



ESTprep: preprocessing cDNA sequence reads

Todd E. Scheetz^{1,2,*}, Nishank Trivedi³, Chad A. Roberts³, Tamara Kucaba⁴, Brian Berger⁴, Natalie L. Robinson³, Clayton L. Birkett³, Allen J. Gavin³, Brian O'Leary³, Terry A. Braun^{1,3,5}, Maria F. Bonaldo⁴, John P. Robinson³, Val C. Sheffield^{4,8}, Marcelo B. Soares^{4,6,7} and Thomas L. Casavant^{1,3,5}

¹Center for Bioinformatics and Computational Biology, ²Department of Ophthalmology, ³Department of Electrical and Computer Engineering, ⁴Department of Pediatrics, ⁵Department of Biomedical Engineering, ⁶Department of Biochemistry, ⁷Department of Physiology and ⁸Howard Hughes Medical Institute, The University of Iowa, Iowa City, IA 52242, USA

Received on October 27, 2002; revised on November 3, 2002; accepted on February 6, 2003

ABSTRACT

Motivation: High accuracy of data always governs the large-scale gene discovery projects. The data should not only be trustworthy but should be correctly annotated for various features it contains. Sequence errors are inherent in single-pass sequences such as ESTs obtained from automated sequencing. These errors further complicate the automated identification of EST-related sequencing. A tool is required to prepare the data prior to advanced annotation processing and submission to public databases.

Results: This paper describes ESTprep, a program designed to preprocess expressed sequence tag (EST) sequences. It identifies the location of features present in ESTs and allows the sequence to pass only if it meets various quality criteria. Use of ESTprep has resulted in substantial improvement in accurate EST feature identification and fidelity of results submitted to GenBank.

Availability: The program is freely available for download from <http://genome.uiowa.edu/pubsoft/software.html>

Contact: tscheetz@eng.uiowa.edu

1 INTRODUCTION

A critical aspect of any large-scale sequencing effort is the production of high quality data. The increasing number of large-scale DNA and cDNA sequencing projects, and the availability of high volume automated sequencing, requires an efficient processing method. Computer software has played a critical role in enabling large-scale sequencing projects.

This paper presents a new program, ESTprep, which has been continuously used for preprocessing in several cDNA-based gene discovery projects at the University of

Iowa since 1997. This program is an essential component of the sequence analysis and processing pipeline. The objectives of ESTprep include identification of expected EST (expressed sequence tag) features and assurance that only high-quality sequences proceed to the later stages of the pipeline. The EST features that are identified include restriction site, cloning vector, polyadenylation tail, library tag, and polyadenylation signal. ESTprep is highly configurable allowing simple incorporation into any high-throughput sequence processing environment.

1.1 Background

Preprocessing includes base-calling, filtering of low-quality sequences, identification of sequence features, and report generation. Most of the existing programs used for preprocessing deal with a single step of preprocessing, such as phred (Ewing *et al.*, 1998) for base-calling, or RepeatMasker (<http://ftp.genome.washington.edu/RM/RepeatMasker.html>) for identifying known repetitive elements and low complexity sequences. There are packages available that perform automated sequence preprocessing, covering numerous aspects discussed here. For example, Gasp (genome automated sequence preprocessor; Wendl *et al.*, 1998) is a suite of utilities that provides functionality for compilation of quality statistics, disk management and report generation. Similarly, the Staden Package (Staden, 1996) includes utilities such as pregap (Bonfield and Staden, 1995), gap assembler (Bonfield *et al.*, 1995), and various other tools for trace viewing, mutation detection, etc. A final example is hopper (Smith *et al.*, 1997) which provides similar functionality including report generation. However, almost all of these sequence preprocessing packages are environment specific, focusing on the needs of the laboratory for which they were developed. Use of such

*To whom correspondence should be addressed.

