ARSDA: A new approach for storing, transmitting and analyzing high-throughput sequencing data

Xuhua Xia[12]

1. Department of Biology, University of Ottawa, 30 Marie Curie, P.O. Box 450, Station A, Ottawa, Ontario, Canada, K1N 6N5. Tel: (613) 562-5800 ext. 6886, Fax: (613) 562-5486, E-mail: xxia@uottawa.ca
2. Ottawa Institute of Systems Biology, Ottawa, Ontario, K1H 8M5 Canada

Keywords: high-throughput sequencing, novel storage solution, expression of paralogous genes, sequence format, ARSDA

Running title: ARSDA for high-throughput sequencing data

ABSTRACT

Two major stumbling blocks exist in high-throughput sequencing (HTS) data analysis. The first is the sheer file size typically in gigabytes when uncompressed, causing problems in storage, transmission and analysis. However, these files do not need to be so large and can be reduced without loss of information. Each HTS file, either in compressed .SRA or plain text .fastq format, contains numerous identical reads stored as separate entries. For example, among 44603541 forward reads in the SRR4011234.sra file (from a *Bacillus subtilis* transcriptomic study) deposited at NCBI's SRA database, one read has 497027 identical copies. Instead of storing them as separate entries, one can and should store them as a single entry with the SeqID_NumCopy format (which I dub as FASTA+ format). The second is the proper allocation reads that map equally well to paralogous genes. I illustrate in detail a new method for such allocation. I have developed ARSDA software that implement these new approaches. A number of HTS files for model species are in the process of being processed and deposited at http://coevol.rdc.uottawa.ca to demonstrate that this approach not only saves a huge amount of storage space and transmission bandwidth, but also dramatically reduces time in downstream data analysis. Instead of matching the 497027 identical reads separately against the *Bacillus subtilis* genome, one only needs to match it once. ARSDA includes functions to take advantage of HTS data in the new sequence format for downstream data analysis such as gene expression characterization. ARSDA can be run on Windows, Linux and Macintosh computers and is freely available at http://dambe.bio.uottawa.ca/ARSDA/ARSDA.aspx.

INTRODUCTION

High-throughput sequencing (HTS) is now used not only in characterizing differential gene expression, but also in many other fields where it replaces the traditional microarray approach. Ribosomal profiling, traditionally done through microarray (ARAVA *et al.* 2003; MACKAY *et al.* 2004), is now almost exclusively done with deep sequencing of ribosome-protected segments of messages (INGOLIA *et al.* 2009a; INGOLIA *et al.* 2009b; INGOLIA *et al.* 2011), although the results from the two approaches for ribosomal profiling are largely concordant (XIA *et al.* 2011). Similarly, EST-based (ROGERS *et al.* 2012) and microarray-based (PLEISS *et al.* 2007) methods for detecting alternative splicing events and characterizing splicing efficiency is now replaced by HTS (KAWASHIMA *et al.* 2014), especially by lariat sequencing (AWAN *et al.* 2013; STEPANKIW *et al.* 2015). The availability of HTS data has dramatically accelerated the test of biological hypotheses. For example, a recent study on alternative splicing (VLASSCHAERT *et al.* 2016) showed that skipping of exon 7 (E_7) in human and mouse *USP4* is associated with weak signals of splice sites flanking E_7. The researchers predicted that, in some species where the splice site signals are strong, E_7 skipping would disappear. This prediction is readily tested and confirmed with existing HTS data, i.e., E_6-E_8 mRNA was found in species with weak splice signals flanking E_7, and E_6-E_7-E_8 mRNA in species with strong splice signals flanking E_7 (VLASSCHAERT *et al.* 2016).

In spite of the potential of HTS data in solving practical biological problems, severe under-usage of HTS data has been reported (TEAM 2011). One major stumbling block in using the HTS data is the large file size. Among the 6472 transcriptomic studies on human available at NCBI/DDBJ/EBI by Apr. 14, 2016, 196 studies each contribute more than 1 Terabytes (TB) of nucleotide bases, with the top one contributing 86.4 TB. Few laboratories would be keen on downloading and analyzing this 86.4 TB of nucleotides, not to mention comparing this study to HTS data from other human transcriptomic studies.

