SCOPE++: Sequence Classification Of homoPolymer Emissions

**James T. Morton1, Patricia Abrudan2, Nathanial Figueroa1, and Chun Liang1,2** §**, John E. Karro1,3,4** §

1Department of 1Computer Science and Software Engineering, 2Biology, 3Microbiology, and 4Statistics, Miami University, Oxford, Ohio, USA.

§Corresponding authors

Email addresses:

JTM: [mortonjt@miamiOH.edu](mailto:mortonjt@miamiOH.edu)

PA: [abrudapa@miamiOH.edu](mailto:abrudapa@miamiOH.edu)

NF: [figuernd@miamiOH.edu](mailto:figuernd@miamiOH.edu)

CL: [liangc@miamiOH.edu](mailto:liangc@miamiOH.edu)

JEK: [karroje@miamiOH.edu](mailto:karroje@miamiOH.edu)

# Abstract

**Background:** mRNA polyadenylation, the addition of a poly(A) tail to the 3’-end of pre-mRNA, is a process critical to gene expression and regulation in eukaryotes. To understand the molecular mechanisms governing polyadenylation and other relevant biological processes, it is important to identify these poly(A) tails accurately in transcriptome sequencing data and differentiate them from artificial adapter sequences added in the sequencing process. But the annotation of these tails is complicated by the presence of sequencing errors and post-transcriptional modifications. While determining that a tail is present in a given transcript fragment, these obfuscations make the problem of boundary identification a challenge; conventional seed-and-extend algorithms struggle to accurately identify these poly(A) tail end-points. Further, all existing tools that we are aware of focus exclusively on the trimming of poly(A) tails, failing to provide the detailed information needed for studying the polyadenylation process.

**Results:** We have created SCOPE++, a tool for finding the precise border of poly(A) tails and other homopolymers in raw mRNA sequence reads. Based on a Hidden Markov Model (HMM) approach, SCOPE++ accurately identifies specific homopolymer sequences in error-prone EST/cDNA data or RNA-Seq data at a speed appropriate for large sequence sets.

**Conclusions:** We demonstrate that our tool can precisely identify poly(A) tails with near perfect accuracy at the speed required for high-throughput applications, providing a valuable resource for polyadenylation research.

# Background

*Alternative polyadenylation* (APA) and *polyadenylated tail length variability* have recently been identified as critical mechanisms in gene expression and regulation [1-6]. But conducting studies on their exact role in the regulation process is complicated by the challenge of collecting precise information on the presence and characteristics of polyadenylated (poly(A)) tails in transcriptome data. While finding a significantly long stretch of adenine bases in a sequence transcript is not difficult, the challenge deepens when you try to account for sequence modifications that could obscure the tail sequence (e.g. base-call errors, the effect of processes such as RNA editing, or sequencing artifacts). Tools such as SeqClean [7], TrimEst [8], or SeqTrim [9] are able to effectively remove poly(A) tails via truncation, but cannot recover the detailed information needed when studying issues related to length variation. Thus a tool is required that is able to identify poly(A) tail boundaries and length, and is robust to disruptions in the homopolymer sequence.

Polyadenylation is a post-transcriptional process in which the 3’-end of a pre-mRNA is cleaved and replaced with a poly(A) tail to form a mature mRNA. Specifically, the polyadenylation protein complex binds to poly(A) signals, then cleaves the sequence at a *poly(A) site*, and finally collaborates with the poly(A) polymerase to perform non-templated adenine addition a few bases downstream of the appropriate poly(A) signals [1]. The poly(A) tail at the 3’ end of the mRNA is the hallmark of mRNA maturation, and also serves as a regulatory signal that is critical for mRNA nucleus-to-cytoplasm transportation, mRNA stability, and protein translation [1,2]. Recent research suggests that many eukaryotic genes employ *alternative polyadenylation* (APA), in which multiple distinctive poly(A) sites are utilized to create different transcript isoforms from the same gene [1-3]. It is clear that APA is an important regulator in eukaryotic gene expression and regulation. For example, 3’-UTR shortening by APA appears to be highly active in cancer cells[4]. To increase our understanding of the underlying molecular and biological mechanisms governing polyadenylation and other relevant processes, RNA-seq data is continually being generated to aid in annotating the junctions of the 3’-UTR and the poly(A) tail [5,10]. Moreover, 3’-end tagging (i.e., addition of non-templated U or C/U-rich tags) and 3’-oligouridylation (i.e., poly(U) tails) have been shown to affect mRNA degradation and are common in many eukaryotic species[6,8]. Studies have also shown that the length of the poly(A) tail has a direct effect on mRNA stability, and mRNAs with short poly(A) tails can be stored in cytoplasm and reactivated later for translation by a re-polyadenylation process that elongates the tail lengths [11,12].

