2074–2083 *Nucleic Acids Research, 2007, Vol. 35, No. 6* *Published online 13 March 2007*
doi:10.1093/nar/gkm081

EST assembly supported by a draft genome sequence: an analysis of the *Chlamydomonas reinhardtii* transcriptome

Monica Jain[1], Jeff Shrager[1], Elizabeth H. Harris[2], Renee Halbrook[1], Arthur R. Grossman[1], Charles Hauser[2,4] and Olivier Vallon[1,3,*]

[1]The Carnegie Institution, Department of Plant Biology, 260 Panama Street, Stanford, CA 94305, USA, [2]Biology Department, Duke University, Durham, NC 27708, USA, [3]Institut de Biologie Physico-Chimique, UMR7141 CNRS/Université Pierre et Marie Curie-Paris6, 13 Rue Pierre et Marie Curie, 75005 Paris, France and [4]St. Edwards University, Department of Biology, Austin, TX 78704, USA

Received December 21, 2006; Accepted January 26, 2007

ABSTRACT

Clustering and assembly of expressed sequence tags (ESTs) constitute the basis for most genome-wide descriptions of a transcriptome. This approach is limited by the decline in sequence quality toward the end of each EST, impacting both sequence clustering and assembly. Here, we exploit the available draft genome sequence of the unicellular green alga *Chlamydomonas reinhardtii* to guide clustering and to correct errors in the ESTs. We have grouped all available EST and cDNA sequences into 12 063 ACEGs (assembly of contiguous ESTs based on genome) and generated 15 857 contigs of average length 934 nt. We predict that roughly 3000 of our contigs represent full-length transcripts. Compared to previous assemblies, ACEGs show extended contig length, increased accuracy and a reduction in redundancy. Because our assembly protocol also uses ESTs with no corresponding genomic sequences, it provides sequence information for genes interrupted by sequence gaps. Detailed analysis of randomly sampled ACEGs reveals several hundred putative cases of alternative splicing, many overlapping transcription units and new genes not identified by gene prediction algorithms. Our protocol, although developed for and tailored to the *C. reinhardtii* dataset, can be exploited by any eukaryotic genome project for which both a draft genome sequence and ESTs are available.

INTRODUCTION

With the development of massive DNA sequencing capacity and powerful assembly algorithms, determining sequences of eukaryotic genomes, once a daunting task, has now become commonplace (1,2). As of November 2006, the Genome Online Database lists 631 eukaryotic genome projects, of which 618 are incomplete (see http://www.genomesonline.org/). Using a shotgun genome sequencing strategy, it is possible to generate, in a matter of weeks, a draft genomic sequence that covers a large fraction of the genome and is distributed over a number of 'scaffolds' of various lengths (many more than there are chromosomes). In spite of its shortcomings, a draft genome sequence is adequate for many purposes, from the description of gene content to medium-range synteny analysis and genetic mapping. A more refined genome sequence, ideally with only a few unsequenced tracts of known length, can only be achieved through more dedicated efforts, involving expensive physical mapping and gap closure procedures. Unless technological breakthroughs simplify these arduous tasks, more and more eukaryotic genomes are likely to remain, for long periods, at an advanced draft stage.

Recently, the Joint Genome Institute has generated a draft genome sequence of the unicellular green alga *Chlamydomonas reinhardtii* (http://genome.jgi-psf.org/Chlre3/Chlre3.home.html). This model organism is being used to study numerous biological processes, in particular photosynthetic CO_2 fixation, and the structure and function of cilia and basal bodies (3). The nuclear genome of *C. reinhardtii* is ~120 Mb partitioned into 17 chromosomes. The latest release of the genome

*To whom correspondence should be addressed. Tel: +33 1 5841 5058; Fax: +33 1 5841 5022; Email: ovallon@ibpc.fr

(version 3.0) consists of 1557 scaffolds totaling 105 Mb of high-quality sequence, interspersed with 15 Mb of sequence gaps. The longest scaffold (scaffold_1) covers >2 Mb, and the 24 largest scaffolds make up 50% of the genome. Using homology-based and *ab initio* prediction programs, with 5′ and 3′ UTRs added (based on EST data), the genome has been populated by gene models of which 15 256 have been selected as best describing their respective loci. Among these, 2238 still contain one or more sequence gaps (A. Salamov, JGI, personal communication).

To enhance the *C. reinhardtii* gene catalog, we have sought to generate a set of experimentally verified transcript sequences by assembling the vast array of expressed sequence tags (ESTs) available for this organism. Because of the diversity of cDNA libraries used in these studies, this data is expected to sample a large fraction of the transcriptome. However, the high rate of sequence errors in ESTs limits the accuracy of such an assembly. In addition, the heterogeneity of the *C. reinhardtii* EST dataset represents a challenge for sequence assembly: while the Kazusa Institute (http://www.kazusa.or.jp/en/plant/chlamy/EST/) (4–6) has chosen the C9 strain, the Chlamydomonas Genome Project (CGP, http://www.chlamy.org/search.html) (7) has used mostly the strain 21gr, and to a lesser extent 137c (used in the genome sequencing project) and the highly polymorphic S1D2 strain used for molecular mapping. Both projects have assembled their data using the program suite CONSED/PHRED/PHRAP (8), but only the CGP project, because it used both 5′ and 3′ end reads, has the potential to generate full-length transcripts. Comparison of the last CGP assembly (termed 20021010) with the draft genome sequence shows a relatively high level of redundancy (multiple contigs mapping to the same genomic region) and of inaccuracies (differences between transcript and genome sequences). As the genome sequence has <1 error in 10 000 bp, inaccuracies can be considered as arising mostly from EST sequencing errors and to a lesser extent from inter-strain polymorphisms.

To overcome these limitations, we have developed an algorithm that makes use of the draft genomic sequence to correct errors and polymorphisms in the EST data. The first step of this procedure is to map ESTs onto the genome and generate a 'ghost' representing the template sequence. Ghosts are then grouped into 'ACEGs' (assembly of contiguous ESTs verified on genome), based on position and orientation on the genome and on paired-end sequence information. Finally, sequence assembly is performed within each ACEG to generate one or several contig(s).

