Arraystar LncRNA Microarrays
2019 New Releases: Human V5.0, Mouse V4.0, Rat V3.0

Highlights
- Most sensitive and best technology for LncRNA profiling, superior to RNA-seq
- Comprehensive and robust full-length LncRNA* collection curated from all major latest databases and landmark publications
- Systematic and specialized LncRNA annotation, including genomic context, epigenomic context*, completeness*, subcellular localization**, mRNA recognition sites...
- Unambiguous, reliable and accurate detection and quantification of LncRNA transcript isoforms otherwise difficult by RNA-seq
- Simultaneous LncRNA and mRNA profiling on the same array for co-expression and correlational expression and regulation

Introduction
LncRNAs are a major RNA class in the transcriptome [1]. These noncoding RNAs are transcribed from genomic sites either in association with a protein coding gene nearby or in the intergenic regions as LincRNAs (Fig. 1), with functions in gene expression regulation by multiple mechanisms, either in cis or in trans, at transcriptional or post-transcriptional levels (Fig. 2). LncRNAs are a key player in a wide range of biological systems and diseases. Cutting edge LncRNA science has resolved many long standing mysteries in, for example, chromosomal inactivation, developmental and differentiation programming, and diseases of unknown etiology. In general, LncRNAs exhibit more restricted cell type-specific expression compared to mRNAs, making LncRNAs a class of higher specificity biomarker. With the broadened horizon and modern paradigm of studying gene regulation, the science of gene expression profiling has now gone beyond past mRNA-only to encompass both classes of the coding and non-coding RNAs.

Arraystar is the leader in LncRNA expression profiling technologies, using LncRNA Microarrays as the best performing platform to systematically profile LncRNAs together with mRNAs. To date, these microarrays have been an empowering tool and invaluable resource in LncRNA research touting many high impact publications. To incorporate rapid scientific advances and new data, Arraystar has now released new Human V5.0 and Mouse V4.0, and Rat V3.0 LncRNA Expression Microarrays.

Consolidated, comprehensive, robust, most up-to-date full-length LncRNA contents*

Unlike well-established protein coding genes, publically available LncRNAs are often sparsely annotated, partial in scope and scattered in collection. Large proportions of reported "LncRNAs" tend to be incomplete at 5’ or 3’ ends. Also, RNA-seq reads are not uniform in covering the 5’ and 3’ ends. These inaccurate and truncated LncRNA

---

* Applicable to Human V5.0 ** Applicable to Human V5.0 and Mouse V4.0
annotations can have a profound impact on downstream uses of the data, such as misinterpreting mRNA fragments as lncRNAs, unreliable transcript abundance estimate by FPKM, and misidentification of lncRNA promoter sites [1,2].

Arraystar maintains high quality proprietary transcriptome and lncRNA databases that extensively collect lncRNAs through all major external data sources, knowledge-based mining of scientific publications, and our lncRNA collection pipelines. Especially, we place premium attention on full-length lncRNAs collection. Full length lncRNAs as annotated or experimentally supported in the public databases are compiled with high priority. The lncRNAs in Arraystar proprietary transcriptome databases and published lncRNA studies are carefully assessed by supporting evidence for their sequence completeness: 5’ ends by host gene histone marks [3-5], CAGE cluster [6-9], and DNA hypersensitivity (DHS) [5]; 3’ ends by poly(A)-position profiling (3P-Seq) [10]. Additionally, lncRNA candidates are evaluated for protein coding potentials by a combination of prediction methods [11-13]. Only the lncRNAs that pass these assessments are curated into the full length lncRNA collections (Fig. 3).

Towards systematic and functional annotation of LncRNAs

The Arraystar LncRNA microarray package includes systematic and detailed lncRNA annotations, subclassification, and analyses to gain insight into the complex biological functions of the lncRNAs. LncRNAs with reported biological processes or associated with human diseases are researched, annotated and cross referenced. This rich source of information helps to unravel functional roles and molecular mechanisms of the LncRNAs.

