Development of PrimePCR™ Assays and Arrays for lncRNA Expression Analysis

Jan Hellemans,[1] Pieter Mestdagh,[1] James Flynn,[2] and Joshua Fenrich[2]
[1] Biogazelle NV, Technologiepark 3, B-9052, Zwijnaarde, Belgium.
[2] Bio-Rad Laboratories, Inc., 2000 Alfred Nobel Drive, Hercules, CA 94547, USA

Real-Time PCR

Bulletin 6990

Abstract

The PrimePCR portfolio of real-time PCR gene expression assays has been expanded to include assays and array plates for human long noncoding RNA. lncRNAs play important regulatory roles in both normal and disease biology, but they also present certain challenges for qPCR analysis. Here we discuss the methodology used for the design of PrimePCR Assays and Arrays.

Introduction

Multiple distinct classes of expressed genes can be distinguished within our transcriptome. Messenger RNAs (mRNAs) are by far the most studied and best understood type of transcripts. It has become apparent, however, that a large fraction of our transcriptome does not code for proteins but is functional nonetheless. Two classes of noncoding RNAs can be distinguished within this group — small RNAs, including microRNAs (miRNAs), and long noncoding RNAs (lncRNAs). The latter constitutes a recent addition to our understanding of the transcriptome and significantly increases the number of known genes. In humans, the number of annotated noncoding genes now outnumbers that of protein-coding genes (Figure 1).

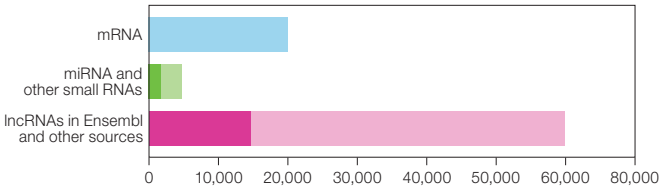

Fig. 1. Estimates of human genes by class. Estimated number of human mRNA genes (■), miRNA genes (■), and other small RNAs (■). The number of human lncRNAs in Ensembl (■) and other sources (■).

lncRNAs were first described as a new transcriptional unit during a large-scale sequencing of mouse full-length cDNA libraries in 2002 (Okazaki et al. 2002). They are defined as transcripts of at least 200 nucleotides in length, an arbitrary cutoff that has been in use since 2007 (Kapranov et al. 2007), and lack an obvious open reading frame. The majority of characterized lncRNAs are generated by the same transcriptional machinery as mRNAs, as evidenced by RNA polymerase II occupancy and histone modifications associated with transcription initiation and elongation (Guttman et al. 2009). lncRNAs share other mRNA features, such as a 5′ terminal methylguanosine cap, and are often spliced and polyadenylated. They may be located intergenically or overlap with protein coding genes in a sense or antisense orientation. More than half of mammalian coding genes have complementary antisense lncRNAs (Katayama 2005). lncRNA expression levels are typically lower than those of mRNAs (Figure 2), with many lncRNAs showing expression restricted to particular cell types or developmental contexts. Generally, lncRNAs have low sequence conservation across species.

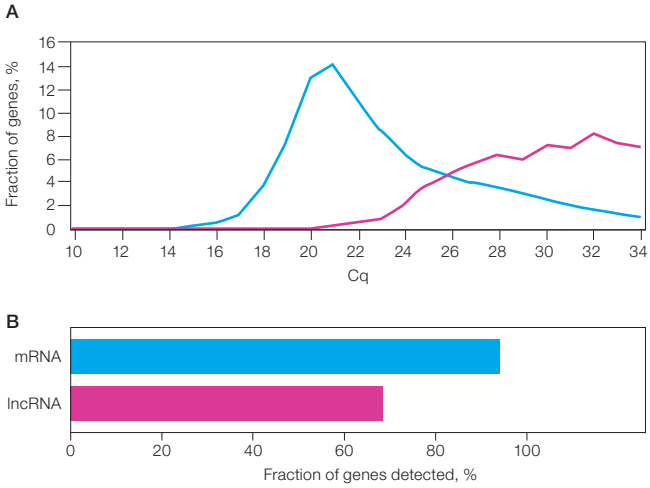

Fig. 2. Comparison of mRNA and lncRNA expression. A, expression landscape of 20,000 mRNAs and a selection of 5,000 lncRNAs measured on MAQC-A cDNA* using qPCR. mRNA (—); lncRNA (—). **B**, fraction of genes that were detected in MAQC-A cDNA. * Microarray Quality Control (MAQC)-A cDNA is a set of commercially available reference RNA samples from pooled human cell lines, containing large numbers of differentially expressed genes (Shi et al. 2006).

