

Supplementary material for

dbEssLnc2.0: exploring disease-associated essential long non-coding

RNAs in human cell lines

Supplementary methods - Essential lncRNA gene annotations

We process the CRISPR based datasets and the ClinVar based datasets separately. For the CRISPR based datasets, four primary datasets that are collected from literatures were strategically merged. Due to differences in experimental designs and statistical approaches, the essentiality criteria varied across different primary CRISPR datasets. To identify essential lncRNA genes, which impacts cell growth and proliferation, we adhered to the specific growth score cutoff values or FDR thresholds, which are established by each individual study. Specifically, for the CRISPR-i screening [1], lncRNA genes with a growth score over 7 were considered significantly affecting cell growth. This set was subsequently filtered, retaining only those with a negative phenotype score. All lncRNA genes with a positive phenotype score were excluded, as a positive phenotype score suggests an enhanced proliferation. In the CRISPR-splice screening [2], lncRNA genes with a cell viability score greater than 2 were considered essential. For the CRISPR-delete [3] and the CRISPR-CasRx [4] screenings, lncRNA genes with an FDR < 0.25 are regarded as essential. The corresponding cell line information with annotations were also collected for the essential lncRNA genes. All genomic coordinates on hg19 [1,3] were converted to hg38 using LiftOver v483. All essential lncRNA gene records from CRISPR based screenings were pooled together to generate a dataset of 1190 human essential lncRNA genes. For the ClinVar based dataset, the variants with lethal phenotype, which were collected as we have mentioned, were mapped directly to the entire NONCODE v6 database (<http://www.noncode.org/>) and the LncBook v2.0 database (<https://ngdc.cncb.ac.cn/lncbook/>) using BEDTools v2.30.0. A total of 1319 putative disease-associated essential lncRNA genes were collected.

We curated 2509 human essential lncRNA genes. These lncRNA genes were further annotated by mapping them to multiple public reference databases, including NONCODE v6.0, NCBI Genes (<https://www.ncbi.nlm.nih.gov/gene/>), LncBook v2.0, and GENCODE v47 (<https://www.encodegenes.org/human/>). When mapping the CRISPR-based essential lncRNA genes, the genomic coordinates of genes, transcripts and exons of a given lncRNA gene were used as query ranges against other reference databases. The mapping required that, for a given lncRNA gene, at least one query range be exactly identical to, or entirely contained within, at least one corresponding reference range on the same strand. When an essential lncRNA gene were mapped to multiple genes in reference databases, we calculate the proportion of cumulative length of overlapping exon regions in the mapped reference region, as follows:

$$p = \frac{q}{r}, \quad (1)$$

where q is the cumulative length of overlapping exon regions in the mapping, r the total length of mapped reference region containing overlapping exon regions. The mapped reference lncRNA genes were prioritized according to p in Eq(1). Genomic sequences were extracted from the ensemble database, after the mapping procedure. When mapping the ClinVar based essential lncRNA genes, we first used the ID conversion tool of the LncBook v2.0 database to supply NCBI id, gene name and ensemble id. Genomic sequences were obtained from the LncBook v2.0 and NONCODE v6.0 databases.

Finally, we updated all records in dbEssLnc1.0 [5] to the standard of v2.0. Gene summaries were obtained from NCBI Genes database for all essential lncRNA genes in dbEssLnc2.0. Tissue specific expression values were obtained across 32 tissues from the LncExpDB (<https://ngdc.cncb.ac.cn/lncexpdb/downloads>). For any missing value or annotation, an "N.A." (Not Available) was marked as a placeholder.

Supplementary Methods: Prompt Engineering and Iterative Refinement for AI-Assisted

Rule Set Generation

This document details a multi-layered rule set developed through extensive manual curation combined with AI-assisted recommendations from Google Gemini 2.5 Pro, designed for the automated classification of phenotype descriptions in the ClinVar database.

The primary goal of this rule set is to systematically identify phenotypes associated with severe, life-threatening genetic disorders, thereby facilitating the inference of clinical severity for related genetic variants. Detailed raw materials and code scripts are stored at https://github.com/auggieyd/dbesslnc2.0/tree/master/data_curated/clinvar_map/phenotype.

S1. Prompt Engineering

The classification rule set was developed through an iterative prompt engineering process with a Large Language Model (LLM), Gemini 2.5 Pro, designed to emulate an expert-driven refinement workflow. This methodology transformed a general request into a highly specific and clinically nuanced classification system. The key stages of this process are outlined below.