The explosive growth of HTS data in recent years has caused serious problems in data storage, transmission and analysis(LEINONEN *et al.* 2011; KODAMA *et al.* 2012). Because of the high cost of maintaining such data, coupled with the fact that few researchers had been using such data, NCBI had planned the closure of the sequence read archive a few years ago (TEAM 2011), but continued the support only after DDBJ and EBI decided to continue their effort of archiving the data. The incident highlights the prohibitive task of storing, transmitting and analyzing HTS data, and motivated the joint effort of both industry and academics to search for data compression solutions (JANIN *et al.* 2014; ZHU *et al.* 2015b; NUMANAGIC *et al.* 2016).

A SOLUTION WITH A NEW SEQUENCE FORMAT

HTS data files do not need to be so huge. 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 sequence reads of 50 nt each, but 195310 sequences are all identical (TGTTATCACGGGAGACACACGGCGGGTGCTAACGTCCGTCGTGAAGAGGG), all mapping exactly to sites 929-978 in *E. coli* 23S rRNA genes (The study did use the MICROBExpress Bacterial mRNA Enrichment Kit to remove the 16S and 23S rRNA, otherwise there would be many more). There are much more extreme cases. For example, one of the 12 HTS files from a transcriptomic study of *Escherichia coli* (SRR922264.sra), contains a read with 1,606,515 identical copies among its 9,690,570 forward reads. There is no sequence information lost if all these 1,606,515 identical reads are stored by a single sequence with a sequence ID such as UniqueSeqX_1606515 (i.e., SeqID_CopyNumber format which I dub as FASTA+ format with file type .fasP). Such storage scheme not only results in dramatic saving in data storage and transmission, but also leads to dramatic reduction in computation time in downstream data analysis. At present, all software packages for HTS data analysis will take the 1,606,515 identical reads separately and search them individually against the *E. coli* genome (or target gene sequences such as coding sequences). The SeqID_CopyNumber storage scheme reduces the 1,606,515 searches to a single one.

A huge chunk of SRA data stored in NCBI/DDBJ/EBI consists of ribosome profiling data (INGOLIA *et al.* 2009a; INGOLIA *et al.* 2009b; INGOLIA *et al.* 2011), which is obtained by sequencing the mRNA segment (~30 bases) protected by the ribosome after digesting all the unprotected mRNA segments. Mapping these ribosome-protected segments to the genome allows one to know specifically where the ribosomes are located along individual mRNAs. In general, such data are essential to understand translation initiation, elongation and termination efficiencies. For example, a short poly(A) segment with about eight or nine consecutive A immediately upstream of the start codon in yeast (*Saccharomyces cerevisiae*) genes is significantly associated with ribosome density and occupancy (XIA *et al.* 2011), confirming the hypothesis that short poly(A) upstream of the start codon facilitates the recruitment of translation initiation factors but long poly(A) would bind to poly(A)-binding protein and interfere with cap-dependent translation. Sequence redundancy is high in such ribosomal profiling data and the FASTA+ format can lead to dramatic saving in the disk space of data storage and time in data transmission.

ARSDA

I developed software ARSDA (for Analyzing RNA-Seq Data, Fig. 1a) to alleviate the problem associated with storage, transmission and analysis of HTS data. ARSDA can take input .SRA files or .fastq files of many gigabytes, build an efficient dictionary of unique sequence reads in a FASTA/FASTQ file, keep track of their copy numbers, and output them to a FASTA+ file in the SeqID_CopyNumber format (Fig. 1b). Both fixed-length and variable-length sequences can be used as input. In addition, I have implemented functions in ARSDA to take advantage of the new sequence format to perform gene expression, with the main objective to demonstrate how much faster downstream data analysis can be done with data in FASTA+/FASTQ+ format. However, ARSDA does include a unique feature in assigning shared reads among paralogous genes that I will describe below. ARSDA also includes sequence visualization functions for global base-calling quality, per-read quality and site-specific read quality (Fig. 1c-d), but these functions are also available elsewhere, e.g., FastQC (ANDREWS 2017) and NGSQC (DAI *et al.* 2010) and consequently will not be described further (but see the attached QuickStart.PDF).

Converting HTS data to FASTA+/FASTQ+

The output from processing the SRR1536586.sra file (with part of the read matching output in Table 1) highlights two points. First, many sequences in the file are identical. Second, although the transcriptomic data characterized in SRR1536586 have undergone rRNA depletion by using

Ambion's MICROBExpress Bacterial mRNA Enrichment Kit (POBRE and ARRAIANO 2015), there are still numerous reads in the transcriptomic data that are from rRNA genes. This suggests that storing mRNA reads separately from rRNA reads can dramatically reduce file size because most researchers are not interested in the abundance of rRNAs.

