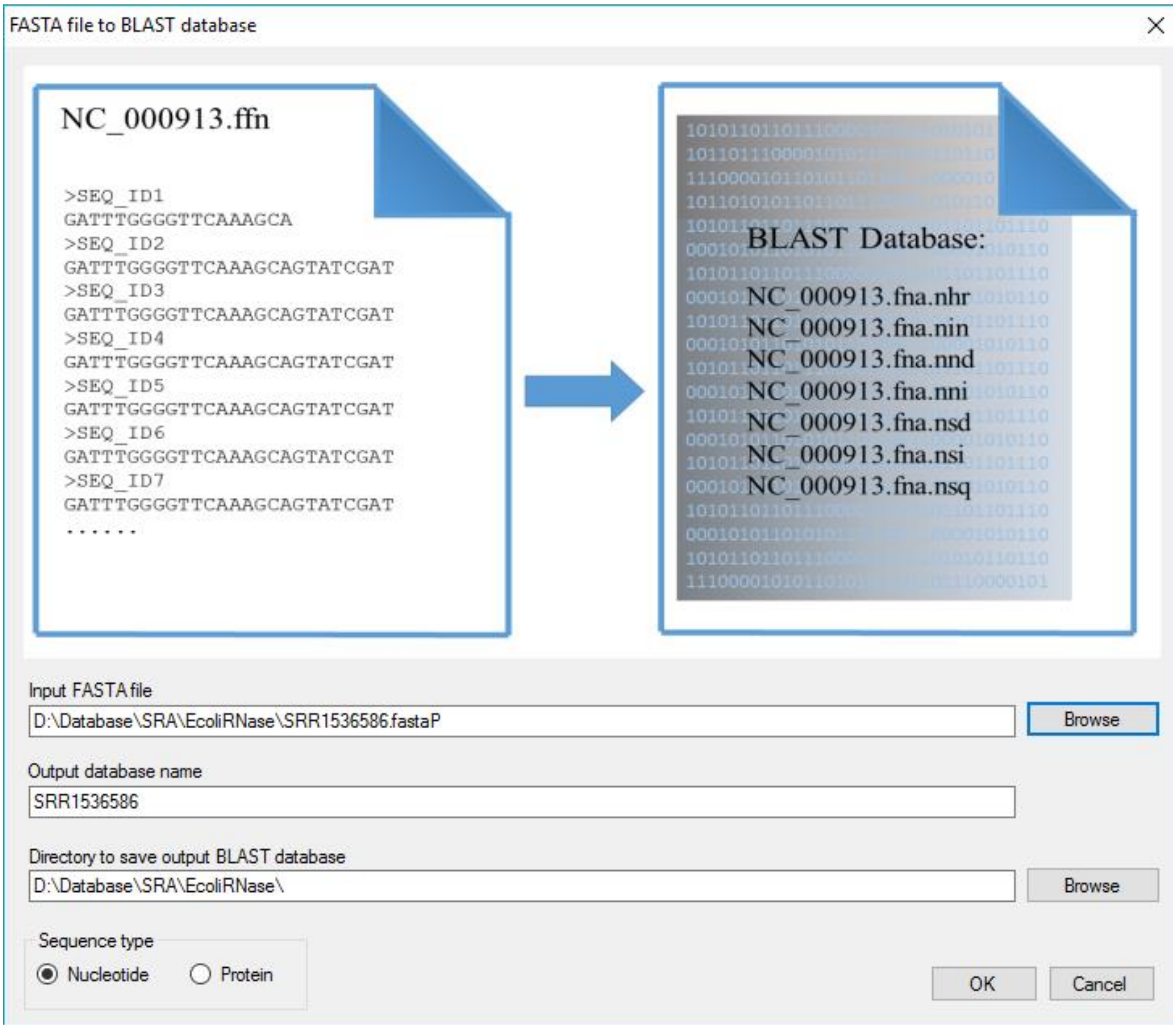
Characterizing gene expression involves a process of assigning reads to genes and normalize the read count for each gene to FPKM (Fragment per kilobases per million matched reads, sometimes ‘fragment’ is replaced by ‘read’ leading to RPKM). The normalization allows comparisons not only between genes of different sequence lengths but also between experiments with different number of total matched reads.