Stage 1: Initial Broad Prompting and Objective Definition

The process commenced with a high-level directive, instructing the AI to assume the role of a clinical geneticist and develop a system to identify lethal phenotypes from the provided ClinVar data (unique_phenotypes.txt).

Prompt 1: "Act as an expert clinical geneticist. Analyze the provided pheno.txt file containing phenotype descriptions from ClinVar. Develop a defined rule set to filter for phenotypes that are lethal. Lines containing semicolons should be parsed as multiple, independent phenotypes."

Stage 2: Concept Scoping by Refining the Definition of "Lethality"

The initial output was overly inclusive. The definition of "lethal" was subsequently constrained to enhance precision, focusing on conditions with severe, early-onset prognoses.

Prompt 2 (Refinement): "The initial definition of 'lethal' is too broad. Refine the rules to specifically target 'highly lethal' phenotypes, defined as conditions that typically cause death in the perinatal, neonatal, infantile, or early childhood periods and for which effective curative treatments are generally unavailable."

Stage 3: Case-Based for Rule Correction and Enhancement

To move beyond simple keyword matching and address complex clinical semantics, specific phenotype strings were presented to the AI.

Prompt 3.1: "The current classification system fails to identify 'Leukoencephalopathy, progressive, infantile-onset, with or without deafness' as lethal. Analyze the existing rules to identify the logical failure. Propose and implement a more robust, pattern-based rule (e.g., combining severity, onset, and system keywords) to correctly classify this phenotype and similar complex descriptions."

Prompt 3.2: "The classification system is incorrectly flagging phenotypes such as 'Marfan syndrome, mild' and 'lidocaine-induced Brugada syndrome' as high-risk. These represent non-lethal clinical scenarios. Explain the semantic importance of modifiers like 'mild,' '-induced,' and 'without [non-critical feature].'. Update the rule set to implement a high-priority exclusion logic that correctly interprets these mitigating modifiers."

Stage 4: Final Rule Set Architecture

Following multiple cycles of refinement, a final directive was issued to consolidate all learned logic

into a definitive, multi-layered classification architecture.

Prompt 4: "Based on all previous iterations, construct the final, definitive rule set. The output should be structured to include the phenotype, its classification, the specific clinical category, and the rule or pattern that was matched.

S2. Algorithm and Logic Flow

The classification algorithm processes each unique phenotype string through a sequential, prioritized pipeline:

Step 1: High-Priority Exclusion (Rule Zero): The algorithm first applies a set of high-priority exclusion rules to filter out phenotypes that are unambiguously non-severe, represent risk states, or contain mitigating modifiers. A match at this stage results in a "Non-Severe" classification, and no further rules are evaluated.

Step 2: General Exclusion: Phenotypes that pass Rule Zero are then checked against a list of non-lethal, late-onset, or effectively manageable conditions.

Step 3: Confirmed Lethal Classification: The remaining phenotypes are evaluated against high-confidence rules designed to identify phenotypes with near-certain perinatal, neonatal, or infantile mortality.

Step 4: High Severity Classification: Phenotypes not meeting the Confirmed Lethal criteria are assessed against a broader set of rules that capture conditions associated with significant premature mortality, catastrophic clinical events, or a fatal natural history.

Step 5: Default Classification: Any phenotype that does not match an inclusion or exclusion rule is assigned a default classification of "Non-Severe".

S3. The Rule Set

The complete rule set, encompassing Exclusion Rules (Table S1), the Core Lexicon for pattern matching (Table S2), and Inclusion Rules (Table S3), is provided in the Supplementary Material. Please refer to Tables S1, S2, and S3 in supplementary table file for full details.