METHODS

Data collection and pre-processing

Our procedure is summarized in Figure 1. The details about the computational aspect of this algorithm can be found in (21). The draft genome sequence was obtained from http://genome.jgi-psf.org/Chlre3/Chlre3.home.html. Most EST sequences were provided by the CGP, a joint effort of the Carnegie Institution and the Stanford Genome Technology Center. The quality-trimmed ESTs from the Kazusa project were downloaded from GenBank, and additional unpublished ESTs were provided by Saul Purton (University College, London). In addition, we retrieved all *C. reinhardtii* cDNA sequences present in the EMBL database on June 1, 2004, and obtained a few unpublished sequences from individual laboratories. Overall, we collected 246 972 EST and cDNA sequences (Table 1). Each sequence is designated by a name (clone or database entry) followed by a suffix, either .x1 (for 3′ end sequences) or .y1 (for 5′ end sequences). When a CGP clone was sequenced more than once from any end, .x2 .y2 etc were used as the suffix.

In the first step of the procedure, contaminating vector sequences and low-quality ESTs in the CGP dataset were

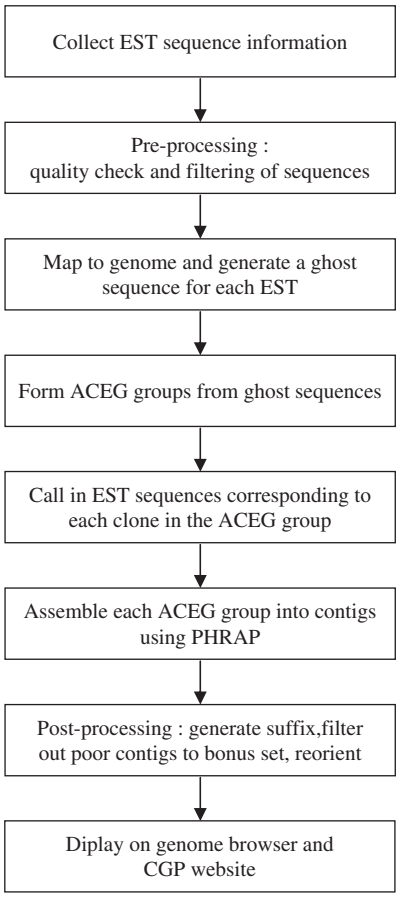

Figure 1. Overall algorithm for ACEG assembly.

Table 1. Source of sequences and remaining numbers after quality-screening and ghost generation

ESTs	Total (input)	After Lucy	After BLAT
CGP Libraries	194 920	145 686	114 809
Kazusa	50 961	50 961	48 044
Genbank	765	765	698
Purton	283	283	262
Private	43	43	42
Total	**246 972**	**197 738**	**163 855**

filtered using TIGR's sequence cleanup program Lucy (http://www.tigr.org/software/) (9). Only sequences containing a minimum of 75 nt with an average probability of error of no greater than 0.02, using a window size of 10, were included in the assembly. Using these quality criteria, 20% (49 234) EST clone reads were rejected from the assembly (see Table 1). Previous assemblies of the CGP data have suffered from EST misnaming (clones sequenced under a wrong ID). To correct this artifact, CGP ESTs were compared among themselves using BLAST to identify cases for which a high number of reads from one plate matched reads from the same well position but in another plate. Based on this analysis, reads from 51 plates were renamed to restore the correct paired-end information (listed on http://www.chlamy.org/search.html).

Mapping ESTs onto the genome and 'ghost' generation

Filtered ESTs were mapped to the draft genomic sequence using the program BLAT (10). The goal was to identify a unique, unambiguous genomic position for each EST and to extract the associated matrix genomic sequence. BLAT alignments are given a score according to Equation (1):

$$\text{Score} = 100 \times (\#\text{matches} - \#\text{mismatches} - \#\text{rep\_matches})/\text{EST\_length}$$

where #rep_matches is the number of positions covered by another match in the same or another scaffold.

An EST sequence was considered to not match the genome and was dropped from the assembly pipeline if:

(a) the EST mapped to different scaffolds with scores of ±10%, or
(b) a segment of the EST mapped to different locations of the same scaffold with coordinates that were offset by >500 nt and a score of ±10%.
(c) the EST spanned >6 kb of scaffold sequence.
(d) the sum of the BLAT HSP lengths was <0.75 the EST length.

Using these criteria an additional 14% of the ESTs were eliminated from the assembly pipeline.

For each EST that passed these criteria, the corresponding BLAT HSPs were used to generate a 'ghost' sequence which represents the sequence on the genome that corresponds to the mature transcript. When the BLAT alignment was comprised of more than one HSP, the nature of the gap between consecutive HSPs directed the generation of the ghost sequence. If the gap was ≤20 nt on both the cDNA (dC) and on the genome (dG), it was treated as a sequencing error or polymorphism in the cDNA. In this case, the genome sequence defined by the start of the first HSP and the end of the second HSP (including the gap sequence) was used to generate the ghost sequence. In cases where $dC = 0$ and $dG \geq 20$, the mismatch was assumed to come from an intron, and the genomic sequence between the HSPs was not included in the 'ghost' sequence. When dG was >20 but dC was not equal to 0, this was interpreted to indicate either a sequencing error straddling a junction between two exons or a short exon that was not identified in the BLAT analysis. In this case we considered that the missing sequence, although uncertain, was of the length described by the EST, and we introduced in the ghost sequence a number of unidentified bases (N) equal to dC. The ghost genomic sequence was assigned the clone ID of the corresponding EST sequence preceded by a 'g'. The genomic position of the ghost is stored in its defline.

We tried to eliminate sequences originating from contamination of cDNA libraries by genomic DNA. Here, 351 contaminating clones were identified in the CGP database based on two criteria: (1) absence of introns (all ghosts from these clones consisted of a contiguous stretch of sequence), and (2) presence on the genomic sequence of a XhoI restriction site (CTCGAG) within 20 nt downstream of their 3′ end (XhoI is the enzyme used for cloning into Lambda-ZAP). Note that our procedure relies on analysis of 3′ ends, and will therefore fail to identify genomic contaminants in the Kazusa ESTs.