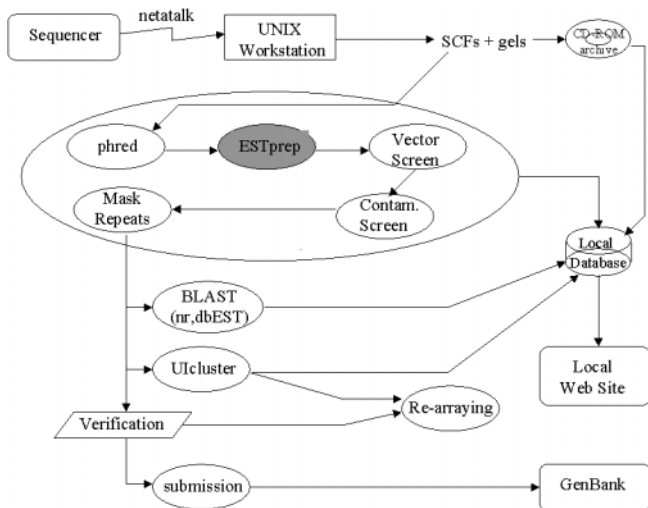


Fig. 1. Data-flow diagram of sequence processing pipeline at the University of Iowa Coordinated Laboratory for Computational Genomics.

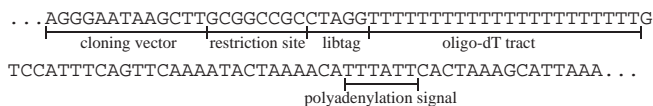


Fig. 2. Features expected in a 3' EST sequence. The labeled segments represent cloning vector, restriction site (NotI shown), polyadenylation tail, polyadenylation signal, and library tag, respectively, within an EST sequence.

programs typically requires changes in implementation and design of either the program or the protocols within the laboratory itself.

ESTprep was designed to satisfy the requirements of cDNA-based gene discovery projects in general. Figure 1 shows a data-flow diagram of the sequence processing pipeline in which ESTprep is a component (Scheetz and Casavant, 2003). The pipeline consists of five distinct stages: data collection, quality screening, feature annotation, clustering, and submission to public databases. The quality screening and feature identification stages are the main focus of ESTprep.

An EST is a partial copy of an mRNA transcript obtained from a cDNA clone. High-throughput gene discovery projects typically sequence large numbers of ESTs to identify a comprehensive set of transcribed genes. Figure 2 shows an EST sequence that has been annotated for several of these features.

An EST can be generated by sequencing from either end of a cDNA clone. A 5' EST is in the same linear orientation as the progenitor mRNA molecule. A 3' EST

is in the opposite direction, the reverse complement of the progenitor mRNA sequence. Regardless of direction, the most prominent and first feature to be identified in an EST is the restriction site; a short sequence that specifies a restriction enzyme cleavage site. Figure 2 shows a common set of features for 3' ESTs, including the NotI restriction site, GCGGCCGC. Prior to the restriction site is a subsequence of the cloning vector or polylinker. These two features should be present in every EST, regardless of direction. ESTs lacking a detectable restriction site are rejected. The two most common reasons for lacking a detectable restriction site are an abundance of errors early in the sequence, or a delay in beginning data acquisition. In 5' ESTs the restriction site and cloning vector are the only two features present, and the restriction site is immediately followed by the sequence obtained from the cDNA insert. Unlike 5' ESTs, those derived from the 3' end normally contain several additional features. The restriction site may be immediately followed by a library tag, if present. Library tags provide useful information regarding the tissue of origin for sequences from pooled cDNA libraries. To prevent the loss of tissue source annotation, several bases are synthetically introduced. This is referred to as the library tag. At the University of Iowa, we use a novel method for creation and robust identification of library tags (Gavin *et al.*, 2002). This method employs short oligos and provides increased accuracy in identification of tissue source with little added processing time. In addition to the library tag, a polyadenylation tail may be present. In 3' ESTs, this is a stretch of the letter T immediately following the library tag or restriction site. A polyadenylation signal may be found from 11 to 30 base pairs from a detected polyadenylation tail. The library tag, polyadenylation tail and signal features may not always be present in every 3' sequence. ESTprep identifies all of these features which might be present in an EST sequence and then judges the high quality region of insert in the remainder of the sequence. Hence, it acts as a 'high-pass' filter for the pipeline, removing low quality sequences from further processing.