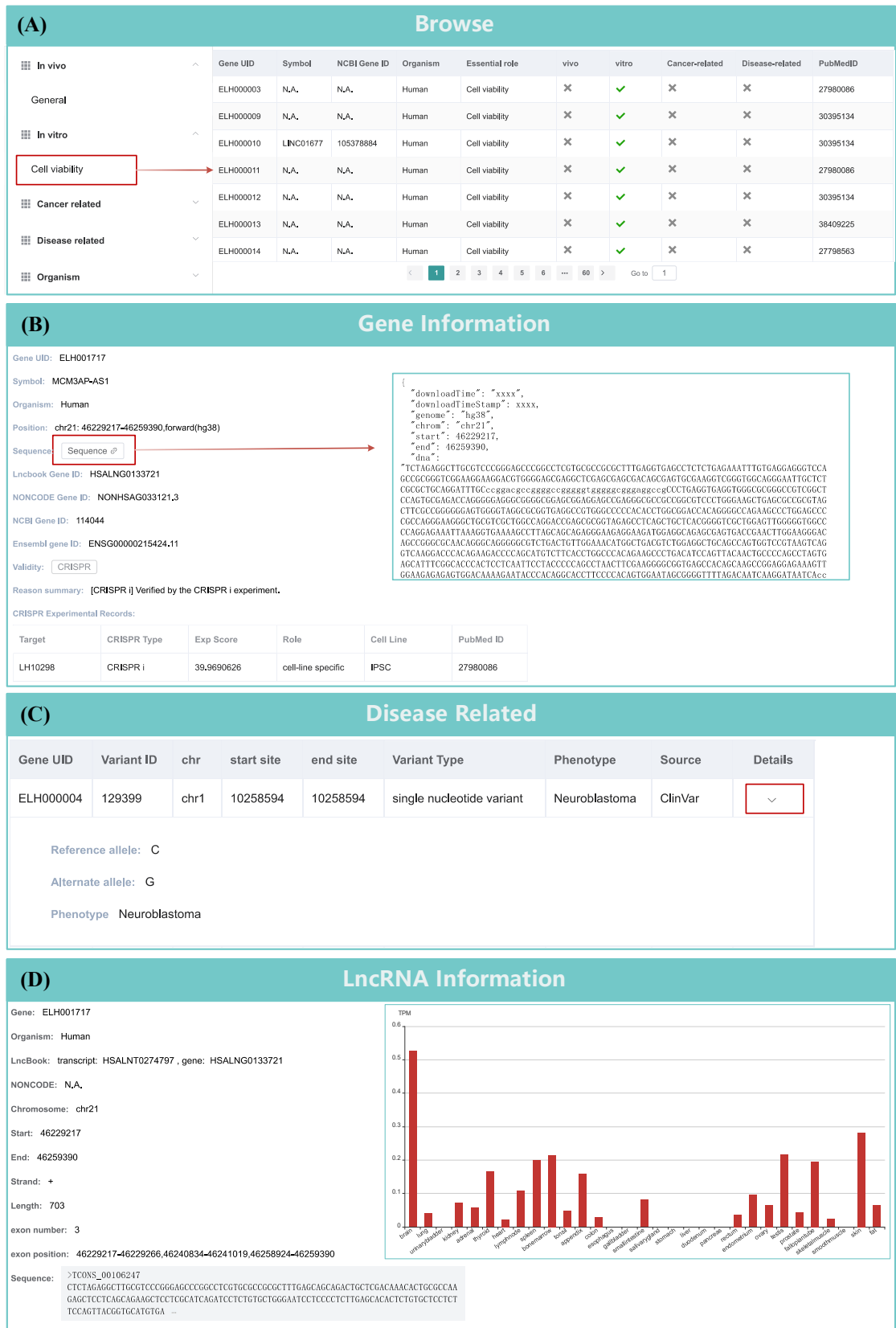
S4. Final list of lethal phenotypes

Detailed lists of the lethal phenotypes are provided in Table S4 of the supplementary table file.

References

- 1 Liu S J, Horlbeck M A, Cho S W, Birk H S, Malatesta M, He D, Attenello F J, Villalta J E, Cho M Y, Chen Y, Mandegar M A, Olvera M P, Gilbert L A, Conklin B R, Chang H Y, Weissman J S, Lim D A. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science*, 2017, 355(6320): eaah7111
- 2 Liu Y, Cao Z, Wang Y, Guo Y, Xu P, Yuan P, Liu Z, He Y, Wei W. Genome-wide screening for functional long noncoding RNAs in human cells by Cas9 targeting of splice sites. *Nature Biotechnology*, 2018, 36(12): 1203–1210
- 3 Zhu S, Li W, Liu J, Chen C H, Liao Q, Xu P, Xu H, Xiao T, Cao Z, Peng J, Yuan P, Brown-M, Liu X S, Wei W. Genome-scale deletion screening of human long non-coding RNAs using a paired-guide RNA CRISPR–Cas9 library. *Nature Biotechnology*, 2016, 34(12): 1279–1286
- 4 Montero J J, Trozzo R, Sugden M, Öllinger R, Belka A, Zhigalova E, Waetzig P, Engleitner T, Schmidt-Supprian M, Saur D, Rad R. Genome-scale pan-cancer interrogation of lncRNA dependencies using CasRx. *Nature Methods*, 2024, 1–13
- 5 Zhang Y Y, Zhang W Y, Xin X H, Du P F. dbEssLnc: A manually curated database of human and mouse essential lncRNA genes. *Computational and Structural Biotechnology Journal*, 2022, 20: 2657–2663

Supplementary Fig. S1



Supplementary Fig. S1. User interface of the dbEssLnc2.0 database. (A) The "Browse" page includes a navigation sidebar on the left and a corresponding category table on the right; (B) The "Gene" page for gene annotations; (C) The alternative "Gene" page for disease-related details; (D) The "Visual" page for

lncRNA information and tissue specific expression profile.