Forming ACEG clusters

Ghosts were grouped into ACEG clusters using their 'genomic position and orientation' (derived from the BLAT mapping) and 'clone name' (used to pair the 5′ and 3′ ends of the same clone). We started by randomly selecting a 5′ ghost and assigning it to an ACEG cluster. All overlapping 5′ ghosts were added to the cluster (stage 1, Figure 2), along with the 3′ ghosts originating from the same clones, unless they were separated from the corresponding 5′ ghost by >20 000 bp (stage 2, Figure 2). The clustering algorithm is recursive: for each new 3′ ghost, overlapping 3′ ghosts and corresponding 5′ ghosts were added to the cluster (stage 3 in Figure 2). Again, for each new 5′ ghost, overlapping 5′ ghosts and corresponding 3′ ghosts were also added to the cluster. The cluster continued to grow like this until no more additional ghosts entered the cluster (stage 4 in Figure 2). Once a cluster was complete, all ghosts belonging to that cluster were removed from the dataset and the process was repeated until all 5′ ghost had been assigned to a cluster. ACEG coordinates were defined by the outermost nucleotide positions of its ghosts.

At this stage, a gene could still be represented by several ACEGs, because some cDNAs in the libraries are cloned in reverse orientation, others are truncated at both the 5′ and 3′ ends, etc. All ACEGs that overlapped by >50% of the length of the shorter ACEG, regardless of their orientation, were fused into a single ACEG.

Assembling ACEGs into contigs

Each ACEG group was then assembled into one or more sequence contigs using the PHRED/PHRAP suite (stage 5 in Figure 2). In order to capture all available sequence information, we used the ghosts plus all the ESTs from the corresponding clones, even those that did not generate a ghost. For CGP ESTs, we used sequence quality values derived automatically from the chromatograms (.phd files). For ESTs from the Kazusa and Purton projects, this data was not available and all positions were given the arbitrary quality value of 30, corresponding to an expected error rate of 1 every 1000 nt. The ghost and EMBL sequences were

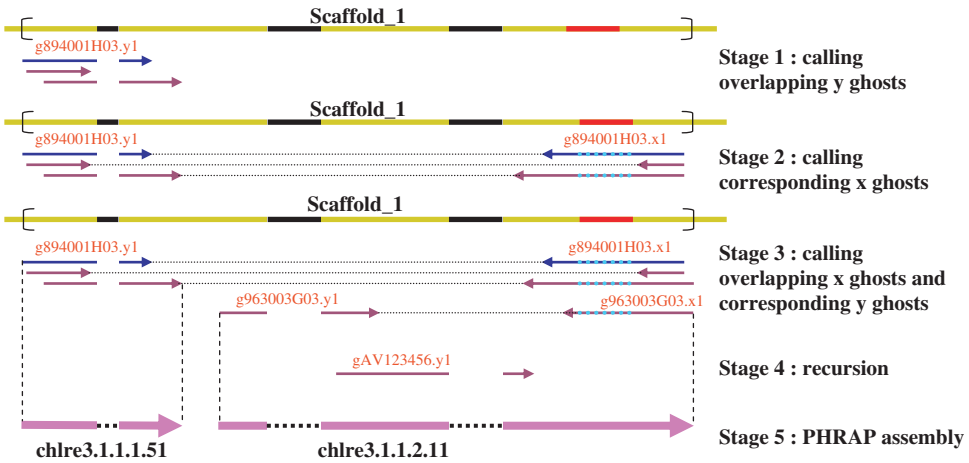

Figure 2. Different stages of ACEG generation. An example is given of a hypothetical gene (bracketted on the scaffold line) split by three introns (black bars) and with two possible polyadenylation sites. Its last exon is interrupted by a sequence gap (red) that leads to stretches of N in some of the ghosts (dotted cyan lines). Thin arrows indicate ghost position and orientation, dotted black lines group paired ghosts from the same clone. Assembly starts with ghost g894001H03.y1 and generates two non-overlapping contigs (purple arrows). Because ESTs are introduced at stage 5, chlre3.1.1.2.11contains the sequence missing in the genome sequence gap.

given the quality value of 35 because of their intrinsic higher accuracy. The PHRAP parameters (http://www.phrap.org/phredphrap/phrap.html) used to control the stringency and completeness of the assembly process were Forcelevel 5; Retain-duplicates ON; Gap-extension penalty −2; Revise_greedy ON. For a small number of ACEGs that failed to assemble after 10 min, PHRAP was re-run with Revise_greedy OFF and Forcelevel 7. For 34 ACEGs with a single ghost, PHRAP generated no contig, and the ghost sequence was taken as the contig.

Post-processing

Once contigs were generated, remaining vector and adaptor sequences were removed using cross-match and text search, respectively. The direction of the contigs with respect to that of transcription was determined based on whether 5′- and 3′-sequences were used by PHRAP in the direct, or reverse, orientation. When necessary, the contig sequences were reverse complemented so that all contigs eventually read in the 5′ to 3′ orientation with respect to the transcript. Each contig was assigned a four part ID preceded by the common prefix 'chlre3'. The ID starts with the scaffold number from which the ACEG was formed, followed by the ACEG number and contig number within that ACEG. Within an ACEG, the number assigned to a contig increases with the number of sequences it uses. To describe the nature of the sequences used, a suffix was added at the end of the ID. In a single contig ACEG, the contig may group both 5′ and 3′ reads (suffix .1), only 3′ reads (suffix .3), or only 5′ reads (suffix .5). When the ACEG contained more than one contig, two-digit suffixes were used. The first digit describes the type of the contig, as above, while the second digit ranks contigs within this type. For example, contig chlre3.12.34.1.11 is the first contig within ACEG #34 of scaffold_12, the one that has the fewest reads; its suffix indicates that it contains sequences from both 5′ and 3′ ends and that other contigs were generated for that ACEG. Finally, contigs for which only EST reads (no ghosts) were used received a suffix starting with 0.9. Because of their intrinsic low quality, these contigs were moved to a separate bonus set, together with other suspected artifactual sequences (see Results section).

RESULTS

ACEG generation

Our starting EST/cDNA dataset comprised a total of 246 972 sequences (Table 1), of which 197 738 (80%) were deemed of sufficient quality to be included in the analysis. 33 883 (~14%) could not be unambiguously mapped onto the genome, either because the length of the match was too short, or because it matched at multiple positions on the genome. Overall, 163 855 sequences were mapped onto the genome and converted into 'ghosts' as described in the Methods section. Note that for 1568 (0.6%) clones, the .y (5′) and .x (3′) ghosts mapped to different scaffolds. This could occur because the gene is split between two scaffolds, or because of errors in mapping the ESTs. For the CGP ESTs, the length of ghosts (651 bp on average) was larger than that of the 'high-quality region' of the chromatograms (526 bp), which stresses the advantage of using a reference genome rather than arbitrary quality criteria when trying to limit the effect of sequence errors.