Data relevant to the role of polyadenylation in gene regulation is best collected from sequenced mRNA and cDNA sequences. However, the precise identification of poly(A) tails embedded within those sequences is a challenge, as the search for a contiguous sequence of adenine bases is complicated by the potential obfuscation of the sequence pattern by base-call errors and the presence of sequencing-induced artificial sequences (e.g., adapters, linkers and primers) added near the poly(A) tails. Such modifications, as well as those from poorly understood biological processes (such as RNA editing and non-templated nucleotide addition [13,14]), can disguise the characteristic adenine sequence and result in impurity in the poly(A) tails, frustrating traditional seed-and-extend algorithms. Certain tools, such as SeqClean [7], TrimEST [8], and SeqTrim [9], are able to reliably remove such tails by identifying one end and truncating it, and are hence are useful for those interested in obfuscated coding sequence. But they are not able to provide the information needed to study the polyadenylation process itself.

In this paper, we introduce SCOPE++, an open-source software tool employing a Hidden Markov Model approach for the precise identification of the boundaries and length of poly(A) tails and other homopolymers in sequence reads. SCOPE++ runs at a speed appropriate for Next Generation Sequencing output sizes, with a capability to self-tailor its computational model to the characteristics of a given dataset through the use of machine learning algorithms. This makes it possible to precisely study tail length, alternative expression, and their roles in regulating gene expression. In particular, our tool is designed to accurately detect poly(A) tails of lower purity, where tail boundaries will be difficult to identify using conventional algorithms.

# Methods

SCOPE++ identifies poly(A) tails through the alignment of sequence reads to a predefined Hidden Markov Model (HMM) topology using the Viterbi algorithm. (For interested readers unfamiliar with HMM, the Viterbi algorithm, or Baum-Welch Training, a useful bioinformatics-oriented overview is presented in *Durbin et al.* [15], with a more technical review in *Rabiner* [16]). Employing sliding windows to initialize HMM parameter values tailored to the dataset, SCOPE++ utilizes the Viterbi algorithm to approximate the most likely position of a poly(A) tail within any given fragment. [15]

**HMM Topology**

SCOPE++ identifies poly(A) tails using a variable-state Hidden Markov Model [15] conforming to the topology illustrated in Figure 1(a), allowing for the identification of both perfect and imperfect poly(A) tails in raw sequence reads. Using sliding windows along a random sampling of the input as training data for initial parameter values (optionally coupled with Baum-Welch training for HMM parameter optimization), SCOPE++ utilizes the Viterbi algorithm to approximate the most likely position of a poly(A) tail within any given fragment. In Figure 1(a) we see the HMM topology, and in 1(b) the decomposition of a fragment. In the alignment of the fragment each base will be assigned to either: the poly(A) state (thus indicating it is a member of a poly(A) tail); a background state (thus indicating it is not); or one of the intermediate A states. These middle states serve as tail boundary bases, which we require to be error free. Increasing the number of such states can increase the precision of boundary identification, at the cost of sensitivity as base-call errors become more likely to appear within the defined end region. Experiments indicated that using four such states (split two and two) achieves a reasonable balance.