ARSDA takes two types of data for input: 1) the transcriptomic data stored in a BLAST database which will be generated from the FASTA+ that we have created previously (You can also create BLAST databases from FASTA files and use them with ARSDA but that will take much more computation time in read matching), and 2) the coding sequences (CDSs) in FASTA format that we can extract from an annotated genomic sequences. ARSDA will then match the reads from the transcriptomic study to the CDSs and output FPKM values for each gene. If you work on model organisms, then the databases from transcriptomic studies may already be available at http://coevol.rdc.uottawa.ca, so all what you need to do is just to take a few clicks to extract CDSs from an annotated genomic sequence.

In this section, we will first learn how to ARSDA to create BLAST databases and how to use another free program DAMBE (Xia, 2013) to extract CDSs sequences from an annotated genome. We then input these two types of data to ARSDA to generate FPKM values.

Convert RNA-Seq data to a BLAST database

In the first section we detailed how to convert a FASTA or FASTQ file to a FASTA+. Suppose that you have already generated a FASTA+ file named SRR1536586.fastaP. We will first convert the SRR1536586.fastaP to a BLAST database to facilitate read-matching. Click 'Database|Create BLAST DB' and browse to and open the SRR1536586.fastaP file (Fig. 10). Click 'OK' and a BLAST database with SRR1536586 as database name is creased in the specified directory.



```
>SEQ_ID1
GATTTGGGGTTCAAAGCA
>SEQ_ID2
GATTTGGGGTTCAAAGCAGTATCGAT
>SEQ_ID3
GATTTGGGGTTCAAAGCAGTATCGAT
>SEQ_ID4
GATTTGGGGTTCAAAGCAGTATCGAT
>SEQ_ID5
GATTTGGGGTTCAAAGCAGTATCGAT
>SEQ_ID6
GATTTGGGGTTCAAAGCAGTATCGAT
>SEQ_ID7
GATTTGGGGTTCAAAGCAGTATCGAT
......
```

