EPICENTRE® Biotechnologies

Analysis of 5′ transcript heterogeneity by high-throughput sequencing of cDNA

We isolated mRNA from *Blumeria graminis* and prepared a cDNA library using a modification of Epicentre's ExactSTART™ Full-Length cDNA Library Cloning kit. Analysis of the 5′ ends of the cDNA by 454 pyrosequencing yielded approximately 250,000 expressed sequence tags (ESTs) from one run of sequencing. The data also showed marked heterogeneity of the 5′ ends of the transcripts, including the addition of non–template-encoded bases.

Epicentre's ExactSTART technology enables the user to selectively tag the exact 5′ end of any RNA species present in a total RNA population. The ExactSTART Full-Length cDNA Library Cloning kit was originally designed to create a directionally cloned, full-length cDNA library for the identification of transcription start sites and other analysis of the 5′ ends of transcripts. For example, the kit has been used to discover alternative transcription initiation sites in a cDNA library produced from *Saccharomyces cerevisiae* (Vaidyanathan, R. *et al.* ExactStart™ Full-Length cDNA Library Cloning Kit: a rapid and efficient method to synthesize full-length cDNA for cloning and accurate mapping of transcription initiation and polyadenylation sites. *EPICENTRE Forum* 14.2, 4–5; 2007).

The powdery mildew fungus *Blumeria graminis* is one of the most important pathogens of cereal crops and can reduce crop yields by as much as 40%. One of the experimental challenges posed by *B. graminis* is that it can only be grown on its host; thus, the supply of biological material is very limited and may be contaminated by host tissues. One of the requirements of genome annotation is a collection of full-length cDNA sequences from as many diverse stages of the organism as possible. The advent of high-throughput DNA sequencing platforms has revolutionized the depth at which transcriptomes can be analyzed, and the development of robust and efficient protocols for generating cDNA that can be introduced directly in the sequencing pipeline is of huge importance. Here we describe minor modifications to adapt the ExactSTART technology to enable compatibility with 454 sequencing (**Fig. 1**).

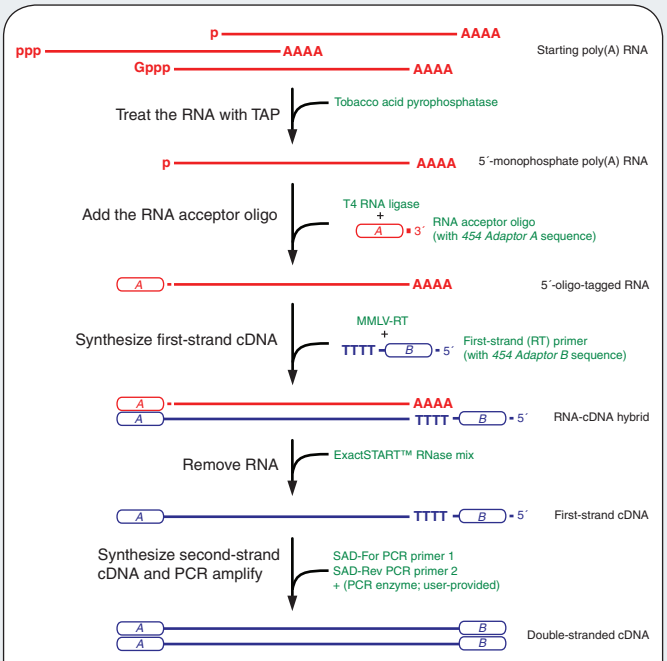

Figure 1 | Schematic representation of the modified ExactSTART cDNA synthesis procedure.

Preparation of cDNA for sequencing

We dissected epiphytic mycelia from barley leaves infected with *Blumeria graminis* f. sp. *hordei* using procedures previously described[1] and extracted total RNA from the fungal structures obtained from approximately 200 infected primary leaves[2], yielding approximately 150 µg of total RNA. Of this sample, we processed 40 µg of total RNA using Epicentre's mRNA-ONLY™ Eukaryotic mRNA Isolation kit to remove non-mRNA species.

Then we synthesized cDNA, following the ExactSTART Full-Length cDNA protocol (**Fig. 1**). The 5′ cap structure was removed from the mRNA using tobacco acid pyrophosphatase in a 10 µl reaction. In the following step, the RNA acceptor oligo was replaced with a custom-made oligoribonucleotide compatible with the 454 Adaptor *A* sequence (5′-GCCUCCCUCGCGUUAUCAGA-3′) and ligated to the decapped mRNA. Next, cDNA was synthesized using 20 µl of the ligated RNA

Pietro D Spanu[1] & Ken Doyle[2]

[1]Department of Life Sciences, Imperial College London, South Kensington Campus, London, UK. [2]Epicentre Biotechnologies, Madison, Wisconsin, USA. Correspondence should be addressed to K.D. (ken.doyle@epibio.com).