While the conversion of FASTA/FASTQ files to FASTA+ files is time-consuming, it needs to be done only once for data storage, preferably at NCBI/EBI/DDBJ, and the resulting saving in storage space, internet traffic and computation time in downstream data analysis is tremendous. The file size is 1.49 GB for the original FASTQ file derived from SRR1536586.sra, but is only 66 MB for the new FASTA+ file, both being plain text files.

I have further created BLAST databases from the processed HTS files in FASTA+ format for model species such as *E. coli, B. subtilis, S. cerevisiae* and *Caenorhabditis elegans* and deposited them at coevol.rdc.uottawa.ca. The sequence ID in these BLAST databases are in the form of SeqID_CopyNumber. These files reduce the computation time for quantifying gene expression from many hours to only a few minutes (less than two minutes for my Windows 10 PC with ani7-4770 CPU at 3.4GHz and 16 GB of RAM). This eliminates one of the key bottleneck in HTS data analysis (LIU *et al.* 2016) and would make it feasible for any laboratory to gain the power of HTS data analysis. I attach the gene expression characterized by ARSDA for the 4321 *E. coli* K12 coding sequences as supplemental file SRR1536586_GB.txt. A part of it is reproduced in Table 2.

One of the frequently used sequence-compression scheme is to use a reference genome so that each read can be represented by a starting and an ending number on the genome (BENOIT *et al.* 2015; KINGSFORD and PATRO 2015; ZHU *et al.* 2015a). This approach has two problems. First, many reads do not map exactly to the genomic sequence because of either somatic mutations or sequencing errors, so representing a read by the starting and ending numbers leads to loss of information. Second, RNA-editing and processing can be so extensive that it becomes impossible to map a transcriptomic read to the genome (ABRAHAM *et al.* 1988; LAMOND 1988; ALATORTSEV *et al.* 2008; LI *et al.* 2009; SIMPSON *et al.* 2016). Furthermore, there are still many scientifically interesting species that do not have a good genomic sequence available.

Assigning sequence reads to paralogous genes

One of the most fundamental objectives of RNA-Seq analysis is to generate an index of gene expression (FPKM: matched fragment/reads per kilobases of transcript per million mapped reads) that can be directly compared among different genes and among different experiments with different total number of matched reads (MORTAZAVI *et al.* 2008). The main difficulty in quantifying gene expression arises with sequence reads matching multiple paralogous genes (TRAPNELL *et al.* 2013; ROGOZIN *et al.* 2014). This problem, which has plagued microarray data analysis, is now plaguing RNA-Seq analysis. Most publications of commonly used RNA-Seq analysis methods (LANGMEAD *et al.* 2009; TRAPNELL *et al.* 2009; LANGMEAD *et al.* 2010; ROBERTS *et al.* 2011; LANGMEAD and SALZBERG 2012; TRAPNELL *et al.* 2012; DOBIN *et al.* 2013; ROBERTS *et al.* 2013; DENG *et al.* 2014) often avoided mentioning read allocation to paralogous genes. The software tools associated with these publications share two simple options for handling matches to paralogous genes. The first is to use only uniquely matched reads, i.e., reads that match to multiple genes are simply ignored. The second is to assign such reads equally among matched genes. These options are obviously unsatisfactory. Here I describe a new approach which should substantially improve the accuracy of HTS data analysis such as gene expression characterization.

Allocating sequence reads to paralogous genes in a two-member gene family

We need a few definitions to explain the allocation. Let L_1 and L_2 be the sequence length of the two paralogous genes. Let $N_{U.1}$ and $N_{U.2}$ be the number of reads that can be uniquely assigned to paralogous gene 1 or 2, respectively (i.e., the reads that matches one gene better than the other). Now for those reads that match the two genes equally well, the proportion of the reads contributed by paralogous gene 1 may be simply estimated as

$$P_1 = \frac{N_{U.1}}{N_{U.1} + N_{U.2}} \tag{1}$$

Now for any read that matches the two paralogous genes equally well, we will assign P_1 to paralogous gene 1, and (1-P_1) to paralogous gene 2. In the extreme case when paralogous genes are all identical, then $N_{U.1} = N_{U.2} = 0$, and we will assign 1/2 of these equally matched read to genes 1 and 2. We should modify Eq. (1) to make it more generally applicable as follows

$$P_1 = \frac{0.01 + N_{U.1}}{0.02 + N_{U.1} + N_{U.2}} \tag{2}$$

where 0.01 in the numerator and 0.02 in the denominator are pseudocounts. The treatment in Eq. (2) implies that when $N_{U.1} = N_{U.2} = 0$ (e.g., when two paralogous genes are perfectly identical), then a read matching equally well to these paralogous genes will be equally divided among the two paralogues.