Genomic context. LncRNAs are systematically classified based on their genomic relationships with the nearest protein coding genes into Intergenic (LncRNA), Intronic, Bidirectional, Sense-overlapping, Antisense and Pseudogene LncRNAs (Fig. 1). These subclasses help dissect various cis- or trans-regulatory functions on the target genes transcriptionally or post-translationally (Fig. 2).

Epigenomic context. LncRNAs can be transcribed in and regulated by a promoter or enhancer region with characteristic promoter or enhancer epigenetic marks [5]. Many active promoter and enhancer regions are themselves transcription units, capable of generating functionally active noncoding RNAs for these cis-regulatory DNA elements. The LncRNAs are thus classified into promoter-lncRNAs (p-lncRNA) and enhancer-lncRNAs (e-lncRNA) based on the epigenomic context (e.g. DNase I hypersensitive sites). The p-lncRNAs are further grouped into intergenic and divergent p-lncRNAs based on their genomic context (Fig. 4). p-lncRNAs are often positively correlated with transcription of their protein-coding genes under the same
promoters. e-IncRNAs often trap TF proteins to the local sites, modify the local chromatin environment, and organize three-dimensional nuclear topology domains for correct activation of the target gene program.

Fig 4. Promoter and enhancer IncRNA categories based on the epigenomic and genomic context. IncRNAs are classified into intergenic p-IncRNA, divergent p-IncRNA, e-IncRNA, and other; based on their TSS and DNase I hypersensitive sites (DHS) in the promoter (marked by H3K4me3), enhancer (marked by H3K4me1, H3K27ac and H3K36ac), or dyadic regulatory (enhancer-promoter alternating states) regions.

Completeness* The sequence completeness of the IncRNA 5’ and 3’ ends is important for many IncRNA follow-up studies, e.g., the location of the accurate IncRNA transcription start site (TSS), the promoter region, or CRISPR-Cas screen targeting site design. Here, the IncRNA end completeness statuses are annotated as: Complete 5’ end, Complete 3’ end, and Full length (complete both 5’ and 3’ end).

Subcellular localization** The molecular functions of IncRNAs are tightly coupled with their subcellular localization [14-16]. For example, IncRNAs localized in the nucleus or chromatin often regulate the gene expression by epigenetic modification and transcription. IncRNAs in the cytoplasm are more likely involved in translation regulation or miRNA sponging such as competing endogenous RNAs (ceRNA)[17-20].

miRNA recognition site Predicted or experimentally identified microRNA sites on the IncRNAs are annotated to indicate potential post-transcriptional regulatory functions in the miRNA regulatory network, such as acting as competing endogenous RNAs (ceRNA).

Highly conserved IncRNAs Certain IncRNA genes harbor ultraconserved regions (UCR) or ultraconserved non-coding elements (UCNE) that do not vary in sequence across species, which imply these sequences being biologically indispensable [21-27]. As many IncRNAs regulate target genes by cis-mechanism, human IncRNAs syntenic to orthologous IncRNAs in other species are also collected even with modest homology, as their genomic context with the neighboring target genes, rather than the sequence conservation, can be more relevant in gene regulation [28].

Tissue specific IncRNAs* The function of a IncRNA can be directly or indirectly related to and indicated by the tissue or cell type in which it is specifically expressed. In Human LncRNA Microarray, 6,059 cell lineage and cancer associated IncRNAs are annotated.

Disease-associated IncRNAs* LncRNAs known to be associated with diseases, such as cataloged in LncRNADisease [29,30], are annotated for clinical and translational investigations.

Coding potential for small peptides* Although most IncRNAs are noncoding, some IncRNAs can contain small open reading frames (smORFs) to encode small peptides [31], as predicted or experimentally detected as cataloged in LncRNAWiki [32].

*LncRNAs known to be associated with diseases, such as cataloged in LncRNADisease [29,30], are annotated for clinical and translational investigations.