BIO-RAD

Although the exact function of most lncRNAs is still unknown, it has become clear that they exert important regulatory functions in normal biology and are implicated in various disease states.

Assay Design

The nature of lncRNAs and the relative immaturity of our understanding pose unique challenges for qPCR assay design. PrimePCR lncRNA Assays were designed with these considerations in mind:

Choice of reference database

Sequences from multiple genome repositories served as the basis for assay design. Ensembl (version 84), with ~15,000 lncRNAs, was a primary source as it is well structured and actively curated. RefSeq, a widely used alternative, was used for the 3,000 Ensembl lncRNAs with linked accessions. LNCipedia (Volders et al. 2013), with its comprehensive collection of lncRNAs, served as an additional source of sequence information. Combining these sources provided a foundation for creating optimal assay designs while considering all relevant genomic information.

Secondary structures

lncRNAs use multiple mechanisms to exert their regulatory functions. Often their secondary structures are important for interacting with proteins and regulatory domains. These secondary structures may also interfere with PCR amplification. Regions of the target sequence exhibiting strong secondary structures were therefore avoided while designing assays.

Specificity

Recent insights into transcriptome complexity can be used to help predict the specificity of qPCR assays. Specificity was evaluated against 20,000 well-curated protein coding genes and a comprehensive set of ~60,000 lncRNAs in LNCipedia (which also includes all annotated Ensembl lncRNAs). An additional challenge is the antisense nature of many lncRNAs. This means that locus specificity may not suffice to assure gene specificity. Due to typically lower lncRNA expression levels, even low fractions of (unspliced) RNA from the opposite strand may impact the measurement of lncRNA expression levels. Assays are designed to be specific and not have complete overlap with another gene on the same locus.

Transcript coverage

Assays are designed so that the majority of transcripts (>2/3) can be detected. With RefSeq, many genes have only one or two transcripts; assays are designed to detect at least half.

Other design parameters

Other settings, such as preferred amplicon size and GC content, targeting of gene regions that occur in all or most transcripts, and the avoidance of common SNPs overlapping primer or probe binding regions, are important design considerations taken into account in the design of all PrimePCR Gene Expression Assays (Hellemans et al. 2012). By utilizing the same parameters used to design mRNA assays, PrimePCR lncRNA Assays are designed to perform optimally under the same reaction conditions and amplification protocol.

PrimePCR designs are optimized to meet the largest number of design and in silico validation criteria possible. However, for some genes it is theoretically not possible to meet all criteria; for example, when a gene completely overlaps another gene on the opposite strand.

Assay Validation

The core algorithm behind the design of PrimePCR Assays has been extensively validated during the design of over 60,000 assays for protein coding genes in human, mouse, and rat samples. For lncRNAs, this design engine was updated and improved to handle the additional complexity of long noncoding gene target designs. Over 5,000 lncRNA assays designed using this improved engine were validated by qPCR and next generation sequencing (NGS). Lab procedures are identical to those used for validation of mRNA assays (Hellemans et al. 2012), with additional analysis for specificity against spliced and unspliced transcripts. When available, validation data can be located on the PrimePCR Assays web page.

Efficacy of intron-spanning designs to avoid co-amplification of genomic DNA (gDNA)

When possible, assays are designed to span introns to reduce the chance of amplifying gDNA. For the majority of intron-spanning assays, no gDNA amplification could be observed. However, at high gDNA input concentrations (2.5 ng/reaction), about 25% of these assays had Cq values below the single molecule cutoff (cycle 35 in the experimental conditions used, Figure 3). Because of potential partial homology with other genomic regions, this design approach provides only a limited guarantee against gDNA co-amplification. As such, it is recommended that gDNA elimination be performed prior to cDNA synthesis. Ideally, a control such as the PrimePCR gDNA Contamination Assay should also be used to monitor gDNA contamination.