The topology of the HMM is fairly simple, but requires the fragment conform to a certain template form. Specifically, any embedded tail must have *x* error-free adenine bases at either end (where *x=2* in Figure 1(a)), and the tolerance of non-adenine tail bases will be dictated by probabilities assigned in the poly(A) state. There is the possibility of a non-tail adenine sitting by chance close to the tail and being thus included in the tail. However, there is virtually no way to distinguish such an aberrant base using only sequence information (or even establish whether or not that A should be outside the tail without wet-lab conformation), hence any tool will likely suffer from this. Two homolopolymers sitting in close proximity could be merged (with the intervening bases labeled as errors within the tail). However, the occurrence of several consecutive non-A bases will result in a poor fit to the model, and hence are unlikely to be accepted. In practice, the occurrences of two homolopolymers actually occurring close enough to be a problem appears to be very smaller.

**Parameter Estimation**

Starting with this fixed topology and a set of fragments, we needed to estimate several parameters for the HMM which may be dependent on the characteristics of the biological data or the sequencing processes. Specifically, we needed transition probabilities (the probability of moving between the homopolymer and background regions in the fragments), and emission probabilities (the base distribution for each fragment type). Note that the probabilities involving the end-segment states (the “A” states) are fixed: each of these states must transition to the next state with a probability of 1. In other words, we required that *x* bases at each end of the tail be adenine(using a user specified *x* with a default value of 2).

Given a large set of input data and the fixed HMM topology, we estimate the remaining parameters by taking a sampling of the fragments, quickly determining the approximate location of the poly(A) tails, and using that data for the basis of the estimation Specifically: we apply a sliding-window scoring filter of user-specified width *M* to a random sampling of the input data. Each window is scored by calculating the difference in the number of adenine and non-adenine bases. (Or thymine v. non-thymine: the tool is able to search for poly(T) tails as well, though for purposes of this discussion we will confine ourselves to discussing poly(A) searches.) As illustrated in Figure 2: we identify a window *WCenter*with the highest such score and, if the score is larger than a pre-determined threshold, we treat this as a potential subsequence of a poly(A) tail. We then find the first window to the left, *WLeft*, whose score is equal to the threshold value, and identify a *WRight* in a symmetric fashion. Finally, we merge together all windows to form one contiguous sequence representing a hypothetical poly(A) tail to be use for parameter training (as illustrated in Figure 2). (We note that this method intentionally tends towards the over-estimation of the actual boundaries, allowing us to compensate for the presence of base-call errors at the ends of the poly(A) tail.) From these putative sets we can then estimates the HMM parameters: transition probabilities are based on mean tail length and the assumption that the length is geometrically distributed, while emission probabilities are sampled directly from the contents of the putative fragments.

Following this, estimates can be further refined using the Baum-Welch algorithm [15], which is designed to optimally fit an HMM model to a given training dataset. In practice, we find that use of Baum-Welch increases runtime with a negligible improvement in results – it appears that our initial estimate of the model parameters are fairly close to the estimates returned by Baum-Welch, hence the algorithm is generally not worth the text time.

Once we have HMM parameter estimates, SCOPE++ independently applies the Viterbi algorithm [15] to each sequence, getting back a the best fit of the sequence to our model (and thus an assigned of each base to one of the characterizing states). While Viterbi can be time consuming in many applications, in this situation that is not the case. Specifically: while the Viterbi algorithm generally takes *O(mn2)* time (where *m* is the length of the sequence and *n* is the number of states in the model), the structure of our model allows us to reduce that to *O(mn)* time. As the number of states employed is essentially fixed, in practice the runtime is linear in the size of the sequence fragment.

Benchmark Sets: Given the lack of benchmark sets on which to compare tools, we have developed a “semi-synthetic” benchmark set: namely, we have taken a set of real sequences with human-identified tails and artificially removed the tails of all impurities (that is: we convert all non-adenine bases to adenine). The result is a set of synthetic sequences with characteristics highly correlated to real data. While the human annotation is likely the subject of human error, by cleaning them of impurities we have essentially turned the human-identified tails into actual (simulated) tails that should be identified by our tool. The non-tail portions of our synthetic fragments are direct copies of actual biological sequences, hence contain within them any sequence characteristics that might have an effect on the performance of the tool. We can then introduce simulated base-call error at a controlled rate, providing large datasets with known tail locations on which to test and compare different tools. There has been some work on experimentally identifying tails[17,18], but the information produced by the studies describes tail length *distribution*, not the individual tail positions for which we are looking.