LncRNA transcript isoforms

LncRNAs, just like miRNAs, can be alternatively processed as transcript isoforms and have distinct functions. Arraystar LncRNA Microarrays use “transcript-specific” probes that hybridize to the splice junctions or exon sequences that are unique to each transcript isoform from the same gene (Fig. 5). Compared with the microarrays, RNA-seq performs poorly in transcript-specific profiling due to short sequencing reads, low isoform/splice junction read coverage, and inherent computational complexity.

Fig 5. Arraystar LncRNA Microarray transcript-specific probes unambiguously and accurately detect and quantify transcript isoforms BCL-XL, BCL-XS, and ENST412972 having distinct oncogenic functions. The “Gene-specific” probes not designed for IncRNA isoforms cannot make such distinction. The arrows indicate the transcription direction.
Why use arraystar LncRNA microarray over RNA-seq for LncRNA profiling?

LncRNAs often express and function at low abundance, buried in other classes of abundant RNAs (Fig. 6A). There are serious limitations of RNA-seq for LncRNA profiling.

LncRNA quantification. For more detection of the presence of a LncRNA, a few reproducible sequencing reads should suffice. But for quantification, at least hundreds read counts are required to reliably represent the RNA level [33] (Fig. 6A). LncRNAs are generally ~10X less abundant than mRNA[34]. RNA-seq quantification at these low LncRNA levels is unacceptably poor and not nearly sufficient for differential expression analysis [2,35] (Fig. 6C, 6D). Even if the sequencing coverage is increased to an unaffordably deep coverage (dotted curve; several hundred times the normal RNA-seq coverage at 20 mil), a large proportion (40%) of transcripts can never be reliably quantified [33] (Fig. 6B). Additionally, FPKM (Fragments Per KiloBase of transcript per Million mapped reads) calculation in RNA-seq depends on accurate LncRNA transcript model lengths, many of which still lack completeness in LncRNA annotation [1]. In contrast, LncRNA Microarray oligo probes hybridize the target RNA at high affinity, independent of other abundant RNAs. The microarrays are highly sensitive and accurate even for low abundance LncRNAs [37](Fig. 6D).

LncRNA transcript isoforms. LncRNAs often have multiple transcript isoforms and function differently in complex genomic and regulatory relationships with their target mRNA genes. Profiling LncRNAs at transcript-specific level is important. However, RNA-seq coverage for the splice profiles is weak and non-uniform, particularly for non-predominant isoforms [2] (Fig. 7). Even at saturating coverage, accurate reconstruction of transcript isoform is inherently challenging due to the missing connectivity information with the short reads in distant exons on the same RNA fragment [2]. These make reconstructing LncRNA transcript isoforms and quantification very difficult [38-41]. For LncRNA Microarrays, the transcript-specific array probe design is based on well-established transcript models for each LncRNA isoform, which is unambiguous and highly accurate in isoform detection and quantification (Fig. 5).

LncRNA annotation and analysis. Unlike well established and curated protein coding genes, RNA-seq raw data are still in need of well-researched and consolidated reference bases for mapping and annotation, which are not readily publically available. Arraystar Microarray LncRNA contents are based on the foundation of high quality proprietary Arraystar LncRNA transcriptome databases that extensively collect LncRNAs through all major public databases and repositories, knowledge-based mining of scientific publications, and our LncRNA collection pipelines. The microarray annotation and analyses are, rich, detailed, and comprehensive, unrivaled by any other profiling platforms.

Fig 6. (A) The median LncRNA expression level is approximately 10X lower than that of mRNAs (based on GENCODE data) [34]. (B) Top 1% of the highest expressed genes, such as housekeeping genes, occupy ~40% of RNA-seq signal. Lowly expressed LncRNAs receive very little sequencing coverage [35]. (C) In a typical mRNA-seq depth at 40 million reads, < 10% LncRNAs can be reliably quantified [36]. (D) While quantitative error becomes unacceptably high for RNA-seq when the RNA level is low, microarray continues to perform very well [37].