2 SYSTEMS AND METHODS

The primary objective of ESTprep is to identify a set of features within ESTs. As mentioned earlier, the only required feature of an EST is the restriction site. All other features are optional. Typically, vector sequence is found upstream of the restriction site. In addition, 3' ESTs may also include a library tag, polyadenylation tail and polyadenylation signal.

Figure 3 shows an overview of the steps used by ESTprep in processing each sequence. The inputs to the process are the sequence files in FASTA

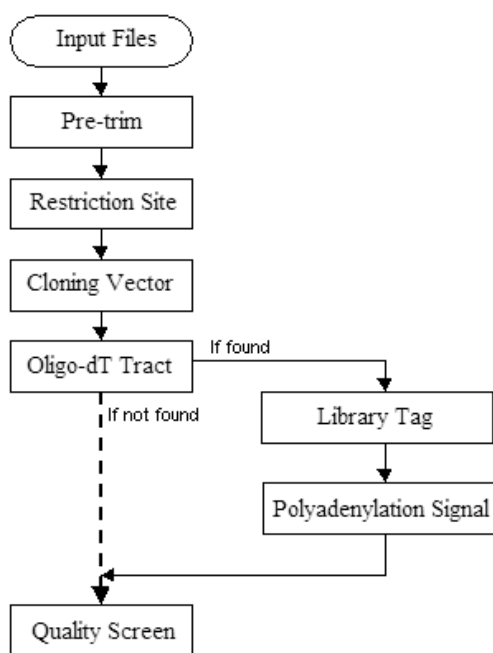


Fig. 3. An overview of steps involved in ESTprep while processing a sequence. Only 3' annotated ESTs require all steps. Oligo-dT tract, library tag and polyadenylation signal searching stages are skipped for 5' ESTs.

format, and the associated quality files obtained from phred. Parameters required for analysis are read from a parameter file. Figure 4 presents a complete parameter file used in processing ESTs from the UI-R-A0 cDNA library. Output of the analysis for each sequence is saved in a tagged-format file using the input file name with a suffix of *.prpsmry*. Additional (human-readable) output can also be printed to the standard output with or without recapitulating the input sequence and quality file contents. Examples of the human-readable and tagged-format files are provided in Figures 5 and 6 respectively.

The fundamental component of the processing performed by ESTprep is the sequence comparison to expected features (e.g. the restriction site). Rather than using a more generic alignment algorithm, such as Smith and Waterman (1981), a heuristic match function is used which accommodates the particular expected error modes of each feature, and its juxtaposition to other features. This function identifies a subsequence of length M , in which at least N bases are in common between the two sequences. Although this can be more computationally intensive for some larger sequences, for all features of interest (e.g. restriction sites and polyadenylation signals), this approach is adequately efficient and provides more useful feedback for annotation purposes.

3 ALGORITHM

3.1 Pre-trimming

Often the initial sequence of an EST is of very low quality. This can hinder accurate feature detection unless low-quality sequence is pre-trimmed. ESTprep trims this region of high error rate prior to identification of sequence-based features. The identification of the region to be trimmed is performed using the parameters provided in the *prep.params* file (shown in Fig. 4) for an acceptable number of low-quality bases within a specified length of the initial subsequence. Using the default parameters as shown in Figure 4, there must be at least nine bases within a window of 20 that are of phred quality value of 10 or more. The left-most position in the sequence such that this criterion is met upstream (to the left lexically) is the trim site.

3.2 Restriction site detection with vector validation

Locating the restriction site is the foundation of the feature identification process. If a restriction site is not found, further analysis is discontinued and the sequence is abandoned. To validate the identified restriction site, ESTprep looks for an associated segment of the vector sequence upstream.