# Results

In assessing the quality of SCOPE++ results we look at two metrics: its ability to correctly identify poly(A) tails (sensitivity), and its precision in identifying tail boundaries (precision). While we looked at the specificity of SCOPE++ and other tools in our initial analysis, we found them to be all near perfect; avoiding false-positives is not a difficult task for any tool, and not worth discussion here. To assess a tool’s precision, we measured result quality in the ways: *% correct*, *average trim error* and *sum of squares error*. The first is the fraction of tails with correctly identified end points. The second looks at the average directed distance from the estimated endpoint to the actual input then the endpoint is correct. The third is computed by looking at the square of trim error. The last is important because the seriousness of boundary error increases super-linearly with the error (e.g. being off by four bases is more than twice as bad as being off by two bases), given the effect of such error on downstream analysis.

At is reasonable to ask whether the complexity of an HMM approach is warranted given the apparently simple nature of the problem, so we start by testing SCOPE++ against a basic string search. Using our semi-synthetic dataset (see Methods), we compare against am algorithm that searches for maximal substrings *s* that that: (1) *s* is no shorter than a fixed value *m*, and (2) the fraction of non-A bases in *s* is no greater than a fixed value *p*. We find that while this basic algorithm can effectively detect the presence of poly(A) tails, it cannot match SCOPE++ in boundary detection accuracy. Experiments indicate that an effective parameter assignment is *m=10* and *p=0.15*, and with those values we find both tools have near-perfect sensitivity (and specificity) for data subjected to a base-call error rate ranging from 0 to 0.1 errors per base. But the simple algorithm suffers in boundary precision. With a 0% errors rate (i.e. perfect tails), the simple tool is able to correctly identify boundaries less than 25% of the time, and over-extends the boundaries by an average of 8 bases; under these conditions SCOPE++ returns perfect results. Bumping the error rate improves the simple tail results slightly (as more errors in the tail prevents over-extension). But even at a 5% error rate the tool can only identify 34% of the boundaries correctly, and over-extends by an average of 4.5 bases. SCOPE++ correctly identifies end points 73% of the time, and those boundaries it misses are short by an average of 0.6 bases. In short: the simple algorithm is useful if we are merely interested in determining if a poly(A) tail is present, but the complexity of SCOPE++ is required if precision is important.

Our second experiment is to determine whether the complexity of the HMM is appropriate structural match for the biological phenomenon. To test this we compare SCOPE++ as described (using *x=2*, thus requiring each tail start and end with 2 As) against a basic two state HMM (containing a “poly(A)” sate and a “non-poly(A)” state). Using the same training procedure for the simplified HMM, the tool tends to assign all bases to the non-poly(A) state. He simplified HMM is able to identify only about 10% of the fragments having tails (as opposed to SCOPE++’s 100%), even when dealing with perfect tails: while tail identification is simple, it is not that simple. We conclude from this both that the complexity of our model is in some way describing the appropriate biologically-dictated sequence structure – at least beyond what the trivial HMM.

Having justified the idea behind the general approach, , we now move to a comparison to against existing tools. For this we looked at the TrimPoly module of SeqClean [7] and TrimEST [8]; SeqTrim [9] was not used due to its slower runtime. Table 1 displays the average sensitivity and specificity of SCOPE++, TrimPoly [7], and TrimEST [8] over the six different semi-synthetic data sets (described in Methods). For each simulation we start with our benchmark dataset of 500 “cleaned” sequences, and for each sequence we randomly generated 200 simulated sequences by stochastically introducing error into the poly(A) tail, as well as 200 sequences without tails by randomly sampling from the non-tail portion of the sequence. Altogether, this gives us 10,000 tail-containing fragments derived directly from actual fragments and 10,000 tail-omitted fragments based on similar base distributions. On these results we saw almost perfect sensitivity and specificity in both the SCOPE++ and TrimPoly tools.