A ghost is a concatenation of the genomic BLAT HSPs that correlate to an EST. Positions of discrepancy between the two sequences likely represent sequence errors or inter-strain polymorphisms; our procedure systematically uses the genomic sequence for those positions. When a complete alignment is impossible, missing sequence is replaced by the appropriate number of unidentified bases. This happens when there are very short exons, when the quality of the EST data is too poor to allow alignment, or when the EST straddles a sequence gap on the genome. Such tracts of undetermined sequence were found in 9.1% of the ghosts.

Based on position and orientation on the genome, ghosts were grouped into 12 063 ACEGs, each associated with genome coordinates derived from the outermost positions of its ghosts. Within each ACEG, ghost and EST sequences were assembled using the PHRAP program suite. The reason for reintroducing the ESTs at this stage was that we wanted to capture the cDNA information lost at the stage of ghost generation because of sequence gaps on the genome. This however could impact negatively on the accuracy of the contig sequences. To counteract this effect, we gave the ghost an arbitrary quality value of 35, which is higher than that of the ESTs, so that sequence errors and polymorphisms within EST sequences would not be incorporated into the contigs. Also, PHRAP parameters were fine-tuned based on sample assemblies with ACEGs from scaffold_1; this process helped limit redundancy among contigs while still capturing some information on potential splice variants. During the final processing, the contigs were reoriented to read in the 5′ to 3′ direction and given a suffix indicating the type of sequences that were used in their generation (as discussed in the Methods section). The ACEGs were then reviewed for several types of possible artifacts. There were 3080 contigs that used exclusively EST information, to the exclusion of ghosts (suffix .9 or .9n). They were placed in a separate 'bonus' set, together with a few contigs showing poly-A stretches at both ends (putative chimeric cDNAs) or that were suspected to represent genomic contamination. Finally, 1817 contigs of a separate set of ACEGs, generated from 3′ ghosts that had not been used by the main ACEG generation algorithm because they had no corresponding 5′ ghost, were also placed into the bonus file. The full bonus set of 4931 contigs is offered as a secondary source of information, as it is largely redundant with the main set. It is not analyzed in this manuscript.

The main results consist of 15 857 contigs contained within 12 063 ACEGs, an average 1.3 contigs per ACEG. By comparison, the previous assembly, using half as many ESTs, generated 8628 ACEs (assembly of contiguous ESTs; no genomic data was used to improve the quality of the assembled sequence) and 14 410 contigs (average 1.7 contigs per ACE). The majority of our ACEGs (59%) consist of a single contig, and only 13 contain more than four contigs (Figure 3). As expected, ACEGs with more contigs have, on average, more reads associated with them. The 5′ and 3′ read composition of contigs is shown in Table 2. The prevalence of ACEGs with a .5 over a .3 suffix is explained by the fact that the Kazusa library contains only 5′ reads.

We examined, in detail, 35 randomly chosen contigs (among 5874) combining 5′ and 3′ reads (Supplementary Table 1, sheet 1). Seventeen of the contigs were found to represent full-length cDNAs of *bona fide* protein-coding genes. The others were incomplete or corresponded to transposons or other types of non-coding sequences. Therefore, we conclude that our assembly describes full-length transcripts for roughly 3000 *C. reinhardtii* genes.

The mean vector-trimmed length of our contigs is 934 nt, and the median is 750 nt. The distribution (Figure 4) is clearly bimodal, with the first peak dominated by .5 and .3 contigs, and the second peak by .1 contigs (see contig length in Table 2). Most of the .5 and .3 contigs have only one or two reads (average 1.8, compared to 15.7 for .1 contigs). Interestingly, Kazusa and CGP ESTs are not randomly distributed among ACEGs (data not shown) and the two libraries thus nicely complement each other. Overall, the Kazusa library appears to contain a lesser proportion of 5′-truncated ESTs.

Most of the longest contigs (insets of Figure 4, maximum 6005 nt) are comprised of both 5′ and 3′ reads. Many of these long contigs correspond to known cDNAs from EMBL that were included in the starting dataset. Among

Table 2. Repartition of contigs between various categories, based on suffix type. The median contig length is indicated for each category

Contig composition	In ACEGs with a single contig (median length)	In ACEGs with several contigs (median length)
Both 5′ and 3′ reads	2894 (1338 nt)	2980 (1372 nt)
Only 5′ reads	4195 (487 nt)	4010 (692 nt)
Only 3′ reads	1 (715 nt)	1777 (754 nt)
Total	**7090 (663 nt)**	**8767 (750 nt)**

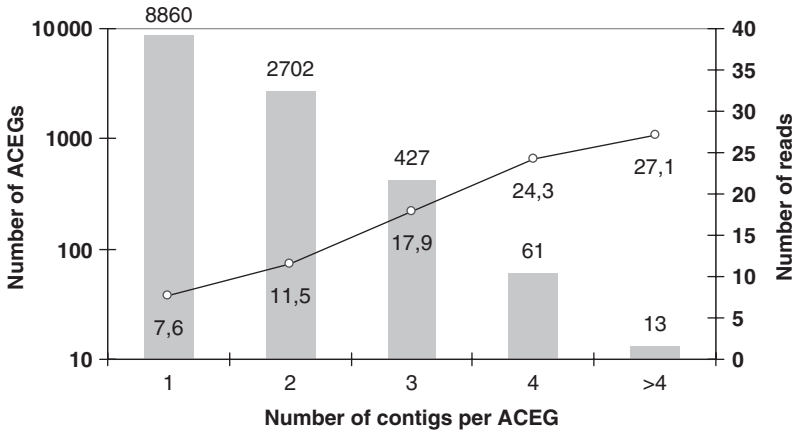

Figure 3. Number of ACEGs (bars) and average number of reads (open circles) as a function of number of contigs in the ACEG.