The restriction site identification process begins by first searching for perfect matches. Only if none are found, are imperfect restriction sites identified. The tolerance for errors within the restriction site is a user-configurable parameter. The default criteria (see Fig. 4) require finding the site in the first 200 bases with one gap (insertion or deletion) and up to two substitution errors. No more than two errors total (gap plus substitution) are permitted by default. Once the set of restriction site candidates has been identified, ESTprep attempts to locate a region of vector sequence immediately upstream of the candidate sites. The best restriction site candidate with significant leading vector is selected as the identified restriction site. In some cases, the distribution of errors within the restriction site prohibits accurate identification. In such a situation, a longer feature is assessed allowing amortization of the errors. This feature contains the n -base restriction site plus the preceding n bases of vector sequence. In doing so, ESTprep further relaxes the error parameters to approximately double those permitted for a site of length n . A restriction site identified using this expanded motif is categorized as a weak site. Similarly, if less than the specified amount of vector sequence is found (as specified in parameters), the sequence is tagged as containing weak vector. It is left to a later stage of the pipeline shown in Figure 3 to reject or accept based on this annotation.

```

# params for A0 library
# Sequence_Direction: FORWARD/REVERSE
Sequence_Direction: FORWARD
# Echo_Sequence: 0 -> no output; 1 -> results only; 2 -> results and input
Echo_Sequence: 2
# Quality-Params: threshold n_below_threshold in_m min_good
Quality-Params: 10 8 20 100
# Restriction-Site-Tag: must-find-before n-allowed-missing n-allowed-wrong ...
# ... n-allowed-wrong-or-missing tag-value
Restriction-Site-Tag: 200 2 1 2 GCGGCCGC
# Cloning-Vector: min-nbases-to-find vector-segment
Cloning-Vector: 6 GCCAAGCTAAAATAACCTCACTAAAGGGAATAAGCTT
# LibTags: n-library-tags n-allowed-wrong n-allowed-missing
# tag1 tag2 ...
# tissue1 tissue2 ...
LibTags: 9 0 0
TTCCA TAGAG CACAC CAAAC ACAAC ATGTG GAGA TCAC AAG
Lung Brain Liver Kidney Heart Placenta Spleen Ovary Muscle
# polyAtail: char-to-look-for minlength percent-of-'A's max-dist-from-restrsite
# ... lengthofTail2Keep window-Internal-primed percent-of-'A's-window
polyAtail: T 10 .95 20 18 18 .65
# polyAsignal: min-dist-from-polyA-tail max-dist-from-polyA-tail-end
# ... #of-recognized-forms ...
# ... n-allowed-missing n-allowed-wrong n-allowed-wrong-or-missing form1 form2 ...
polyAsignal: 11 30 2 0 0 0 TTTATT TTTAAT
polyAsignal-alt: 14 TTTACT TTTATA TTTATG TTTATC TTTAGT TTTAAA TTTCTT CTTTTT TGTTTT AGCCCC TATATT TGTATT TCTATT TTCATT

```

Fig. 4. Example of a parameter file used by ESTprep.

3.3 Oligo-dT tract detection

As mentioned above, identification of a putative polyadenylation tail is specific to 3' ESTs. The strategy used is to identify an oligo-dT tract — a window of maximum density of the nucleotide specified on the polyAtail line of the prep.params file shown in Table 4 ('T' for 3' ESTs). The other parameters on that line specify the minimum length, the minimum density of the polyA letter, and the maximum distance allowed between the end of the restriction site and the beginning of the polyadenylation tail. Due to the procedures used in cDNA library construction, the length of the oligo-dT tract is minimized, hence typically a stretch of 20–30 T's is present rather than hundreds. The search for the oligo-dT tract is begun downstream from the identified restriction site. If ESTprep fails to identify a suitable region using these parameters, the search continues for evidence of a 'weak' oligo-dT tract. Here, a weak oligo-dT tract is characterized by a significant increase in the local density of the nucleotide 'T'. In these cases, the end of the tract is difficult to determine accurately. Therefore, although the library tag *may* be identified, no attempt is made to identify the polyadenylation signal.

3.4 Library tag detection

A list of possible library tags, and their associated tissues is also provided to ESTprep through the parameter file, along with a specification of the error tolerance

for the tags. The detection algorithm searches for the closest match to each provided library tag upstream of the oligo-dT tract. The best match is reported as the library tag. Two sets of parameters are configurable for library tag detection. The first is a description of the set of tags to be identified, including the number of tags. The second parameter set specifies the possible error correction properties. Depending on the set of library tags used, it may be possible to detect and correct for an arbitrary number of insertion or deletion errors. Currently, library tags are in use that can detect or correct for single insertion or deletion errors or up to two substitution errors. For complete details on error detecting and correcting library tags, please refer to Gavin *et al.* (2002).