One problem with this treatment is its assumption of $L_1 = L_2$. If paralogous gene 1 is much longer than the other, then $N_{U.1}$ is expected to be larger than $N_{U.2}$, everything else being equal. One may standardize $N_{U.1}$ and $N_{U.2}$ to number of unique matches per 1000 nt, designated by $SN_{U.i} = 1000N_{U.i}/L_i$ (where i = 1 or 2) and replace $N_{U.i}$ in Eq. (2) by $SN_{U.i}$ as follows (MORTAZAVI *et al.* 2008):

$$P_1 = \frac{0.01 + SN_{U.1}}{0.02 + SN_{U.1} + SN_{U.2}} = \frac{0.01 + \dfrac{1000N_{U.1}}{L_1}}{0.02 + 1000\left(\dfrac{N_{U.1}}{L_1} + \dfrac{N_{U.2}}{L_2}\right)} \tag{3}$$

Allocating sequence reads in gene family with more than two members

One might, mistakenly, think that it is quite simple to extend Eq. (3) for a gene family of two members to a gene family with F members by writing

$$P_i = \frac{0.01 + \dfrac{1000N_{U.i}}{L_i}}{0.01F + 1000\sum_{i=1}^{F} \dfrac{N_{U.i}}{L_i}} \tag{4}$$

This does not work. For example, if we have three paralogous genes designated A, B, and C, respectively. Suppose that the gene duplication that gave rise to B and C occurred very recently so that B and C are identical, but A and the ancestor of B and C have diverged for a long time. In this case, $N_{U.B} = N_{U.C} = 0$ because a read matching B will always matches C equally well, but $N_{U.A}$ may be greater than 0. This will result in unfair allocation of many transcripts from B and C to A according to Eq. (4). I outline the approach below for dealing with gene families with more than two members.

With three or more paralogous genes, one may benefit from a phylogenetic tree for proper allocation of sequence reads. I illustrate the simplest case with a gene family with three paralogous genes A, B, and C idealized into three segments in Fig. 3.The three genes shared one identical middle segment with 23 matched reads (that necessarily match equally well to all three paralogues). Genes B and C share an identical first segment to which 20 reads matched. Gene A has its first segment different from that of B and C and got four matched reads. The three genes also have a diverged third segment where A matched 3 reads, B matched 6 and C matched 12. Our task is then to allocate the 23 reads shared by all three and 20 reads shared by B and C to the three paralogues.

One could apply maximum likelihood or least-squares method for the estimation, but ARSDA uses a simple counting approach by applying the following

$$\begin{aligned} P_A &= \frac{3+4}{3+4+20+6+12} = 0.15556 \\ P_B &= (1-P_A)\frac{6}{6+12} = 0.28148 \\ P_C &= (1-P_A)\frac{12}{6+12} = 0.56296 \end{aligned} \tag{5}$$

Thus, we allocate the 23 reads (that matched three genes equally) to paralogous genes A, B and C according to P_A, P_B and P_C, respectively. For the 20 reads that matched B and C equally well, we allocate 20*6/(6+12) to B and 20*12/(6+12) to C. This gives the estimated number of matches to each gene as

$$\begin{aligned} N_A &= 3+4+23P_A = 10.57778 \\ N_B &= 6+23P_B+20\left(\frac{6}{6+12}\right) = 19.14074 \\ N_C &= 12+23P_C+20\left(\frac{12}{6+12}\right) = 38.28148 \end{aligned} \tag{6}$$

These numbers are then normalized to give FPKM (MORTAZAVI *et al.* 2008). The current version of ARSDA assume that gene families with more than two members to have roughly the same sequence lengths. This is generally fine with prokaryotes but may become problematic with eukaryotes.

In practice, one can obtain the same results without actually undertaking the extremely slow process of building trees for paralogous genes. One first goes through reads shared by two paralogous genes (e.g., the 20 reads shared by genes B and C in Fig. 2) and allocate the reads according to P_B = 6/(6+12) = 1/3 and P_C = 12/(6+12) = 2/3. Now genes B and C will have 12.66667 (=6+20*P_B) and 25.33333 (=12+20*P_C) assigned reads, i.e., $N_{U.B}$ = 12.66667 and $N_{U.C}$ = 25.33333. Once we have done with reads shared by two paralogous genes, we go through reads shared by three paralogous genes, e.g., the 23 reads shared by genes A, B, and C in Fig. 2. With $N_{U.A}$ = 7, $N_{U.B}$ = 12.66667, $N_{U.C}$ = 25.333333, and N = $N_{U.A}$ + $N_{U.B}$ + $N_{U.C}$ = 45, so we have