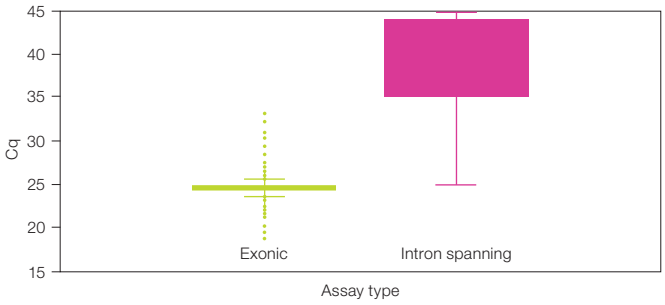

Fig. 3. Detection of human gDNA. Distribution of Cq values obtained from PrimePCR lncRNA Assays with exonic and intron-spanning designs when analyzing 2.5 ng of human gDNA. Exonic (■); intron spanning (■).

Positive no-template controls (NTCs)

Assays were tested with water to determine their tendency to amplify in the absence of template (typically due to primer-dimer formation). Assays that had an NTC Cq value less than 32 were considered failed and were replaced by an alternative design.

Amplification efficiency

In view of the typically low expression levels and sometimes cell type–specific expression of lncRNAs, PCR amplification efficiencies are most effectively determined on synthetic templates. Ideally, efficiencies should be close to 100%. Because of pipetting and measurement uncertainties, efficiencies of good assays are expected to follow a normal distribution of approximately 100%, with the majority in the 90–110% range (Figure 4). Most (99.2%) PrimePCR Assays have efficiency within this range (median ~96%), while 0.5% have acceptable efficiency in the 80–90% or 110–120% range. The remaining 0.3% of assays with efficiencies beyond these extremes were disqualified and replaced with alternative designs.

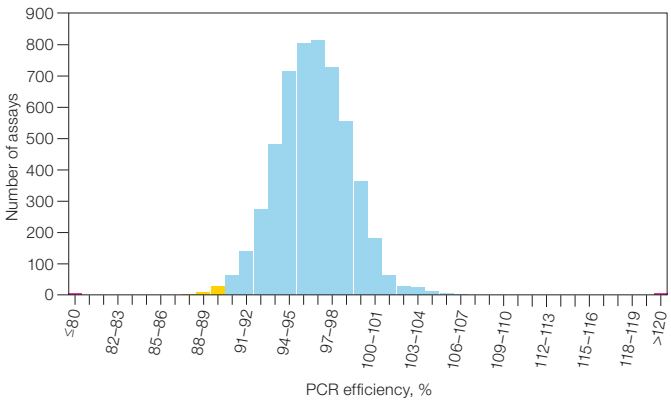

Fig. 4. Efficient amplification with PrimePCR lncRNA Assays. Depiction of the number of PrimePCR lncRNA Assays with ideal PCR efficiency (■), acceptable efficiency (■), and unacceptable efficiency where assays were replaced with alternatives (■).

Specificity

Amplicon sequencing was performed to offer a very sensitive assessment of potential off-target co-amplification. Over 85% of assays generated more than 1,000 reads, resulting in sufficient read depth to allow sensitive detection of off-target amplifications that occur less than 1% of the time. For genes with sufficient expression generating at least 100 reads, 95% of assays were completely specific to the intended target. When the specificity threshold was lowered to >90% specific reads, the number of specific assays grew to 97% (Table 1).

Table 1. Specificity analysis of PrimePCR lncRNA Assays by amplicon sequencing.

Specificity	>100 Reads
100%	95%
>90%	97%

Array Design

One of the challenges in designing lncRNA arrays is our lack of functional insights into the majority of lncRNAs. When it comes to lncRNAs, it will take many years to match the level of understanding that exists for protein coding genes. To overcome the limitation that this poses for the design of arrays, a novel approach was applied to the generation of PrimePCR lncRNA Arrays that associates lncRNAs with sets of coding genes through a combination of two methods.