However, when looking at the *precision* at which boundaries can be identified, we see a different story. We find that SCOPE++ identifies the correct boundaries a significantly higher number of tails (using the 3% simulated error rate, SCOPE++ correctly identified an average of 77% of the sequence boundaries as averaged, as opposed to 28% of for TrimPoly), and has a significantly smaller average sum-of-squares error rate. In short, SCOPE++ is considerably more precise.

We also look at the performance of SCOPE++ and TrimPoly as a function of both tail length and the length of adjacent sequencing adapter sequence (a portion of the fragment added downstream of the tail as an artifact of the sequencing process). We find that, while TrimPoly performs better for identifying very short poly(A) tails (< 20 bp), it is highly sensitive to the length of any adapter remaining on the fragment; SCOPE++ is completely robust to such interference. By augmenting the length of tails or adjacent sequencing adapters fragments of real data in simulation, we can examine the effects of tail length and adapter length on sensitivity (see Figure 3(a)). We observed a significant decrease in the sensitivity of TrimPoly as the adapter length grows beyond the length of the tail – a factor having no effect on the significance of SCOPE++. In Figure 3(b) we plot sensitivity as a function of tail length for fixed adapter segment length. Once again, while SCOPE++ can handle any such values, TrimPoly suffers in the presence of adapter segments longer than 250 bp – falling to a sensitivity of less than 0.4. In Figure 4 we verify this assertion with real data – data set ERR125556 [CITE]. Figure 4(a) shows the number of fragments identified as containing a tail and an adapter of length ≥ m (with adapter lengths for TrimPoly identified sequences determined by an ad-hoc post processing scan), while in Figure 4(b) we see the same information as a percentage of the total identified (i.e. the sensitivity). The findings are consistent with the simulation: as the tails shift deeper into the fragment, the relative ability of TrimPoly to identify them diminishes significantly.

**Large Dataset Validation**

SCOPE++ was run against a 17 GB Arabidopsis dataset [CITE: Arabidopsis]. The Arabidopsis data set was developed using poly(T) tag sequencing and it was estimated that around 60% of the reads contained poly(T) tails. Our tool discovered that 59.6% of all of the reads contained poly(T) tails, which was very close to the initial estimate. A quick search for sequences containing short homopolymers returns only 126 clear false negatives, all of which were present in low quality regions and probably would have been trashed anyway.

# Availability and Requirements

Our software can be found at <https://github.com/mortonjt/SCOPE>, with code to be distributed as open source. SCOPE++ was implemented in C++11, and has been tested using the gnu g++ compiler (v. 4.6) on both OS X and Ubuntu Linux.

# Conclusions

Here we have presented SCOPE++, a novel approach for identifying imperfect homopolymers when end-point precision is important. We have compared SCOPE++ to simpler tools, and in doing so established the necessity of a more complex approach to what initially seems like a simple problem. Unlikely other tools, SCOPE++ is able to identify both end boundaries of a tail (as opposed to simple trimming) and is able to train on data-specific model parameters. Our tests have shown that the tool performs on par with existing poly(A) trimming tools in terms of speed, while showing considerably more accuracy in terms of identifying end-point boundaries and predicting tails buried further away from the fragment ends.

The success of SCOPE++ also indicates some potential for generalizing the approach to related problems, such as the edification of low-complexity repeats or simple artifacts that might need to be weeded out of a sequence in a precise manner. While HMM approaches are frequently time consuming and appear to be over-kill for this sort of problems, in these cases where we can keep the complexity of the model fairly low and the state space small, the same principles might work to keep the runtime within reason.

# Authors' contributions

JM designed, developed and implemented most of the software algorithms. PA conducted human validation on the tests. NF created found the optimal parameters of the software. JK coordinated this project and conducted all of the simulation tests and performed a statistical analysis on the results. LC coordinated this project and conducted human validation on the results. All authors read and approved the final manuscript.