$$P_A = \frac{N_{U.A}}{N} = 0.15556; P_B = \frac{N_{U.B}}{N} = 0.28148; P_C = \frac{N_{U.C}}{N} = 0.56296 \tag{7}$$

$$\begin{aligned} N_A &= 7 + 23P_A = 10.57778 \\ N_B &= 12.66667 + 23P_B = 19.14074 \\ N_C &= 25.33333 + 23P_C = 38.28148 \end{aligned} \tag{8}$$

which are the same as shown in Eq. (6). This progressive process continues until we have allocated reads shared by the largest number of paralogous genes. The gene expression output in the supplemental SRR1536586_GB.txt is obtained in this way.

SOFTWARE AND DATA AVAILABILITY

ARSDA is freely available at http://dambe.bio.uottawa.ca/ARSDA/ARSDA.aspx, together with a QuickStart.PDF file showing HTS file conversion from FASTA/FASTQ file to FASTA+ format, three types of HTS data quality visualization tools, and downstream characterization of gene expression. It is a Windows program but can run on any computer with .NET framework installed (e.g. Macintosh and Linux with MONO activated). The BLAST databases derived from HTS reads for several model species, in which sequence IDs are in the format of SeqID_CopyNumber, are deposited at coevol.rdc.uottawa.ca. One can use these BLAST databases with ARSDA to characterize gene expression or other analysis. Ultimately, it is NCBI/EBI/DDBJ that should store all HTS data in such BLAST databases.

ACKNOWLEDGEMENT

This study is funded by the Discovery Grant from Natural Science and Engineering Research Council (NSERC, RGPIN/261252) of Canada. I thank J. Silke, J. Wang, Y. Wei, and C. Vlasschaert.

ARSDA was demonstrated in two workshops, one organized by Dr. J. Lu of Peking University and another by Dr. E. Pranckeviciene of University of Vilnius. I thank participants for their feedback.

LITERATURE CITED

Abraham, J. M., J. E. Feagin and K. Stuart, 1988 Characterization of cytochrome c oxidase III transcripts that are edited only in the 3' region. Cell 55**:** 267-272.

Alatortsev, V. S., J. Cruz-Reyes, A. G. Zhelonkina and B. Sollner-Webb, 2008 Trypanosoma brucei RNA editing: coupled cycles of U deletion reveal processive activity of the editing complex. Molecular & Cellular Biology 28**:** 2437-2445.

Andrews, S., 2017 FastQC, pp., Babraham Bioinformatics.

Arava, Y., Y. Wang, J. D. Storey, C. L. Liu, P. O. Brown *et al.*, 2003 Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A 100**:** 3889-3894.

Awan, A. R., A. Manfredo and J. A. Pleiss, 2013 Lariat sequencing in a unicellular yeast identifies regulated alternative splicing of exons that are evolutionarily conserved with humans. Proceedings of the National Academy of Sciences of the United States of America 110**:** 12762-12767.

Benoit, G., C. Lemaitre, D. Lavenier, E. Drezen, T. Dayris *et al.*, 2015 Reference-free compression of high throughput sequencing data with a probabilistic de Bruijn graph. BMC Bioinformatics 16**:** 288.

Dai, M., R. C. Thompson, C. Maher, R. Contreras-Galindo, M. H. Kaplan *et al.*, 2010 NGSQC: cross-platform quality analysis pipeline for deep sequencing data. Bmc Genomics 11 Suppl 4**:** S7.

Deng, Q., D. Ramskold, B. Reinius and R. Sandberg, 2014 Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science 343**:** 193-196.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski *et al.*, 2013 STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29**:** 15-21.

Ingolia, N. T., S. Ghaemmaghami, J. R. Newman and J. S. Weissman, 2009a Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324**:** 218-223.

Ingolia, N. T., S. Ghaemmaghami, J. R. S. Newman and J. S. Weissman, 2009b Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. Science 324**:** 218-223.

Ingolia, N. T., L. F. Lareau and J. S. Weissman, 2011 Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell 147**:** 789-802.

Janin, L., O. Schulz-Trieglaff and A. J. Cox, 2014 BEETL-fastq: a searchable compressed archive for DNA reads. Bioinformatics 30**:** 2796-2801.