Input FASTA file: D:\Database\SRA\EcoliRNase\SRR1536586.fastaP

Output database name: SRR1536586

Directory to save output BLAST database: D:\Database\SRA\EcoliRNase\

Sequence type: ● Nucleotide ○ Protein

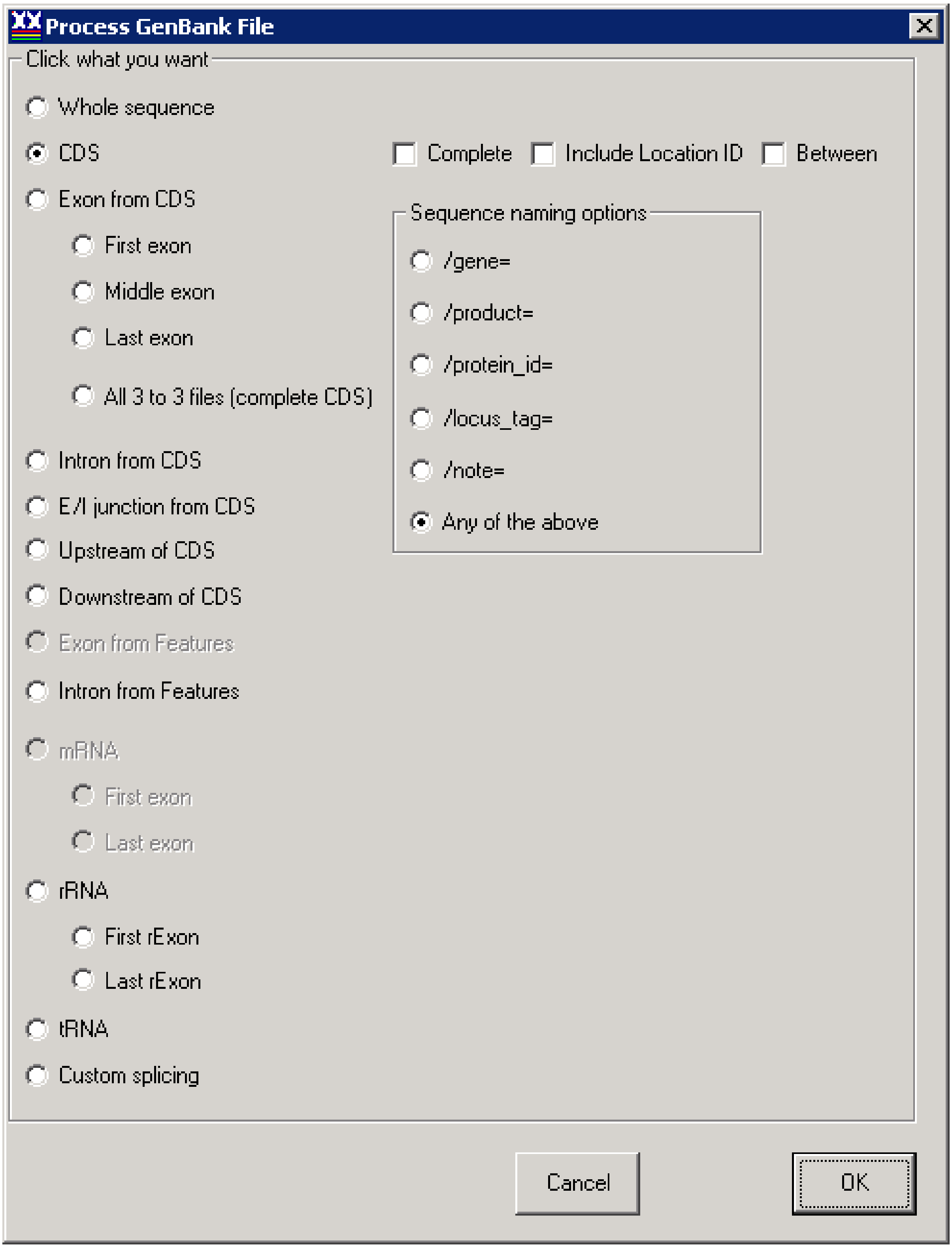
Fig. 10. Input and output for generating a BLAST database from a FASTA file. The latest version has an additional checkbox for processing multiple files.

Extracting gene sequences and save to FASTA format

Suppose we wish to obtain gene expression for *E. coli* K-12 coding sequences (CDSs). We will need to download the relevant genomic file for *E. coli* K-12, extract coding sequences (CDSs) and save it to a file in FASTA format. Obtaining such a file requires only a few mouse clicks by using my DAMBE (Xia, 2013) which is available free at http://dambe.bio.uottawa.ca/DAMBE/dambe.aspx. It takes only a few clicks to install just like ARSDA.

Download *E. coli* GenBank file NC_000913.gbk for E. coli K-12 strain MG1655 which is the closest to the experimental K-12 strain MG1693. Launch DAMBE, click 'File|Open standard sequence file' to open the NC_000913.gbk. DAMBE can extract coding sequences

(CDSs), exons, introns, rRNAs or tRNAs (Fig. 11). Select 'CDS' and optionally include gene location information. It is generally a good idea to include location information (i.e., starting and ending sites of a gene in the chromosome), especially in prokaryotes because closely space genes are often located in the same operon and have similar expression levels. Save the extracted sequences in FASTA format to EcoliCDSGeneLoc.fas (or whatever file name you wish to name it, but remember where it is located).



Process GenBank File

Click what you want

○ Whole sequence
● CDS
○ Exon from CDS
- ○ First exon
- ○ Middle exon
- ○ Last exon
- ○ All 3 to 3 files (complete CDS)

○ Intron from CDS
○ E/I junction from CDS
○ Upstream of CDS
○ Downstream of CDS
○ Exon from Features
○ Intron from Features
○ mRNA
- ○ First exon
- ○ Last exon