APPLICATION NOTES

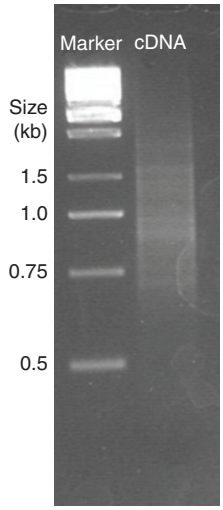

Figure 2 | Size range of cDNA produced. An aliquot of the double-stranded cDNA products was analyzed by agarose gel electrophoresis. The DNA was stained with SYBR® Green (Invitrogen) before electrophoresis. This cDNA was then directly sequenced as described in the text.

Figure 3 | The sequence heterogeneity found at the 5′ end of the sequenced transcripts for a randomly chosen group of sequences corresponding to the same genetic locus. The top sequence is an intron that is spliced out from the mature RNA and is not present in the cDNA. The consensus sequence of the RNA and of the genomic DNA is shown in the blue box. Dots indicate positions where presumed insertions of bases by RNA polymerase occur. A sample of the cDNA sequences obtained by 454 pyrosequencing is shown in the gray box. The actual start sites are shown only for sequences 14 to 40 (after trimming the 3′ end of the SAD-For primer sequence). Gaps in the consensus sequence after position 84 represent insertions into transcripts whose sequences are not shown here.

sample directly in the first-strand synthesis reaction, using a custom-made primer containing the 454 Adaptor *B* sequence (SAD-R poly(T): 5′-GCCTTGCCAGCCCGCTCAG$(T)_{25}$-3′). Second-strand cDNA synthesis and PCR amplification were carried out using Phusion DNA polymerase (NEB) in a 100 µl reaction. The primers used in the amplification reaction were modified to render them compatible with the 454 pyrosequencing protocol (SAD-For: 5′-GCCTCCCTCGCGCCATCAGA-3′; SAD-Rev: 5′-GCCTTGCCAGCCCGCTCAGT-3′). The cDNA yield was approximately 7 µg, and gel electrophoresis showed a DNA smear with a modal distribution of around 900 bp (**Fig. 2**). We sent the DNA to Roche Diagnostics for 454 pyrosequencing (one run on GS-FLX). The sample yielded 247,306 reads, a total of 50.8 megabases, corresponding to an average read length of 205 bases. The data were assembled (using the MIRA assembler; http://www.chevreux.org/projects_mira.html), clustered and combined with the expressed sequence tags (ESTs) available in our own databases and in public repositories. This increased the number of unique *B. graminis* genes identified by cDNA sequencing from 4,584 to 7,727.

Analysis of 5′ heterogeneity

When we compared the cDNA sequences to genomic DNA sequences, it became evident that there was a marked heterogeneity in the populations of RNA. This was true for the actual length of the sequence (possibly reflecting different starts of transcription and/or processing) and of the actual sequence of bases. **Figure 3** illustrates examples of the heterogeneity of one gene, which is representative of these findings. From this it is clear that some bases were added in the transcript at a considerable distance from the beginning of the mature transcript. The majority of these were adenosines, but thymine, cytosine and guanosine were also found. It should be noted that the raw data from the 454 pyrosequencing included four bases from the SAD-For primer, and these were accurately sequenced without exception. Therefore, the heterogeneity found at the 5′ end of the mRNA reflects *in vivo* reality, possibly because of inaccurate transcription by the RNA polymerase in the very first stages of the process. This phenomenon has also been noted by other studies in related fungi (for example, *Magnaporthe grisea*[3]) and other eukaryotic organisms[4].

Conclusions

As shown here, the ExactSTART procedure can be easily adapted to produce full-length cDNA for high-throughput sequencing analysis by modifying the tagging oligonucleotides and PCR amplification primers. The double-stranded, amplified cDNA produced can be directly used in 454 sequencing, providing valuable information about the 5′ heterogeneity associated with transcriptional start sites in many organisms.

ACKNOWLEDGMENTS

The data presented in this application note were obtained in collaboration with T.A. Burgis and J.C. Abbott, Imperial College London.

1. Both, M. *et al.* Transcript profiles of Blumeria graminis development during infection reveal a cluster of genes that are potential virulence determinants. *Mol. Plant-Microbe Interact.* **18**, 125–133 (2005).
2. Chomczynski, P. & Sacchi, N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal. Biochem.* **162**, 156–159 (1987).
3. Gowda, M. *et al.* Robust analysis of 5′-transcript ends: a high-throughput protocol for characterization of sequence diversity of transcription start sites. *Nat. Protoc.* **2**, 1622–1632 (2007).
4. Gowda, M. *et al.* Robust analysis of 5′-transcript ends (5′-RATE): a novel technique for transcriptome analysis and genome annotation. *Nucleic Acid Res.* **34**, e126 (2006).