Kawashima, T., S. Douglass, J. Gabunilas, M. Pellegrini and G. F. Chanfreau, 2014 Widespread use of non-productive alternative splice sites in Saccharomyces cerevisiae. PLoS Genet 10**:** e1004249.

Kingsford, C., and R. Patro, 2015 Reference-based compression of short-read sequences using path encoding. Bioinformatics 31**:** 1920-1928.

Kodama, Y., M. Shumway and R. Leinonen, 2012 The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res 40**:** D54-56.

Lamond, A. I., 1988 RNA editing and the mysterious undercover genes of trypanosomatid mitochondria. Trends Biochem Sci 13**:** 283-284.

Langmead, B., K. D. Hansen and J. T. Leek, 2010 Cloud-scale RNA-sequencing differential expression analysis with Myrna. Genome Biology 11**:** R83.

Langmead, B., and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. Nat Methods 9**:** 357-359.

Langmead, B., C. Trapnell, M. Pop and S. L. Salzberg, 2009 Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10**:** R25.

Leinonen, R., H. Sugawara and M. Shumway, 2011 The sequence read archive. Nucleic Acids Res 39**:** D19-21.

Li, F., P. Ge, W. H. Hui, I. Atanasov, K. Rogers *et al.*, 2009 Structure of the core editing complex (L-complex) involved in uridine insertion/deletion RNA editing in trypanosomatid mitochondria. Proceedings of the National Academy of Sciences of the United States of America 106**:** 12306-12310.

Liu, B., H. Guo, M. Brudno and Y. Wang, 2016 deBGA: read alignment with de Bruijn graph-based seed and extension. Bioinformatics 32**:** 3224-3232.

MacKay, V. L., X. Li, M. R. Flory, E. Turcott, G. L. Law *et al.*, 2004 Gene expression analyzed by high-resolution state array analysis and quantitative proteomics: response of yeast to mating pheromone. Mol Cell Proteomics 3**:** 478-489.

Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer and B. Wold, 2008 Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5**:** 621-628.

Numanagic, I., J. K. Bonfield, F. Hach, J. Voges, J. Ostermann *et al.*, 2016 Comparison of high-throughput sequencing data compression tools. Nat Methods 13**:** 1005-1008.

Pleiss, J. A., G. B. Whitworth, M. Bergkessel and C. Guthrie, 2007 Rapid, transcript-specific changes in splicing in response to environmental stress. Mol Cell 27**:** 928-937.

Pobre, V., and C. M. Arraiano, 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., L. Schaeffer and L. Pachter, 2013 Updating RNA-Seq analyses after re-annotation. Bioinformatics 29**:** 1631-1637.

Roberts, A., C. Trapnell, J. Donaghey, J. L. Rinn and L. Pachter, 2011 Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biology 12**:** R22.

Rogers, M. F., J. Thomas, A. S. Reddy and A. Ben-Hur, 2012 SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data. Genome Biology 13**:** R4.

Rogozin, I. B., D. Managadze, S. A. Shabalina and E. V. Koonin, 2014 Gene family level comparative analysis of gene expression in mammals validates the ortholog conjecture. Genome Biol Evol 6**:** 754-762.

Simpson, R. M., A. E. Bruno, J. E. Bard, M. J. Buck and L. K. Read, 2016 High-throughput sequencing of partially edited trypanosome mRNAs reveals barriers to editing progression and evidence for alternative editing. RNA 22**:** 677-695.

Stepankiw, N., M. Raghavan, E. A. Fogarty, A. Grimson and J. A. Pleiss, 2015 Widespread alternative and aberrant splicing revealed by lariat sequencing. Nucleic Acids Res 43**:** 8488-8501.

Team, G. E., 2011 Closure of the NCBI SRA and implications for the long-term future of genomics data storage. Genome Biology 12**:** 402.

Trapnell, C., D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn *et al.*, 2013 Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol 31**:** 46-53.

Trapnell, C., L. Pachter and S. L. Salzberg, 2009 TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25**:** 1105-1111.

Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim *et al.*, 2012 Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7**:** 562-578.

Vlasschaert, C., X. Xia and D. A. Gray, 2016 Selection preserves Ubiquitin Specific Protease 4 alternative exon skipping in therian mammals. Scientific Reports 6:20039.

Xia, X., V. MacKay, X. Yao, J. Wu, F. Miura *et al.*, 2011 Translation Initiation: A Regulatory Role for Poly(A) Tracts in Front of the AUG Codon in Saccharomyces cerevisiae. Genetics 189**:** 469-478.