# Acknowledgements

Funding: This project was funded partially by the National Science Foundation (No. O953215 to JK) and the NIH-AREA (1R15GM94732-1 A1 to CL).

# Figures

Figure : (Left) A generalization of the HMM topology used for identification. The number of A states between the background and poly(A) states can be varied depending on user priorities, with a larger number of states leading to more precise boundary identification but loss of overall sensitivity. Using a total of four states (split two and two) achieves a reasonable balance in practice. (Right) Example of a sequence read containing an identified poly(A) tail that has been divided into four types of segment: background (bases not part of the poly(A) tail), upstream tail-ends (tail bases within *x* of the upstream tail boundary), interior tail bases (further than at least *x* bases away from either boundary), and down-stream tail-ends.

Figure 2: The sliding window scoring scheme used to provide an estimate of the HMM parameters. (a) Illustration of score calculation. (b) Finding *Wcenter*.

Figure 3: (a) Plot of tool sensitivity as a function of the ratio of adapter segment length to tail length: while SCOPE++ is robust to this ratio, TrimPoly sensitivity deteriorates as the length of the adapter segment becomes longer than that of the tail.



Figure 4: Application of tools to the Triticum aestivum ERR125556 fragment set [19]. With 140,000 tail-containing sequences identified, (a) shows the number of sequences with adapter segments of length ≥ m that are identified by SCOPE++, TrimPoly, or both. In (b) we see the percentage of those sequences that were identified only by SCOPE++ or TrimPoly.

**Table 1:** Results of tests on data-derived simulated sets using SCOPE++, the TrimPoly component of SeqClean [7], and TrimEst [8]; SeqTrim [9] was too slow for testing at the necessary scale. Starting with a set of 500 human-annotated sequenced transcript fragments (Illumina-sequenced Arabidopsis, 454-sequences Chalamydomons, and Sanger-sequence Human [CITE: Conifergdb]), we remove all sequencing errors within the human identified poly(A) or poly(T) tails, then introduce simulated error into those tails to allow the testing of sensitivity and boundary condition at a controlled error rate. For specificity we generate sequences using the base distribution of the non-tail segments of the sequences in the set. Applying each tool to the set, we then have a reference solution that allows us to compute result qualities.

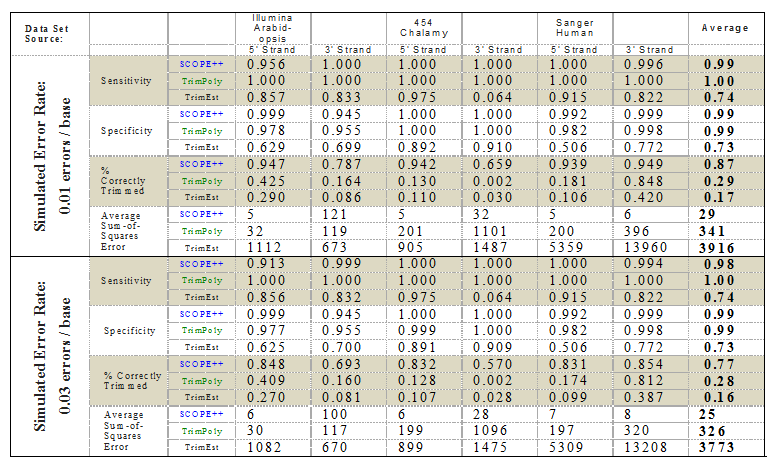


Table 2: In SCOPE++ trim option, poly(A) tails and poly(T) tails can be simultaneously trimmed. The table below lists all possible arrangements of poly(A) and poly(T) tails within a single read and the frequencies in which they appear in the actual dataset [20].