3.5 Polyadenylation signal detection

The polyadenylation signal is a six base motif located downstream of an oligo-dT tract in 3' ESTs. The parameter file (using the polyAsignal tag) provides the location with respect to the oligo-dT tract, at which the polyadenylation signal must be found to be considered valid. Based upon published literature (Chen *et al.*, 1995), the polyadenylation signal is required to start between 11 to 30 bases downstream from the end of the polyadenylation tail by default. This subsequence is compared to the sets of polyadenylation signals provided in the parameter file. By default, only perfect matches to one of the canonical or alternative polyadenylation signals provided

Table 1. Results of ESTprep's quality control criteria for six different gels

Gel	Sequences	Rejected ESTprep		Ambiguous ESTprep		Accepted ESTprep	
		Expert	Expert	Expert	Expert	Expert	Expert
08-03-bkt-2000-96	96	6	7	0	6	90	83
08-02-bkj-2000-96	96	4	5	2	5	86	86
10-04-bqe-2000-96	96	12	13	8	15	73	68
10-04-bqw-2000-96	96	1	1	2	4	92	91
10-25-asi-2000-96	96	4	4	3	3	85	89
10-25-ash-2000-96	96	4	6	4	4	85	86

is accepted as evidence of a polyadenylation signal. Errors may be allowed in the detection of canonical signals using the same specification parameters used in the restriction site. The most common alternative signals, such as those from Beaudoin *et al.* (2000), may easily be added. These are specified using the `polyAsignal-alt` tag.

3.6 Sequence trimming and quality assessment

Once the EST sequence features described above have been identified, the regions of high and low quality within the sequence are determined. The left-most trim site is located such that a portion of the polyadenylation tail is left untrimmed if present in 3' ESTs—18 bases using the parameters from Figure 4. If a polyadenylation tail was not identified, then the trim site is located at the beginning of the restriction site. The right-most trim site is determined based upon the sequence quality criteria specified in the parameter file. Starting from the left trimming location, a sliding window method is employed to identify the point at which the sequence quality falls below the specified threshold. As shown on the `Quality-Params` line, the trim site is the left-most base of a window of size m , in this n of m are below the threshold.

In addition to the feature identification described above, ESTprep also applies additional quality criteria to each EST. The intention of these criteria is to trap objectionable sequences and to provide detailed reports to the user, identifying the cause of the problem and generating debugging data. The quality assessment methods utilize five basic metrics: polyadenylation tail length, length of high-quality sequence, average phred quality, percentage of bases above a specified phred quality value, and the fidelity of cloning vector and restriction site prior to the cDNA insert. The rejection criteria include failure to identify a restriction site, insufficient length of high-quality sequence, and low quality sequence throughout the EST as a whole. Not all quality criteria are associated with phred quality values. For example, a sequence will be rejected if the polyadenylation tail is longer than a user-specified length. This is done because the remaining sequence in ESTs with long tails is unreliable. Empirical

analysis was used to determine the default cutoff value of 70. For further quality assessment, the average quality value of the trimmed portion of the sequence, and the percentage of bases with quality value greater than 20 are examined. Both must satisfy user-specified limits. Various other limits on average quality, bases under threshold quality value, and standard deviation in quality, are employed to identify low quality sequence. The default limits were determined through analysis of empirical data, and may be adjusted as needed without recompilation. A sequence will also be rejected if, despite the pre-trimming steps, a region of low quality remains at the beginning of the sequence. The quality screening step is a crucial component of the EST processing pipeline, providing substantial saving in time and effort spent on sequences which are later utilized in providing other analyses. A detailed report of the statistics may be found in T. Kucaba (in preparation).

4 IMPLEMENTATION

The ESTprep program was first implemented in 1997 and has since incorporated numerous modifications, increasing both robustness and efficiency. The program is written in C and has been successfully tested on several variants of the Unix operating system, Windows 98 (or newer) and Mac OS X. The utilities supporting the program (e.g. the report generation program) were developed in either C or Perl depending upon the nature of the job performed by those utilities. The development of ESTprep has enhanced the efficiency of EST sequence processing significantly.