Zhu, Z., L. Li, Y. Zhang, Y. Yang and X. Yang, 2015a CompMap: a reference-based compression program to speed up read mapping to related reference sequences. Bioinformatics 31**:** 426-428.

Zhu, Z., Y. Zhang, Z. Ji, S. He and X. Yang, 2015b High-throughput DNA sequence data compression. Brief Bioinform 16**:** 1-15.

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 NCBI.

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

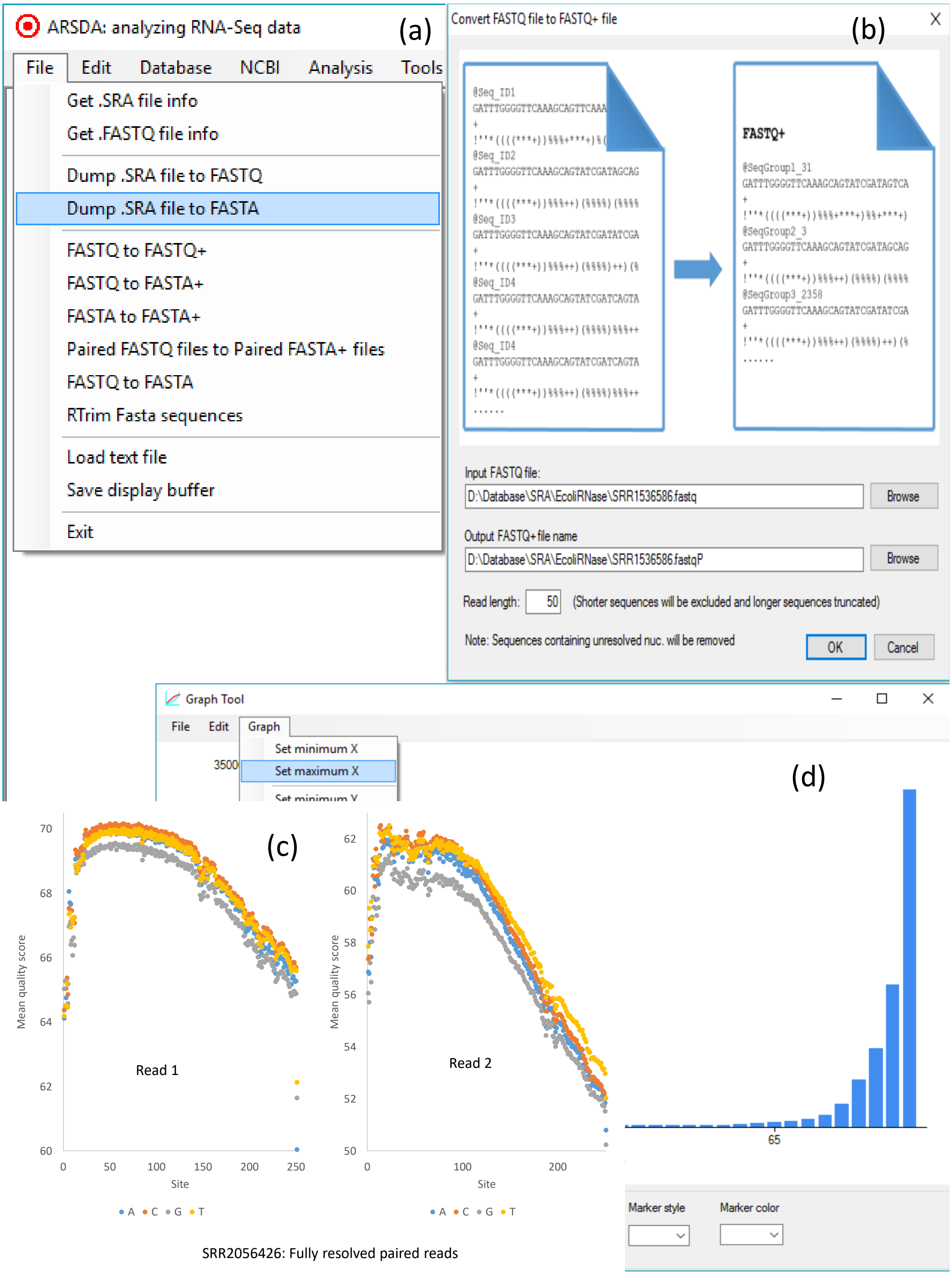
Table 2. Partial output of gene expression, with the gene locus_tag (together with start and end sites) as Gene ID.