○ rRNA
- ○ First rExon
- ○ Last rExon

○ tRNA
○ Custom splicing

□ Complete □ Include Location ID □ Between

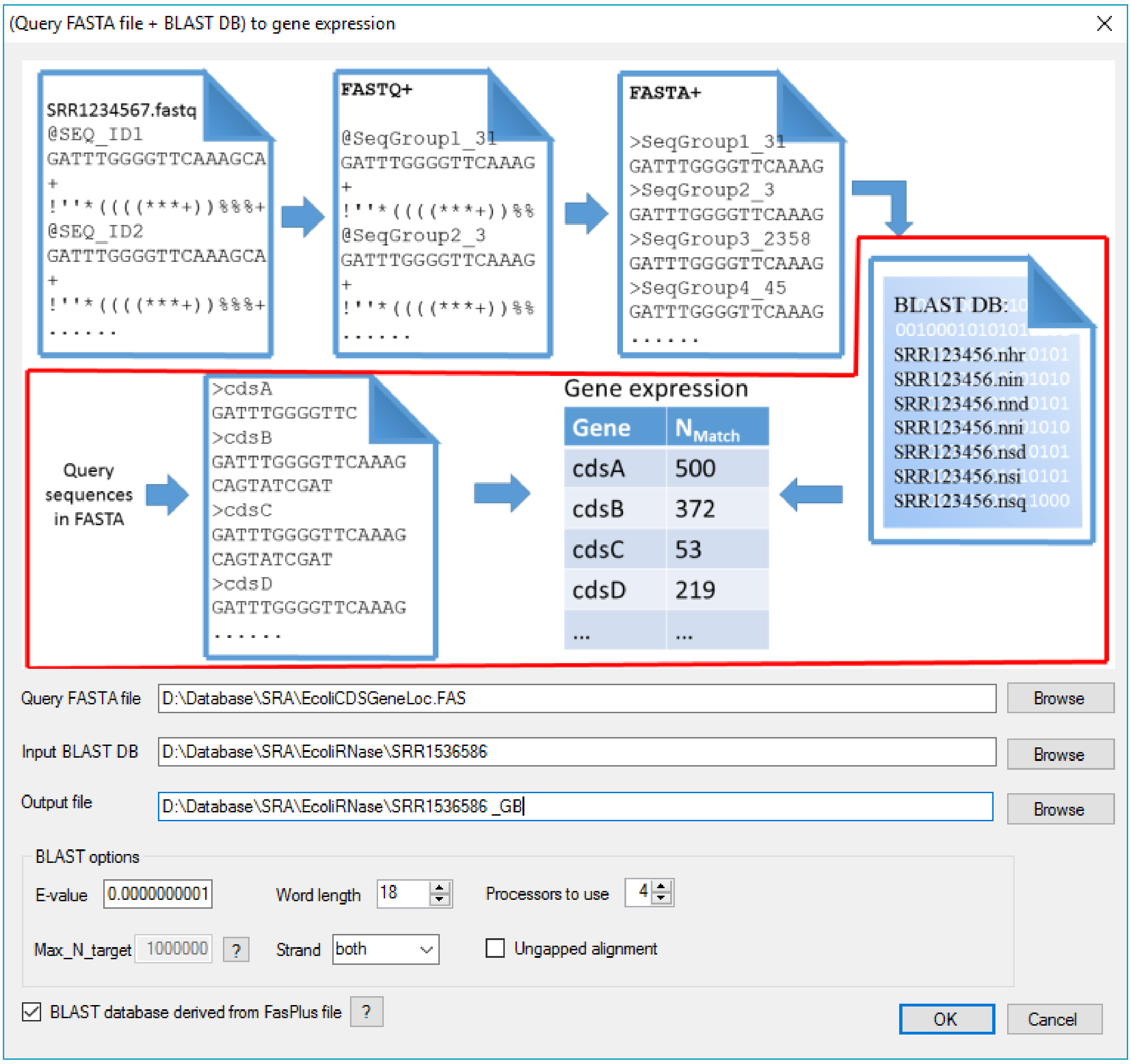
Sequence naming options

○ /gene=
○ /product=
○ /protein_id=
○ /locus_tag=
○ /note=
● Any of the above

Fig. 11. Sequence extraction dialog box in DAMBE for GenBank files.