To summarize the results produced by ESTprep, several support utilities are available. For example, utilities are provided that can be used to collect analysis results for groups of sequences and generate reports. Every time a new batch of sequences is processed, these utilities update the reports, providing monthly, annual and project-wide statistics. The various reports provide information on numbers of sequences that passed the quality criteria, average read length, average quality of good sequence, locations of polyadenylation tail, polyadenylation signal

```

== Quality source file: qual/UI-R-A0-aa-d-01-0-UI.s1.qual ==

== Sequence source file: seq/UI-R-A0-aa-d-01-0-UI.s1.seq ==
0000000001111111112222222222333333333344444444445
12345678901234567890123456789012345678901234567890
-----
GGCATGCTAATATTACCTCACTAAAGGAATAAGCTTGGCGACGCCAC
ACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTATTATTTATAGTG
AAGCTTTATTTAACTCAGGAATGGATACACAATGACACAAGGGATCATA
AATTGTTATATGAAAGAGATAATAACAAGTGGATTGTGCTATAAATCCAG
AAACTGCATGTGCTTTCTGGATCTTAAGACAAGAGTCAGGATGGATGATA
GAGGGAAGGGACTGGATAAAAAACCTGAAGGGGATTGCAAGAAGCAACACA
GTACAAGCCAAAATGCCTGATCTATTCAGGAAGGAATGAAGCAGAGCTGG
TAAGTGGGTGCGAGAGGCAAAAACATTGAGAACTGGCATCGTATCAAT
GTCCTTTGGGTGAACCCAGAGATTTTCAGGTTAAAGTTCTGTAAAATGGTTG
TCAAGAATAGAAACAACCTGCATGCGGGCCAGGCCCTCTCCAACACAAGCT
CGTTTTCTGCTGAAAGAAANTATCACTTACTAAGTTTCATCCATCANAA
AGTGGCAGGGTCAAAACATCTCTGGGTGGGAACTCTTGCACTCATGCANG
CCAAAACCTGAAATGTTCCAAAGTTTGAATGGGTTTTAAAAACAAAACAC
TTTTTTTTGGGTTTTTCCCCCCCTTTTTCCCGGGGGGAAAATTTTTTAAA
AGGGGTTAAAAGGTTTTTTCCCGGTGGAAAATTTTTTTTTAAA

0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 2
1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0
-----
7 7 7 9 8 10 13 17 9 8 6 6 6 7 6 6 6 6 8 13
19 20 30 30 24 19 18 14 14 20 27 27 24 18 17 20 21 24 29 23
13 13 7 7 7 12 16 24 25 27 28 33 30 40 35 56 56 56 56 56
56 56 56 56 56 56 56 56 56 56 56 56 56 56 56 56 56 56 56
47 47 47 37 32 32 27 26 26 27 27 28 29 29 27 29 32 35 35 35
35 33 35 35 36 42 42 44 44 35 35 35 29 29 29 25 30 15 24 24
42 39 39 39 31 31 21 37 37 42 42 35 36 36 35 35 35 37 42 37
42 42 42 42 42 40 40 40 40 40 45 36 35 35 30 33 33 35 35 35
35 35 40 46 42 50 44 42 44 50 50 50 37 37 37 42 42 37 37 40
40 40 40 42 42 42 42 37 35 35 35 35 35 35 33 35 27 27 13 27
27 37 33 42 42 42 42 40 40 40 40 37 37 42 46 42 35 35 39 40
40 40 37 37 35 35 33 28 26 42 33 33 42 42 42 30 42 30 37 35
35 35 37 37 37 42 42 36 36 22 30 18 31 27 40 40 35 30 27 31
27 29 25 31 25 31 35 38 36 42 37 37 42 35 33 33 30 34 34
40 40 42 42 42 30 30 15 20 25 27 27 30 37 27 30 28 27 31 24
21 14 7 7 8 13 17 21 29 25 30 33 33 33 33 34 34 35 56 44
42 42 42 42 44 33 40 28 28 28 42 35 34 34 38 36 37 29 27 18
13 13 10 11 12 23 16 33 33 31 33 36 42 37 37 42 42 38 35 35
29 29 22 27 33 30 27 29 27 35 31 35 30 27 27 30 33 37 33 33
17 17 19 33 31 36 38 35 35 34 42 34 34 42 42 42 42 42 42 42
42 35 34 33 29 29 27 27 29 32 40 28 27 15 15 11 13 21 26
25 37 37 37 29 24 19 16 25 23 25 25 29 30 30 29 28 21 21 19
13 13 21 29 27 27 18 12 12 19 24 21 13 12 20 18 15 19 12 12
21 23 20 25 24 15 23 18 18 20 29 17 17 13 12 10 12 9 9 9
9 9 10 12 25 40 40 29 18 15 15 15 21 23 21 16 16 24 28 31
31 29 25 29 29 25 25 21 18 9 6 6 11 6 7 15 10 8 4
0 4 9 9 9 9 8 6 4 4 4 13 6 7 7 7 9 10 13 16
25 16 15 8 6 9 4 0 4 10 8 10 10 10 8 13 7 7 7 7
7 8 8 8 6 6 6 6 6 7 8 9 9 13 12 12 10 10 18 10
10 8 8 8 9 9 8 8 8 8 8 8 11 13 12 12 13 4 0 4
4 4 7 13 8 8 7 6 6 6 6 6 7 7 6 7 7 7 7 7
7 7 7 7 6 6 6 6 6 6 7 9 8 8 7 7 7 7 7 7 7
8 6 6 6 6 6 6 6 6 6 6 12 16 26 19 14 10 10 7 6
6 6 7 6 9 9 9 12 12 15 7 6 6 6 6 6 6 6 6 6
6 8 9 16 9 10 7 7 7 7 7 7 8 6 6 7 7 7 7 7 10
10 10 15 15 15 10 10 10 10 9 10 7 7 7 9 9 12 8 8 8
8 12 15 12 6 6 6 6 6 6 8 8 10 7 8 8 11 10 10 8
8 8 8 8 8
!!Restriction Site (GCGGACGC) of Length 8, with 1 Wrong,
0 Missing, and 0 Inserted Bases Found at Position: 40
+Length of Cloning Vector Found Before Restr Site: 29
+startPolyA: 53; endPolyA: 94; PolyALength: 42
+Library Tag Found. Ideal: CACAC Actual: CACAC
+PolyA Signal (TTTATT) of Length 6 Found at Position: 105
+Average Quality (trim) =31.15
-Trimmed Sequence from 77 to 511.
+-----