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
…	…	…	…	…

FIGURE CAPTIONS

Fig. 1. User interface in ARSDA. (a) The menu system, with database creation under the ‘Database’ menu, gene expression characterization under the ‘Analysis’ menu, etc. (b) Converting a FASTQ/FASTA file to a FASTQ+/FASTA+ file. (c) Site-specific read quality visualization. (d) Global read quality visualization.

Fig. 2. Allocation of shared reads in a gene family with three paralogous genes A, B and C with three idealized segments with a conserved identical middle segment, strongly homologous first segment that is identical in B and C, and a diverged third segment. Reads and the gene segment they match to are of the same color.



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

```
@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

SRR2056426: Fully resolved paired reads

Fig. 1

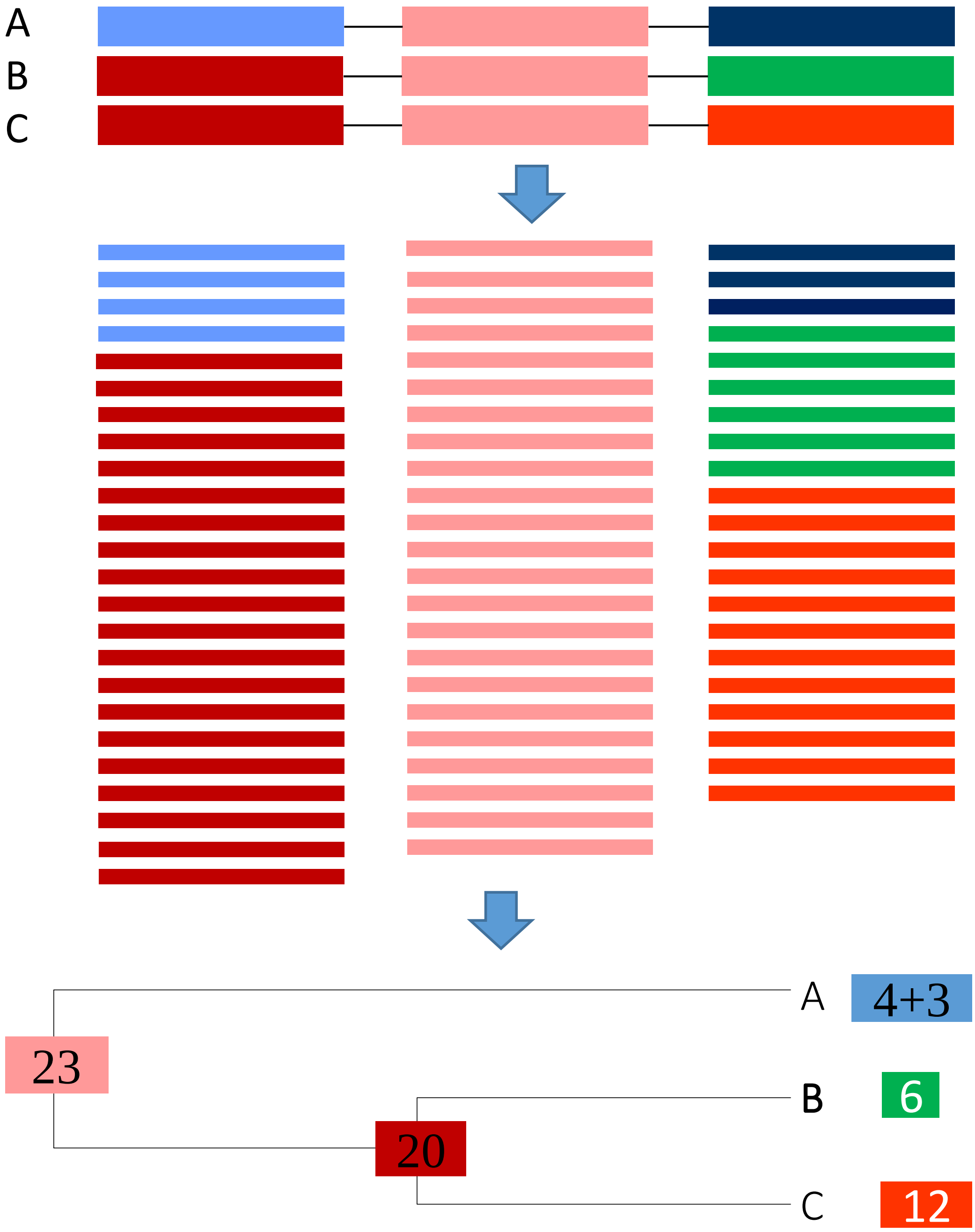


Fig. 2