Supplementary Fig. S2.

(A)

Search from Database

Human

MC_%1

Search results

| Gene UID | Symbol | NCBI Gene ID | LncBook Gene ID | NONCODE Gene ID | Organism | PubMedID | General | Cell-viability | Cancer-related | Disease-related |
|-----------|------------|--------------|-----------------|-----------------|----------|----------|---------|----------------|----------------|-----------------|
| ELH002463 | MCPH1-AS1 | 100507530 | HSALNG0063153 | NONHSAG049421.3 | Human | 27798563 | ✗ | ✓ | ✗ | ✗ |
| ELH001717 | MCM3AP-AS1 | 114044 | HSALNG0133721 | NONHSAG033121.3 | Human | 27980086 | ✗ | ✓ | ✗ | ✗ |
| ELH002210 | LINC00472 | 79940 | HSALNG0051162 | NONHSAG094269.2 | Human | 30522853 | ✗ | ✗ | ✓ | ✗ |

(B)

Blast

>TCONS_00106247
CTCTAGAGGCTTGGCTCCCGGGAGCCCGGCTCGTGGCGCGGCTTTGAGCAGCAGACTGCTCGACAAACACTGCGCCAAAGAGCTCCTCAGCAGAAGCTCCTCGCATCAGATCCTGTGTGTGGGAATCCTCCCTC
TTGAGCACACTCTGTCTCTTCCAGTTACGGTGCATGTGAAGCAATGGTATGGGAAATTTGTTGCAGAAGGATGAAAAGGCTTTATTGCCAACTGAACACAGGACTCACCCTGTAGATACCTGCAGAGCA
CTGAAGCTCCTGGAGGCTCTCTTTGAGTCTGGAGATTTCTCCACGAGAAACAAGTCCACTAAGTGGGCACAGACATCCTCAGCAGAACGGGCCACAGGACCTCTGGTCTGTCTCTACTGCATTCTAGAAA
CAGGGCAATCAGCATGGAAGACACTGCACTTGGGGCCACAGACACTGAGGGCTTGGTTGAAAAGTCCAAGACTCAGTCAGGGCGGCTGGCTCAGCCTGTAATCCAGCACTTTGGAAGCCGGAAGCGGTGGATC
ATGAGGTCAAGAGATCCAGCATCCTGGCTAACATGGTGAACCCCTGTCTCTACTAAAAATACAAAAAATTAGCCTGGTGTGGTGGCGGGCGCCTGTAGTCCAGCTACTCGGAGGCTGAAGCAGGAGAATGAC
GTGAAGCCGGGAGGCAGA

e-value: 1e-7

word size: 11

Blast

Example

| Detail | Subject seq id | Percentage(identical matches) | Alignment length | e-value | Bitscore |
|--------|----------------|-------------------------------|------------------|---------|----------|
| ▼ | TCONS_00106247 | 100.000 | 703 | 0.0 | 1299 |