|  |  |  |
| --- | --- | --- |
| Trim type | Number identified | Percent identified |
| [TTTT.....................]: | 45577831 | 53.1196% |
| [.....................TTTT]: | 679815 | 0.792303% |
| [AAAA.....................]: | 15734 | 0.0183375% |
| [.....................AAAA]: | 83200 | 0.096967% |
| [TTTT............TTTT]: | 2864767 | 3.3388% |
| [AAAA............AAAA]: | 340 | 0.000396259% |
| [TTTT............AAAA]: | 2733703 | 3.18605% |
| [AAAA…….....TTTT]: | 1059 | 0.00123423% |
| No homopolymers | 33845914 | 39.44636558% |

[1] N.J. Proudfoot, Ending the message: poly(A) signals then and now, Genes & Development. 25 (2011) 1770–1782.

[2] D. Xing, Q.Q. Li, Alternative polyadenylation and gene expression regulation in plants, WIREs RNA. 2 (2010) 445–458.

[3] D.C. Di Giammartino, K. Nishida, J.L. Manley, Mechanisms and consequences of alternative polyadenylation, Mol. Cell. 43 (2011) 853–866.

[4] C. Mayr, D.P. Bartel, Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells, Cell. 138 (2009) 673–684.

[5] X. Wu, M. Liu, B. Downie, C. Liang, G. Ji, Q.Q. Li, et al., Genome-wide landscape of polyadenylation in Arabidopsis provides evidence for extensive alternative polyadenylation, Proceedings of the National Academy of Sciences. 108 (2011) 12533–12538.

[6] Y.S. Choi, W. Patena, A.D. Leavitt, M.T. McManus, Rna. 18 (2012).

[7] DFCI Gene Indices Software Tool, (n.d.).

[8] P. Rice, I. Longden, A. Bleasby, EMBOSS: the European Molecular Biology Open Software Suite, Trends Genet. 16 (2000) 276–277.

[9] J. Falgueras, A.J. Lara, N. Fernández-Pozo, F.R. Cantón, G. Pérez-Trabado, M.G. Claros, SeqTrim: a high-throughput pipeline for preprocessing any type of sequence reads, BMC Bioinformatics. 11 (2010) 38.

[10] F. Ozsolak, P. Kapranov, S. Foissac, S.W. Kim, E. Fishilevich, A.P. Monaghan, et al., Comprehensive Polyadenylation Site Maps in Yeast and Human Reveal Pervasive Alternative Polyadenylation, Cell. 143 (2010) 1018–1029.

[11] C. Bonaïti, S. Parayre, F. Irlinger, Novel extraction strategy of ribosomal RNA and genomic DNA from cheese for PCR-based investigations, Int. J. Food Microbiol. 107 (2006) 171–179.

[12] R. Mendez, J.D. Richter, Translational control by CPEB: a means to the end, Nat. Rev. Mol. Cell Biol. 2 (2001) 521.

[13] Y.W. Cheng, L.M. Visomirski-Robic, J.M. Gott, Non-templated addition of nucleotides to the 3' end of nascent RNA during RNA editing in Physarum, Embo J. 20 (2001) 1405–1414.

[14] Y. JIN, Nontemplated nucleotide addition prior to polyadenylation: A comparison of Arabidopsis cDNA and genomic sequences, Rna. 10 (2004) 1695–1697.

[15] R. Durbin, S.R. Eddy, A. Krogh, G. Mitchison, Biological Sequence Analysis, Cambridge University Press, 1998.

[16] L. Rabiner, B.H. Juang, An introduction to hidden Markov models, ASSP Magazine, IEEE. 3 (1986) 4–16.

[17] A.O. Subtelny, S.W. Eichhorn, G.R. Chen, H. Sive, D.P. Bartel, Poly(A)-tail profiling reveals an embryonic switch in translational control, Nature. 508 (2014) 66–71.

[18] H. Chang, J. Lim, M. Ha, V.N. Kim, TAIL-seq: genome-wide determination of poly(A) tail length and 3' end modifications, Mol. Cell. 53 (2014) 1044–1052.

[19] Triticum Aestivum, (n.d.).

[20] SCOPE++: Github Repository, (n.d.).