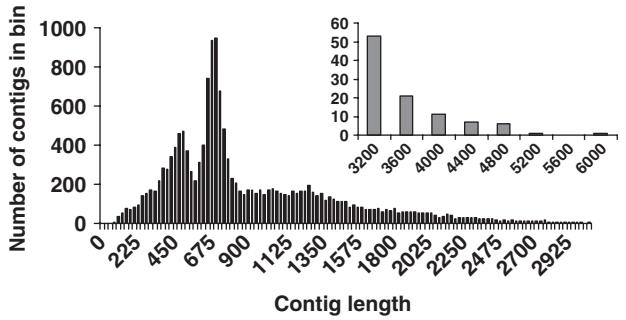

Figure 4. Distribution of contig lengths. Data has been placed into bins of 25 units in width. The inset is an enlarged display (with 400 units in width) of the longest contigs.

the interesting exceptions, chlre3.37.30.7.11 (4955 nt) arose from an accumulation of ESTs in an expressed pseudogene (estExt_fgenesh2_pg.C_370117). The longest protein-coding gene identified exclusively by EST information was the *DEH1* DEAD-box helicase (4569 nt).

Sequence redundancy among ACEG contigs

To estimate the level of redundancy in our ACEG assembly, we analyzed the results of a BLAST search using the main contig set both as query and subject, with a cutoff of $E = 10^{-30}$. We found that 10 596 contigs (67% of total) were unique, i.e. did not hit any other contig. This is a marked improvement compared to the previous assembly. About half of these non-redundant contigs were the sole contig within their ACEG, while the others had sister contigs but there was no overlap among the contigs. For 5261 contigs, BLAST revealed similarity to other contigs in the assembly. Two categories of redundancy are considered: internal redundancy (matches to a contig of the same ACEG) and external redundancy (matches to a contig of another ACEG). For each category and subcategory (see below), a set of contig matches was chosen at random and the ACEGs associated with these contigs were analyzed in detail, comparing the genome information available on the JGI browser and the EST information. Our aim was to determine whether this redundancy was a real property of the gene, or was due to an artifact in our dataset or assembly procedure.

ACEGs with internal redundancy are presented in Supplementary Table 1 (sheets 2–5) and its accompanying description. In addition to a couple of cases of direct or inverted repeats within the contig, we observed 997 cases of overlap between contigs in the same ACEG. In search for potential cases of alternative splicing, we focused the analysis on the 531 ACEGs where the alignment of the two contigs was discontinuous (i.e. showed several HSPs) and was in the plus orientation (the two contigs read on the same strand of the genome). Somewhat arbitrarily, we tried to distinguish between: (1) biologically meaningful alternative splicing or alternative sites of transcription initiation, giving rise to multiple transcripts with reasonable coding capacity, and (2) 'mis-splicing', i.e. improper processing of the pre-mRNA, where one of the cDNAs contains premature stop codons or appears otherwise non-functional. In 30 ACEGs examined (Supplementary Table 1, sheet 2), we found 8 occurrences of mis-splicing, 11 of alternative splicing and 2 of alternative transcription start sites. Extrapolating to the 531 ACEGs in that set, we estimate that our dataset will contain ~230 cases where multiple transcripts are produced from a single gene (Supplementary Table 1, sheet 5).

In addition to alternative splicing, this analysis of internal redundancy also revealed cases of inter-strain polymorphism (in the region of the transcript where genome sequence was not available), artifacts occurring during cDNA cloning (contamination by genomic DNA, cDNA cloned in reverse orientation) or shortcomings of the assembly procedure. Interestingly, our internal redundancy analysis sampled three cases in which a minor cDNA was associated with the opposite strand of a well-expressed gene (an ABC transporter, the methionine adenosyl-transferase gene *METM* and a PBF-2-like transcription factor). The antisense cDNAs showed canonical intron splicing and sometimes a poly-(A) tail, but did not seem to code for a protein. Whether these have a regulatory function or are the result of spurious transcription of processed pseudogenes remains to be determined.

External redundancy, i.e. match between contigs of different ACEGs, was analyzed separately (Supplementary Table 1, sheets 6 and 7). This category arose mostly from similarity among transposon sequences, simple nucleotide repeats and gene families. Thus, 2490 contigs had hits to ACEGs on other scaffolds, with up to several hundred HSPs for some transposons. These were not examined further. However, we examined in greater detail those cases where the ACEG hit was on the same scaffold. In particular, we looked for cases where a single gene would be erroneously described by two overlapping ACEGs (a common artifact in previous *C. reinhardtii* EST assemblies). Of 10 randomly chosen cases, this was never encountered. However, in addition to two cases of nearby transposons, we found five cases of closely related genes located in the same scaffold (but not overlapping). This is in line with the large number of local gene duplications found in the *C. reinhardtii* genome (13%, Simon Prochnik, JGI, personal communication). The three remaining cases represented instances of overlapping genes: one divergent pair, where the 5′ ends of the transcripts overlapped, and two convergent pairs with overlapping 3′ ends. As gene overlap can be functionally significant, we tried to estimate the frequency of this type of configuration. Among 252 pairs of ACEGS whose contigs showed HSPs in the minus orientation and located near the end of both contigs, we examined 10 randomly chosen examples and indeed found six clear cases of overlapping, converging genes (Supplementary Table 1, sheet 6, section IIc). By extrapolation, our entire dataset must contain roughly 150 such configurations. If the two genes are expressed at the same time and the overlap is long enough to lead to the formation of double-stranded RNA, this could affect transcript accumulation from both genes. Note, however, that the overlapping transcript often was one of several possible transcripts for that gene, usually the one represented by the least number

of ESTs, so that interference would be only partial. We also examined genes whose 5′ ends overlapped in a divergent orientation. Ten examples were examined of the 35 ACEG pairs identified by BLAST, and six cases of overlap were identified (section IId). Thus, overlapping occurred less often in the divergent than in the convergent configuration. Again, in most cases we found that one of the genes had two possible transcription start sites, only one of which lead to transcript overlap. Finally, we identified genes overlapping in a parallel orientation, i.e. with HSPs in the plus orientation and spanning the beginning of one sequence and the end of the other. In the 10 examples examined, out of 165 pairs, we observed 5 cases of tandem overlap (section IIb). We did not systematically analyze the 1054 remaining ACEG pairs, where the HSPs were not near the extremities of the contigs (250 in minus orientation, 804 in plus orientation), but cursory inspection indicates that they also contain a number cases of overlapping genes, including some with the an exon aligning in an intron of the other gene. We conclude that overlapping transcription units are widespread in the *C. reinhardtii* genome, with at least several hundred cases represented in our dataset alone, and that there is a bias towards the overlap of 3′ ends. Note that this analysis is based on BLAST and requires that the overlap be sufficiently large to give a significant match. When only the ACEG coordinates (derived from those of the ghost sequences) are considered, the number of overlapping pairs rises to 724 (Supplementary Table 1, sheet 8). The median length of the ACEG overlap was 109 nt and in 610 cases the overlap was larger than 22 nt, the minimum length for an RNAi effect. Only 411 of these showed blast hits between their contigs, but the others could be significant as well.