```

Fig. 5. Human readable output of ESTprep.

```

CLONENAME: seq/UI-R-A0-aa-d-01-0-UI.s1.seq
RESTRSITEFOUND: LOW
VECTORFOUND: TRUE
VECTORLENGTH: 29
POLYATAILFOUND: TRUE
POLYATAILLENGTH: 42
LIBTAGFOUND: TRUE
LIBTAG: CACAC Liver
POLYASIGNALFOUND: TRUE
POLYASIGNAL: TTTATT
TRIMLOC: 77 511
GOODQUALITY: 31.15
QUAL_FILTERS: 31.15 14.55
STATUS: GO

```

Fig. 6. Tagged format output of ESTprep.

and library tags. These summaries provide important data for troubleshooting and assessing the the underlying cDNA libraries. To facilitate access to these summaries, utilities are also provided to convert the summary reports into HTML formatted files, making them accessible though the web.

Configurability was one of the main design issues in the development of ESTprep. Each feature to be detected can be specified through a set of parameters, including a description of the feature and the acceptable error rates. Typically, these parameters are the same for all sequences within a batch, therefore rather than reading the parameters through the command line, parameters are read from a configuration file. Figure 5 shows a verbose (human readable) set of output for one EST. This data shows the details of the feature identified along with the sequence and corresponding quality values. Figure 6 shows the tagged-field formatted output for the same sequence. This format is used within the context of the pipeline shown in Figure 1.