Fig 7. Compared with better expressed mRNAs, lowly expressed LncRNA isoforms cannot be adequately covered by short RNA-seq reads to reconstruct the exon models nor their quantification[2].
| Table 1. LncRNA Microarray vs RNA-seq for IncRNA profiling |
|---------------------------------|-----------------|
| **LncRNA Microarray** | **RNA-Seq** |
| High sensitivity and quantification accuracy for IncRNAs as low as 1 transcript/cell. | Most IncRNAs at low levels cannot be accurately and reliably quantified. |
| Natively specific for RNA strandedness for both sense and antisense IncRNAs. | Stranded RNA-sequencing library prep required. |
| Unambiguous and specific IncRNA isoform detection/quantification. | Poor sensitivity and accuracy for IncRNA isoforms. |
| Arraystar LncRNA Microarray premium IncRNA collection, annotation and analyses. Entire coding miRNA gene set also included. | Public IncRNA reference databases can be deficient. Systematic IncRNA annotation and analyses are not readily available for the RNA-seq data. |

**Arraystar LncRNA Array Specifications**

<table>
<thead>
<tr>
<th></th>
<th>Human V5.0</th>
<th>Mouse V4.0</th>
<th>Rat V3.0</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total number of distinct probes</strong></td>
<td>60,491</td>
<td>60,641</td>
<td>38,352</td>
</tr>
<tr>
<td><strong>Probe length</strong></td>
<td></td>
<td></td>
<td>60 nt</td>
</tr>
<tr>
<td><strong>Probe selection region</strong></td>
<td>Specific exon or splice junction along the entire length of the transcript</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Probe specificity</strong></td>
<td>Transcript-specific</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Labeling method</strong></td>
<td>3'RNAs are labeled along the entire length without 3' bias, even for degraded RNA at low amount.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>mRNAs</strong></td>
<td>21,174</td>
<td>30,924</td>
<td>27,770</td>
</tr>
<tr>
<td><strong>IncRNAs</strong></td>
<td>40,173</td>
<td>37,949</td>
<td>10,582</td>
</tr>
<tr>
<td><strong>Gold Standard LncRNAs</strong></td>
<td>8,393</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Reliable LncRNAs</strong></td>
<td>30,924</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>mRNA sources</strong></td>
<td>Refseq, UCSC, GENCODE, FANTOM5 CAT</td>
<td>Refseq, Known Gene, GENCODE</td>
<td>Refseq, Ensembl</td>
</tr>
<tr>
<td><strong>LncRNA sources</strong></td>
<td>Databases (up to 2018): FANTOM5 CAT (v1); GENCODE (v25); RefSeq (2018.11); BigTranscriptome (v1); knownGene (2018.11); LncRNAdb; LncRNAWiki; RNAdb; NRED; CLS FL; NONCODE (v5); miTranscriptome (v2)</td>
<td>Databases (up to 2018): GENCODE (v18); RefSeq; KnownGene; GenBank</td>
<td>Databases (up to 2018): Refseq, Ensembl</td>
</tr>
<tr>
<td><strong>Literatures</strong></td>
<td>(up to 2018)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Arraystar LncRNA collection pipelines</strong></td>
<td>(up to 2018)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Array format</strong></td>
<td>8 × 60K</td>
<td>8 × 60K</td>
<td>4 × 44K</td>
</tr>
</tbody>
</table>

**LncRNA Study Roadmap**

![Image of LncRNA Study Roadmap]

**Fig 8.** LncRNA research roadmap for studying the identified differentially expressed IncRNAs, for their regulatory molecular mechanisms, biological functions, and biomarker development.
Selected Publications

Since Arraystar launched the first commercial LncRNA Microarray in 2009, over 400 publications citing our superior LncRNA array service were published in top scientific journals.

NKLIA lncRNA promotes tumor immune evasion by sensitizing T cells to activation-induced cell death. Huang D, et al. *Nature Immunology*, 2018


References