As mentioned above, local gene duplications are numerous on the *C. reinhardtii* genome, where a large proportion of genes show high similarity (at the protein sequence level) to a gene located nearby on the genome. In search of cases where conservation of closely positioned sequences was highest, we concentrated on the 22 cases for which the aligned region corresponded to the end of the contigs, hence possibly the 3′ UTRs (Supplementary Table 1, sheet 7). Not counting transposons, we found 10 examples of strong sequence conservation between related and closely linked genes (or pseudogenes). Interestingly, we also found two cases where the 3′ UTRs showed sequence similarity, but not the CDS. This is probably a consequence of duplication of a part and not an entire gene. There also was a case in which the 5′ and 3′ UTRs of a gene were highly similar, again because of a short local duplication.

Comparisons of ACEGs to gene models

Our main goal in generating ACEGs was to complement the description of *C. reinhardtii* genes provided by the draft genome sequence. The JGI genome annotation pipeline involves the prediction of gene models via a series of *ab initio* or homology-based methods, followed by a choice of most likely models based on a scoring algorithm. Although some of the models were based on ACEG contigs and raw EST data was used to extend some of the gene models at their 5′ and 3′ ends, we found that the preferred model did not always conform to the EST data. We compared all ACEG contigs (15 857) to the set of 'Filtered Gene Models 2' (15 256 transcripts) using BLASTN with $E = 10^{-15}$. An interactive database was generated and has been made accessible online (http://ren.stanford.edu/AcegTool/AcegTool.html); it displays the results either on a per contig or per gene model basis and can also be accessed from the JGI protein pages. Surprisingly, we found that 7109 (59%) ACEGs do not match a gene model, when the requirement for identity was at least 98% and for coverage at least 90% of contig length (Figure 5). Even when the requirement for coverage was lowered to 20%, 3461 ACEGs had no match in the filtered models. Out of 20 randomly chosen ACEGs in that category (Supplementary Table 1, sheet 9), two were found to show a perfect match to a gene model that was not selected in the 'Filtered models' set, and twelve could be considered as extensions of existing gene models (usually 5′ or 3′ extensions). Three more showed reasonable coding capacity and probably represented protein-coding genes that had been completely overlooked by the gene prediction algorithms. We also found two cases of transposons and one example of a non-coding

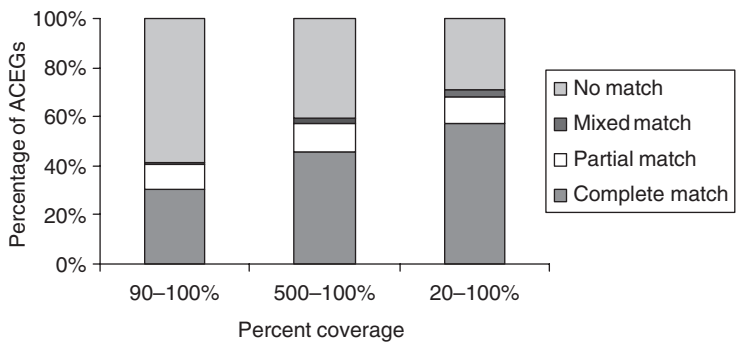

Figure 5. Comparison of ACEGs and gene models. ACEGs show either a complete match to a single gene model (for all contigs, coverage is above cutoff and identity at least 98%), a partial match (some contigs match the model, but others match nothing), a mixed match to several gene models (some contigs match one model, others match another model), or no match at all. Results are displayed for three minimum coverage levels on the ACEG contigs.

RNA with a poly-(A) tail. We conclude that ACEGs could be more thoroughly exploited to create new gene models and to extend existing gene models in the 5′ and 3′ directions. They could also be used to guide the filtering algorithm that chooses the most likely model at a particular locus.

ACEGs bridge sequence gaps in the genome

In order to allow the use of cDNA information spanning sequence gaps in the genome, our algorithm reintroduces the ESTs into the dataset used for PHRAP assembly. We have attempted to determine whether or not this resulted in a gain of information. By mapping ACEGs back to the genome, we found that 649 contigs, representing 568 ACEGs in 113 scaffolds, had BLAT HSPs on both sides of a sequence gap (Supplementary Table 1, sheet 10). For 10 out of 20 randomly selected genes in this set, it was found that the genome sequence has a sequence gap in the transcript that was fully covered by the ACEG contig. The gap ranged from a couple of nucleotides to an entire exon. Extrapolating this data suggests that the ACEG information can fill in gaps for ~280 *C. reinhardtii* genes (out of the 2238 that have gaps). In addition, 185 ACEG contigs mapped within 10 nt of a sequence gap, and in a fraction of these, the contig was found to read into the gap.

In addition to bridging gaps within a scaffold, an assembly procedure using EST paired-end information has the potential to bridge genome scaffolds together, if an expressed gene is split between two scaffolds. We examined 20 of the 55 cases where an ACEG contig has BLAT HSPs on two different scaffolds, with <10% overlap between the hits. Five cases were found where a small scaffold (# 154, 374, 916, 1574 and 2517) could be entirely or partially placed within a gap of a larger scaffold (resp. # 46, 30, 41, 4 and 2).

DISCUSSION

Identifying protein-coding genes in a genome sequence is a daunting task, yet it is a crucial step in making the sequence a useful tool for addressing biological questions concerning gene function. Researchers have used an array of complementary approaches to identify protein-encoding genes, one of which is the systematic sequencing of cDNA clones. Sequencing can be carried out either on both strands of carefully selected cDNAs, with the aim of establishing a complete and reliable cDNA sequence, or on randomly selected clones from various libraries, in which case only end-sequences are collected. This EST approach is easily automated and generates a large number of sequences that can be assembled, using programs like PHRAP, into a smaller number of sequence contigs (11–13). The main drawback of EST assembly is that it does not usually permit the determination of a complete cDNA sequence, because most genes are too large to be covered by end-sequencing. In addition, sequence quality drops towards the end of the sequence reads, which can prevent assembly programs from joining overlapping sequences into a single contig. Even when full-length contigs are generated, they are likely to contain errors, especially in regions where only low-quality data are available.