Quantifying gene expression

Now we have everything needed to characterize gene expression. Close DAMBE and other memory-hungry programs (DAMBE is not memory-hungry) and go back to ARSDA. Click ‘Analysis|Gene expression from BLAST database’ (NOT ‘Gene expression from SRA database’ which is extremely slow). Enter the FASTA file that we have just created from *E. coli* CDSs in the ‘Query FASTA file’ input field, the BLAST database we created in the ‘Input BLAST DB’ field, and whatever file name in the ‘Output file’ field (Fig. 12). Click ‘OK’ and gene expression for the 4321 *E. coli* K12 CDSs will be generated. Part of the output is shown is Table 2.



Gene	N_{Match}
cdsA	500
cdsB	372
cdsC	53
cdsD	219
...	...

Query FASTA file: D:\Database\SRA\EcoliCDSGeneLoc.FAS

Input BLAST DB: D:\Database\SRA\EcoliRNase\SRR1536586

Output file: D:\Database\SRA\EcoliRNase\SRR1536586_GB

BLAST options

E-value: 0.0000000001 Word length: 18 Processors to use: 4

Max_N_target: 1000000 Strand: both □ Ungapped alignment

☑ BLAST database derived from FasPlus file

Fig. 12. Input specification for characterizing gene expression.

Table 2. Partial output of gene expression, with the gene locus_tag (together with start and end sites) as gene ID.

Gene ID	SeqLen	Count	Count/Kb	FPKM
b0001|190_255	66	76	1151.515	389.894
b0002|337_2799	2463	2963	1203.004	407.328
b0003|2801_3733	933	1121	1201.501	406.819
b0004|3734_5020	1287	1782	1384.615	468.82
b0005|5234_5530	297	97	326.599	110.584

b0006|C5683_6459	777	113	145.431	49.242
b0007|C6529_7959	1431	143	99.93	33.836
b0008|8238_9191	954	1561	1636.268	554.028
b0009|9306_9893	588	289	491.497	166.417
b0010|C9928_10494	567	100	176.367	59.716
b0011|C10643_11356	714	13	18.207	6.165
b0013|C11382_11786	405	2	4.938	1.672
b0014|12163_14079	1917	6863	3580.073	1212.186
b0015|14168_15298	1131	1671	1477.454	500.255
…	…	…	…	…

REFERENCES

DENG Q, RAMSKOLD D, REINIUS B, SANDBERG R. 2014 Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science. 343(6167):193-196.

DOBIN A, DAVIS CA, SCHLESINGER F, DRENKOW J, ZALESKI C, JHA S, BATUT P, CHAISSON M, GINGERAS TR. 2013 STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 29(1):15-21.

LANGMEAD B, HANSEN KD, LEEK JT. 2010 Cloud-scale RNA-sequencing differential expression analysis with Myrna. Genome Biology. 11(8):R83.

LANGMEAD B, SALZBERG SL. 2012 Fast gapped-read alignment with Bowtie 2. Nat Methods. 9(4):357-359.

LANGMEAD B, TRAPNELL C, POP M, SALZBERG SL. 2009 Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology. 10(3):R25.

POBRE V, ARRAIANO CM. 2015 Next generation sequencing analysis reveals that the ribonucleases RNase II, RNase R and PNPase affect bacterial motility and biofilm formation in E. coli. Bmc Genomics. 16:72.

ROBERTS A, SCHAEFFER L, PACHTER L. 2013 Updating RNA-Seq analyses after re-annotation. Bioinformatics. 29(13):1631-1637.

ROBERTS A, TRAPNELL C, DONAGHEY J, RINN JL, PACHTER L. 2011 Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biology. 12(3):R22.

TRAPNELL C, PACHTER L, SALZBERG SL. 2009 TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 25(9):1105-1111.

TRAPNELL C, ROBERTS A, GOFF L, PERTEA G, KIM D, KELLEY DR, PIMENTEL H, SALZBERG SL, RINN JL, PACHTER L. 2012 Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 7(3):562-578.

XIA X. 2013 DAMBE5: A comprehensive software package for data analysis in molecular biology and evolution. Molecular Biology and Evolution. 30:1720-1728.

XIA X, XIE Z. 2001 DAMBE: Software package for data analysis in molecular biology and evolution. Journal of Heredity. 92(4):371-373.

Last revised on January 27, 2017, Xuhua Xia