The first method is a forward genetics approach where a pathway is perturbed and differentially expressed lncRNAs are identified. While this approach is less direct compared to a reverse genetics approach (that is, modulating the expression of the lncRNA followed by downstream pathway analysis), the forward genetics approach allows functional analysis of many lncRNAs for each of the pathways under investigation. By performing high-throughput pathway perturbation experiments, gene expression profiling, and integrative transcriptomic analysis, thousands of lncRNAs have been functionally characterized. More specifically, independent MCF-7 cell cultures were perturbed using 180 distinct chemical and genetic treatments, including siRNA-based silencing of 90 different human transcription factors along with drug treatments directed at 90 different protein-coding gene targets.

In a second approach, lncRNAs were linked to mRNA expression using guilt-by-association analysis. This method relies on the assumption that an lncRNA that is co-expressed with a pathway of interest is likely to either share a common upstream regulator with that pathway or be part of that pathway altogether. For this method, the gene expression profiles of various tumor and normal samples were analyzed to form association information between genes.

For each type of experiment, lncRNA and mRNA expression were quantified using total RNA sequencing, enabling detection of transcripts with and without polyadenylation (Figure 5). Perturbations that modulate either lncRNA or mRNA expression were identified by z-score analysis. mRNA z-scores were subsequently used to associate mRNA panels with each perturbation by means of Gene Set Enrichment Analysis (GSEA, Subramanian et al. 2005). For the tumor and normal tissue samples, co-expression of lncRNA and mRNA was analyzed. Panels of mRNA enriched among mRNAs that are positively or negatively correlated to an lncRNA were identified using GSEA. Each lncRNA was matched with one or multiple mRNA array panels.

By combining the lncRNA to mRNA array panel associations derived from these two independent analyses, and by including key genes identified in literature, arrays of lncRNAs were developed using an approach that is both unique and robust.