The availability of genome sequence information can significantly improve the assembly of ESTs. In this article, we have used the draft genome information generated by JGI in two ways. First, we used it to correct sequence errors in our EST collection. 'Error' here means not only sequencing artifacts (undetermined bases, base changes, indels) which for a gene will vary from one EST to another, but also inter-strain polymorphisms that will be present in all ESTs derived from the same strain. Our reference genome sequence is from a strain of the '137c' lineage, while most ESTs derive from the 21gr strain or from the Japanese C9 strain, or even from the highly polymorphic S1D2. The ghost sequences that we have generated are based on the genomic sequence, and the artificial quality value they received was high enough to warrant that the 137c sequence would prevail at the assembly step, even when many high-quality polymorphic ESTs also span that part of the gene. For example, the non-coding chlre3.38.1.4.51 is comprised exclusively of reads from S1D2, with 30–50 SNPs and 7–10 indels, yet its sequence is identical to that found in the 137c strain. Occasionally, a contig from the main set will show many differences from the genome sequence: this has occurred when a S1D2 EST was used for a fraction of its length, but the corresponding ghost sequence did not align well enough and was placed in a different contig (e.g. chlre3.8.131.2.11).

Our second usage of the genome is in the grouping of ESTs. Mapping ESTs to the genome allows for their accurate assignment to a particular gene, much better than a simple sequence comparison would. As a result, highly conserved gene families are better described by ACEGs than they are by the ACEs of the previous assembly (7). For example, of the nine *LHCBM* genes (coding for light-harvesting proteins of photosystem II), only *LHCBM2* and *LHCBM5* were described by a single ACE. All other *LHCBM* genes were matched by contigs from several ACEs, up to five contigs from three different ACEs in the case of *LHCBM3*. Moreover, the closely related and linked *LHCBM4*, *LHCBM6*, *LHCBM8* and *LHCBM9* even had similar matching scores to the same series of contigs from ACEs 20021010.829 and 20021010.1714. This problem is solved by our genome-based assembly protocol; none of the *LHCBM* genes are represented by more than a single ACEG. This can be traced to rejection, at the ghost generation stage, of ambiguous ESTs that match to several locations equally well. A drawback of this stringency is that it can prevent generation of an ACEG, even for a highly expressed gene. For example, *LHCBM3*, in spite of its hundreds of ESTs, was only partially covered by our assembly.

Because *C. reinhardtii* genes often overlap in antiparallel orientation, our grouping procedure had to make combined use of the position, strand and clone name information, to avoid grouping together overlapping ESTs coming from different genes. This was largely successful, as exemplified by the fact that our extensive random sampling did not identify a single case of an

ACEG describing two nearby genes and only one where two overlapping ACEGs described the same gene (see Supplementary Table 1, sheet 6). This is not to say that every ACEG represents a unique gene: especially for large genes, it is not uncommon to see different, non-overlapping ACEGs describing different regions of the transcript. In addition, a substantial fraction of the ACEGs correspond to transposons, genomic contamination in the EST libraries, or non-coding RNAs (see Supplementary Table 1). Overall, we estimate that roughly half of the ~15 000 protein-coding genes in the *C. reinhardtii* genome are at least partially described by our assembly.

Another source of improvement in our protocol was the systematic search and curing of plate numbering errors at the sequencing stage. This, in addition to the gene family issue discussed above and the difficulty in matching error-riddled ESTs, had undoubtedly contributed to redundancy in previous *C. reinhardtii* EST assemblies, including the preliminary ACEG assembly that we generated using version 2.0 of the genome and that was used for generating the oligonucleotide array (see Figure S1 in (14)). Our attempt to eliminate genomic contaminants also was helpful. Our screening procedure (identification of a XhoI site upstream of the 3′ read, requirement for a single HSP in all the ghosts of that clone) is rather demanding, and our sampling has found numerous genomic contaminants that had escaped our screening. Still, we believe that this type of screening should be systematically implemented in EST assembly protocols that make use of restriction at the cloning stage.

Our analysis of the main contig set sheds light on some intriguing aspects of the *C. reinhardtii* transcriptome. For example, numerous examples were found of alternative transcription starts, sometimes causing overlap with upstream genes. Potential alternative splicing has also been identified in a dozen cases, and analysis of the ACEGs listed in Supplementary Table 1, sheet 5, will reveal several hundred more. This of course is an underestimation of the real extent of these phenomena since EST coverage is only limited and our criteria for screening BLAST hits were stringent. Of the five cases of alternative splicing documented in the literature based on cDNA data (15–19), only two are revealed as distinct ACEG contigs.

ACEGs are also an invaluable resource for describing the 5′ and 3′ UTRs of genes and deciding whether the predicted gene models should be extended. In this respect, they are more useful than raw EST data. For example, ACEG coordinates can be used to attribute an ACEG contig to a gene, even if it does not overlap with an existing gene model, simply because the existence of a correlated EST at the other end of the gene precisely sets the gene boundaries (see for example, gene model gwH.2.337.1). Also, since ACEG generation integrates read orientation, their use will not force incorporation of ESTs belonging to another gene overlapping in the reverse orientation, as is often encountered in 'EST-extended' genewise or fgenesh models (see for example, the 3′ end of gene model estExt_fgenesh2_pg.C_20044). A better description of UTRs is necessary, especially if we want to understand the full meaning of the extensive gene overlap revealed by our analysis. Gene overlap, in particular in the antiparallel orientation, offers the possibility of regulatory mechanisms that can now be explored at the genome level. Transcriptional interference is increasingly recognized as an important feature of eukaryotic genomes (20).

Another area for which ACEGs can provide invaluable information is in the identification of non-coding RNAs. Besides ribosomal DNA, our ACEG assembly was found to contain putative polyadenylated precursors for several spliceosomal snRNAs (not shown), as well as a possible micro-RNA precursor (Supplementary Table 1, sheet 2). A systematic analysis of ACEGs that are not associated with gene models might uncover other types of non-coding RNAs.