5 DISCUSSION

An expert analyzed several batches of sequences to compare with the automated quality assessment used in ESTprep. Table 1 shows the result of the quality control analysis for six sets of 96 ESTs. The automated results obtained by ESTprep were compared with those of manual analysis of the chromatograph files by an expert. The assessment was performed using a double-blind analysis. Neither the program nor the expert knew of each others' classification before the analysis. Sequences which were on the border line for rejection were marked 'ambiguous' in the expert analysis. The *rejected* columns show the number of sequences rejected by ESTprep and by expert analysis. Similarly the *ambiguous* column provides the number of sequences which were marked as questionable by the expert analysis, and from those, the

number of sequences rejected by ESTprep. The *accepted* columns show the total number of sequences accepted by both analyses. Overall, ESTprep was able to identify 31 out of 36 sequences which were marked as rejected in the expert analysis. Of the 37 sequences marked as questionable by the expert, analysis ESTprep rejected 19 sequences. This confirms the 'ambiguous' nature of the sequences.

In summary, the number of sequences that were not identified as low-quality by expert analysis but were rejected by ESTprep was 15 out of 576, giving an estimate of false negative predictions of 2.6%. Only five sequences that passed the quality criteria utilized in ESTprep were identified as low-quality by the expert. Thus, the false positive rate is estimated at 5/576, or less than 1%. The utilization of heuristic, rule-based methods was the primary cause of differences between the expert assessment and that provided by ESTprep. Reduction of the prevalence of the false positives is possible, but significantly increases the false negative rate. A conservative set of defaults was used for these analyses. The choice of a conservative approach aids in removing potentially contaminated sequences from further processing. The quality-based trimming used in ESTprep was compared to phred's automated trimming (data not shown). ESTprep-based trimming was found to be comparable to that of phred using the `-trim_alt` option with the appropriate restriction site.

ESTprep has been under continuous development, testing and production use for more than 5 years. It is well suited to the needs of large-scale sequence preprocessing, providing a time- and cost-effective solution. ESTprep is under continuous refinement to improve robustness and enhance its functionality. The program has been used at the University of Iowa to process over 500 000 of EST sequences from various organisms and enjoys feedback from the cDNA construction groups of Bento Soares. The use of ESTprep has significantly reduced the time and

complexity involved in numerous gene discovery projects at the University of Iowa.

ACKNOWLEDGEMENT

We would also like to thank Dylan J. Tack, Chris Moressi, Barry Gackle, and Andy Choi for providing feedback through extensive use of ESTprep.

REFERENCES

- Beaudoing,E., Freier,S., Wyatt,J.R., Claverie,J. and Gautheret,D. (2000) Patterns of variant polyadenylation signal usage in human genes. *Genome Res.*, **10**, 1001–1010.
- Bonfield,J.K., Smith,K.F. and Staden,R. (1995) A new DNA sequence assembly program. *Nucleic Acids Res.*, **23**, 4992–4999.
- Bonfield,J.K. and Staden,R. (1995) Experiment files and their application during large-scale sequencing projects. *DNA Sequence*, **6**, 109–117.
- Chen,F., MacDonald,C.C. and Wilusz,J. (1995) Cleavage site determinants in the mammalian polyadenylation signal. *Nucleic Acids Res.*, **23**, 2614–2620.
- Ewing,B.G., Hiller,L., Wendl,M.C. and Green,P. (1998) Basecalling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.*, **8**, 175–185.
- Gavin,A.J., Scheetz,T.E., Roberts,C.A., O'Leary,B., Braun,T.A., Sheffield,V.C., Soares,M.B., Robinson,J.P. and Casavant,T.L. (2002) Creation and identification of pooled library tissue tags for est-based gene discovery. *Bioinformatics*, **18**, 1162–1166.
- Scheetz,T.E. and Casavant,T.L. (2003) Informatics for efficient EST-based gene discovery in normalized and subtracted cDNA libraries. *Technical Report: TR-CLCG-030129*.
- Smith,T.M., Abajian,C. and Martin,C. (1997) HOPPER: software for automating data tracking and flow in DNA sequencing. *Comput. Appl. Biosci.*, **13**, 175–182.
- Smith,T.M. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
- Staden,R. (1996) The Staden sequence analysis package. *Mol. Biol.*, **5**, 233–241.
- Wendl,M.C., Dear,S., Hodgson,D. and Hiller,L. (1998) Automated sequence preprocessing in a large-scale sequencing environment. *Genome Res.*, **8**, 975–984.