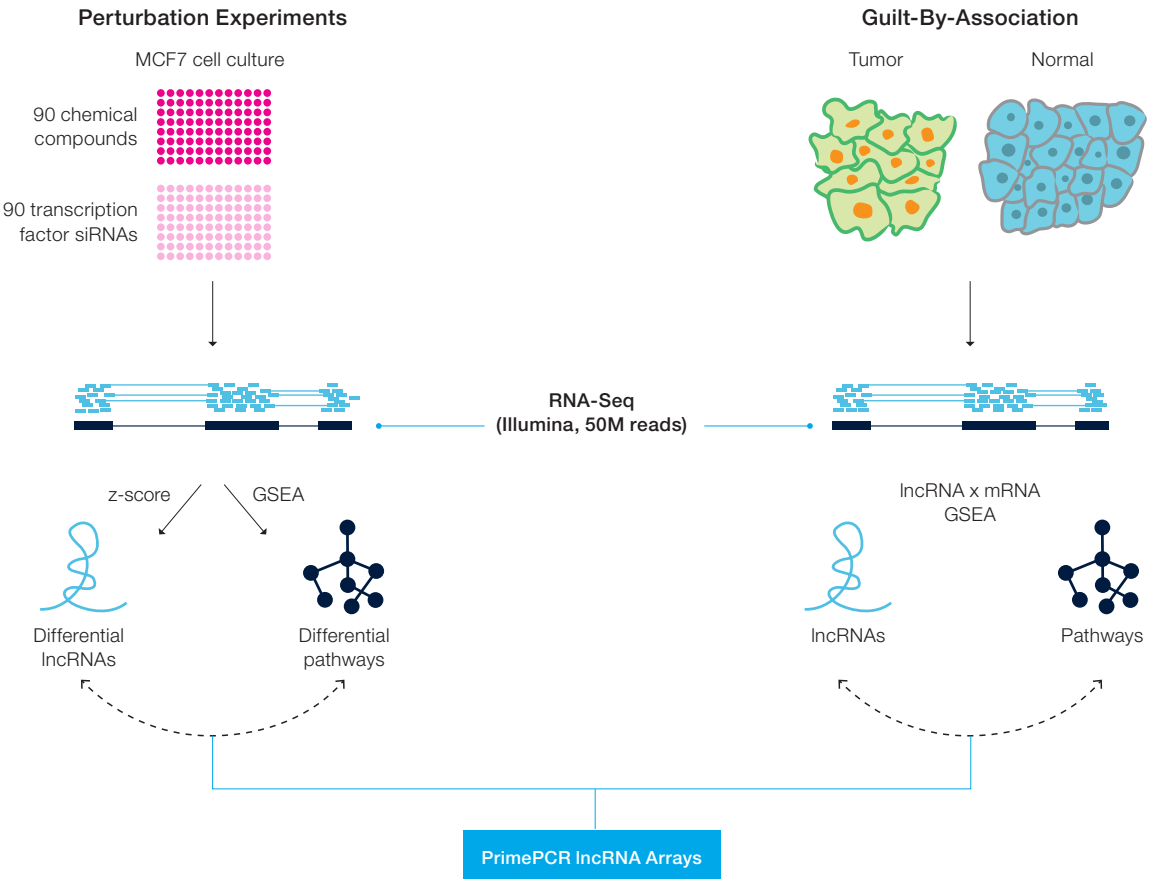

Fig. 5. Associating lncRNAs with pathway perturbations. Workflow used to associate the differential expression of lncRNAs and mRNAs, using 180 perturbations of MCF-7 cells, and a comparison of expression profiles in tumor and normal cells.

Summary

PrimePCR lncRNA Assays and Array Plates were meticulously developed to provide researchers with a reliable solution for analyzing long noncoding RNA expression by real-time PCR. Through the consideration of many factors during assay development and the use of a novel approach for array design, Bio-Rad aims to help accelerate research in this exciting and growing field.

References

Guttman M et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458, 223–227.

Hellemans et al. (2012). PrimePCR™ Assays: meeting the MIQE guidelines by full wet-lab validation. Bio-Rad Bulletin 6862.

Kapranov P et al. (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 316, 1,484–1,488.

Katayama S et al. (2005). Antisense transcription in the mammalian transcriptome. Science 309, 1,564–1,566.

Okazaki Y et al. (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420, 563–573.

Shi L et al. (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol 24, 1,151–1,161.

Subramanian A et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102, 15,545–15,550.

Volders PJ et al. (2013). LNCipedia: a database for annotated human lncRNA transcript sequences and structures. Nucleic Acids Res 41, D246–D251.

Visit bio-rad.com/lncRNAdesign for more information.

BIO-RAD

Bio-Rad Laboratories, Inc.

Life Science Group

Web site *bio-rad.com* ***USA*** *1 800 424 6723* ***Australia*** *61 2 9914 2800* ***Austria*** *43 1 877 89 01 177* ***Belgium*** *32 (0)3 710 53 00* ***Brazil*** *55 11 3065 7550* ***Canada*** *1 905 364 3435* ***China*** *86 21 6169 8500* ***Czech Republic*** *420 241 430 532* ***Denmark*** *45 44 52 10 00* ***Finland*** *358 09 804 22 00* ***France*** *33 01 47 95 69 65* ***Germany*** *49 89 31 884 0* ***Hong Kong*** *852 2789 3300* ***Hungary*** *36 1 459 6100* ***India*** *91 124 4029300* ***Israel*** *972 03 963 6050* ***Italy*** *39 02 216091* ***Japan*** *81 3 6361 7000* ***Korea*** *82 2 3473 4460* ***Mexico*** *52 555 488 7670* ***The Netherlands*** *31 (0)318 540 666* ***New Zealand*** *64 9 415 2280* ***Norway*** *47 23 38 41 30* ***Poland*** *48 22 331 99 99* ***Portugal*** *351 21 472 7700* ***Russia*** *7 495 721 14 04* ***Singapore*** *65 6415 3188* ***South Africa*** *27 (0) 861 246 723* ***Spain*** *34 91 590 5200* ***Sweden*** *46 08 555 12700* ***Switzerland*** *41 026 674 55 05* ***Taiwan*** *886 2 2578 7189* ***Thailand*** *66 2 651 8311* ***United Arab Emirates*** *971 4 8187300* ***United Kingdom*** *44 020 8328 2000*

Bulletin 6990 Ver A US/EG *17-0815 1217 Sig 1216*