Our study offers for the first time a comprehensive view of the *C. reinhardtii* transcriptome and its structural peculiarities. In addition, it provides a paradigm of general applicability for the genome-aided assembly of EST data, which should be easily applicable to any eukaryotic genome project for which both a draft genome sequence and a comprehensive EST dataset are available.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We are indebted to the US Department of Energy Joint Genome Institute, http://www.jgi.doe.gov/, for supplying the genome sequence data. We thank Saul Purton and various members of the Chlamydomonas community for supplying unpublished sequences, and Stephan Eberhard for critical reading of the manuscript. This work was supported by the CNRS (UMR7141) and by the NSF Chlamydomonas Genome Project, MCB 0235878 awarded to A.R.G. Funding to pay the Open Access publication charges for this article was provided by the Center National de la Recherche Scientifique.

Conflict of interest statement. None declared.

REFERENCES

1. Windsor,A.J. and Mitchell-Olds,T. (2006) Comparative genomics as a tool for gene discovery. *Curr. Opin. Biotechnol.*, **17**, 161–167.
2. Liolios,K., Tavernarakis,N., Hugenholtz,P. and Kyrpides,N.C. (2006) The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res.*, **34**, D332–D334.
3. Grossman,A.R., Harris,E.E., Hauser,C., Lefebvre,P.A., Martinez,D., Rokhsar,D., Shrager,J., Silflow,C.D., Stern,D. *et al.* (2003) Chlamydomonas reinhardtii at the crossroads of genomics. *Eukaryot. Cell*, **2**, 1137–1150.
4. Asamizu,E., Nakamura,Y., Sato,S., Fukuzawa,H. and Tabata,S. (1999) A large scale structural analysis of cDNAs in a unicellular green alga, Chlamydomonas reinhardtii. I. Generation of 3433 non-redundant expressed sequence tags. *DNA Res.*, **6**, 369–373.
5. Asamizu,E., Miura,K., Kucho,K., Inoue,Y., Fukuzawa,H., Ohyama,K., Nakamura,Y. and Tabata,S. (2000) Generation of

expressed sequence tags from low-CO_2 and high-CO_2 adapted cells of Chlamydomonas reinhardtii. *DNA Res.*, **7**, 305–307.
6. Asamizu,E., Nakamura,Y., Miura,K., Fukuzawa,H., Fujiwara,S., Hirono,M., Iwamoto,K., Matsuda,Y., Minagawa,J. *et al.* (2004) Establishment of publicly available cDNA material and information resource of Chlamydomonas reinhardtii (Chlorophyta), to facilitate gene function analysis. *Phycologia*, **43**, 722–726.
7. Shrager,J., Hauser,C., Chang,C.W., Harris,E.H., Davies,J., McDermott,J., Tamse,R., Zhang,Z. and Grossman,A.R. (2003) Chlamydomonas reinhardtii genome project. A guide to the generation and use of the cDNA information. *Plant Physiol.*, **131**, 401–408.
8. Gordon,D., Abajian,C. and Green,P. (1998) Consed: a graphical tool for sequence finishing. *Genome Res.*, **8**, 195–202.
9. Chou,H.H. and Holmes,M.H. (2001) DNA sequence quality trimming and vector removal. *Bioinformatics*, **17**, 1093–1104.
10. Kent,W.J. (2002) BLAT – the BLAST-like alignment tool. Genome Res., **12**, 656–664.
11. Zhang,L.D., Yuan,D.J., Zhang,J.W., Wang,S.P. and Zhang,Q.F. (2003) A new method for EST clustering. *Yi Chuan Xue Bao*, **30**, 147–153.
12. Kalyanaraman,A., Aluru,S., Kothari,S. and Brendel,V. (2003) Efficient clustering of large EST data sets on parallel computers. *Nucleic Acids Res.*, **31**, 2963–2974.
13. Udall,J.A., Swanson,J.M., Haller,K., Rapp,R.A., Sparks,M.E., Hatfield,J., Yu,Y., Wu,Y., Dowd,C. *et al.* (2006) A global assembly of cotton ESTs. *Genome Res.*, **16**, 441–450.
14. Eberhard,S., Jain,M., Im,C.S., Pollock,S., Shrager,J., Lin,Y., Peek,A.S. and Grossman,A.R. (2006) Generation of an oligonucleotide array for analysis of gene expression in Chlamydomonas reinhardtii. *Curr. Genet.*, **49**, 106–124.
15. Fukuzawa,H., Miura,K., Ishizaki,K., Kucho,K.I., Saito,T., Kohinata,T. and Ohyama,K. (2001) Ccm1, a regulatory gene controlling the induction of a carbon-concentrating mechanism in Chlamydomonas reinhardtii by sensing CO_2 availability. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 5347–5352.
16. Falciatore,A., Merendino,L., Barneche,F., Ceol,M., Meskauskiene,R., Apel,K. and Rochaix,J.D. (2005) The FLP proteins act as regulators of chlorophyll synthesis in response to light and plastid signals in Chlamydomonas. *Genes Dev.*, **19**, 176–187.
17. Beligni,M.V., Yamaguchi,K. and Mayfield,S.P. (2004) Chloroplast elongation factor ts pro-protein is an evolutionarily conserved fusion with the s1 domain-containing plastid-specific ribosomal protein-7. *Plant Cell*, **16**, 3357–3369.
18. Li,J.B., Lin,S., Jia,H., Wu,H., Roe,B.A., Kulp,D., Stormo,G.D. and Dutcher,S.K. (2003) Analysis of Chlamydomonas reinhardtii genome structure using large-scale sequencing of regions on linkage groups I and III. *J. Eukaryot. Microbiol.*, **50**, 145–155.
19. Schroda,M., Vallon,O., Whitelegge,J.P., Beck,C.F. and Wollman,F.A. (2001) The chloroplastic GrpE homolog of Chlamydomonas: two isoforms generated by differential splicing. *Plant Cell*, **13**, 2823–2839.
20. Shearwin,K.E., Callen,B.P. and Egan,J.B. (2005) Transcriptional interference – a crash course. *Trends. Genet.*, **21**, 339–345.
21. Jain,M., Holz,H., Shrager,J., Vallon,O., Hauser,C. and Grossman,A.R. *18th International Conference on Pattern Recognition. (ICPR'06)*. IEEE.