Organism: Human

Gene UID: ELH001717

transcript_id: TCONS_00106247

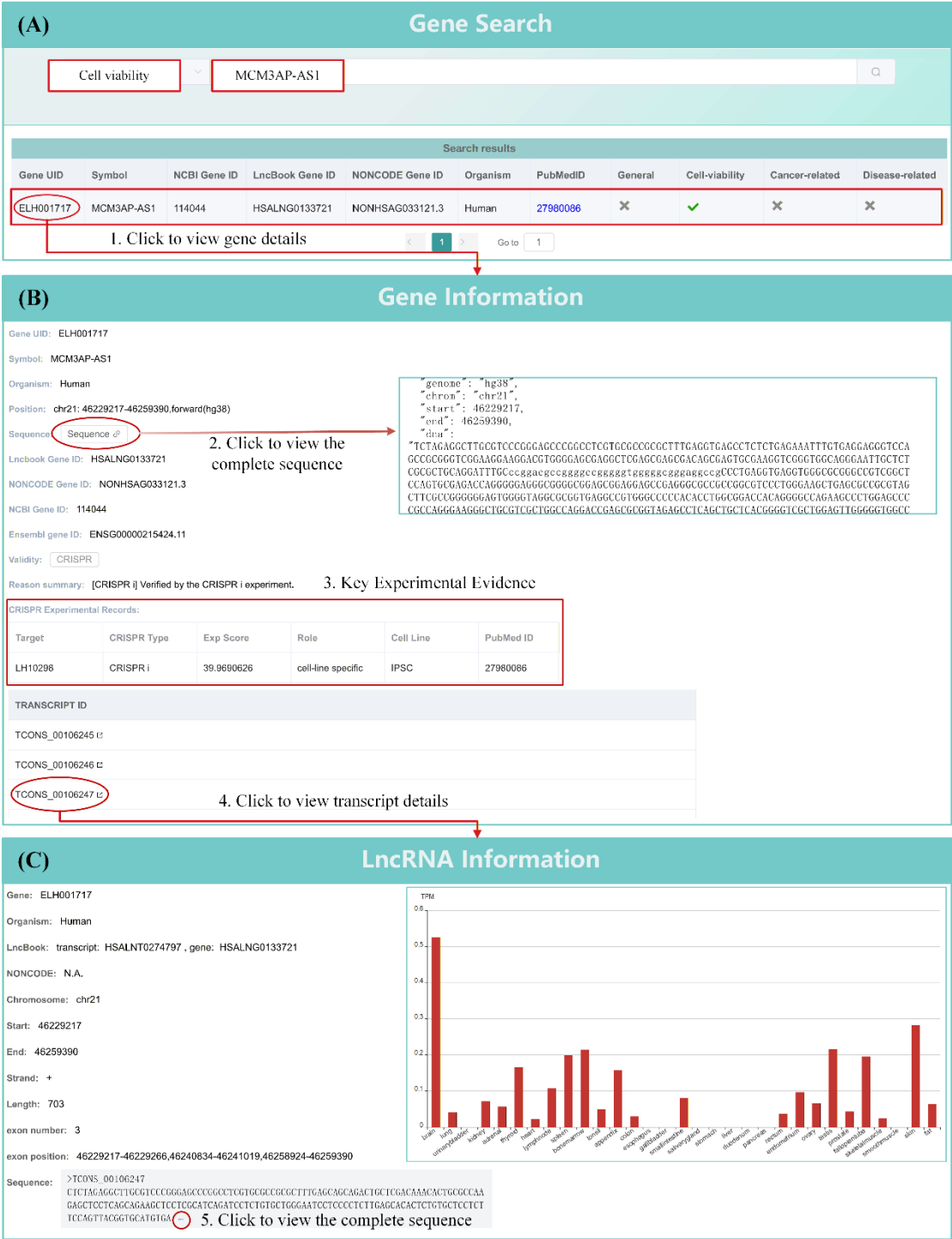
Position: chr21:46229217-46259390,forward(hg38)

LncRNA Sequence Length: 703

Sequence: >TCONS_00106247
CTCTAGAGGCTTGGCTCCCGGGAGCCCGGCTCGTGGCGCGGCTTTGAGCAGCAGACTGCTCGACAAACACTGCGCCAAAGAG
CTCCTCAGCAGAAGCTCCTCGCATCAGATCCTGTGTGTGGGAATCCTCCCTCTTGAGCACACTGTGTCTCTTCCAGTTA
CGGTGCATGTGA

Supplementary Fig. S2. Search interface and online BLAST Service. (A) Users can search for lncRNAs by keywords like "gene UID," "LncBook ID," or "gene name." This search refines using parameters such as species ("Human," "Mouse") or "Cell viability." The interface features auto-completion and supports wildcard queries ('_' for single-character matching; '%' for multiple-character sequences). Results appear in a table, allowing direct navigation to detailed gene pages via "Gene UID" or original publications via "Literature ID." (B) Complementing this, an integrated BLAST service facilitates sequence similarity searches for homologous lncRNAs, with results displayed in a concise, expandable list.

Supplementary Fig. S3



Supplementary Fig. S3. Use-case example demonstrating a typical workflow in dbEssLnc2.0. (A) The user performs a search for the lncRNA MCM3AP-AS1 using the search interface, filtering by essential role ('Cell-viability'). The results are displayed in a table, from which the user can navigate to the detailed gene page by clicking the "Gene UID". (B) The detailed gene page for MCM3AP-AS1 (ELH001717) provides comprehensive, integrated annotations. This includes general gene information with cross-references to other major databases, specific CRISPR i experimental data confirming its essentiality in the iPSC cell line. (C) Navigating to the transcript "LncRNA Information" page reveals further details, including the exon structure, transcript sequence, and an integrated tissue expression profile.