Gong *et al. BMC Genomics* (2017) 18:434
DOI 10.1186/s12864-017-3817-0

BMC Genomics

SOFTWARE Open Access

lncRNA-screen: an interactive platform for computationally screening long non-coding RNAs in large genomics datasets

Yixiao Gong[1,2], Hsuan-Ting Huang[3], Yu Liang[4], Thomas Trimarchi[5], Iannis Aifantis[1,2*] and Aristotelis Tsirigos[1,2,6*]

Abstract

Background: Long non-coding RNAs (lncRNAs) have emerged as a class of factors that are important for regulating development and cancer. Computational prediction of lncRNAs from ultra-deep RNA sequencing has been successful in identifying candidate lncRNAs. However, the complexity of handling and integrating different types of genomics data poses significant challenges to experimental laboratories that lack extensive genomics expertise.

Result: To address this issue, we have developed lncRNA-screen, a comprehensive pipeline for computationally screening putative lncRNA transcripts over large multimodal datasets. The main objective of this work is to facilitate the computational discovery of lncRNA candidates to be further examined by functional experiments. lncRNA-screen provides a fully automated easy-to-run pipeline which performs data download, RNA-seq alignment, assembly, quality assessment, transcript filtration, novel lncRNA identification, coding potential estimation, expression level quantification, histone mark enrichment profile integration, differential expression analysis, annotation with other type of segmented data (CNVs, SNPs, Hi-C, etc.) and visualization. Importantly, lncRNA-screen generates an interactive report summarizing all interesting lncRNA features including genome browser snapshots and lncRNA-mRNA interactions based on Hi-C data.

Conclusion: lncRNA-screen provides a comprehensive solution for lncRNA discovery and an intuitive interactive report for identifying promising lncRNA candidates. lncRNA-screen is available as open-source software on GitHub.

Keywords: lncRNA, Comprehensive pipeline, Data integration, Fully automated, Interactive report

Background

The landscape of transcription in organisms is now known to be complex and pervasive, producing a wide range of small and long RNA species with a variety of biological functions discovered so far [1–3]. One of the least characterized yet largest class of RNA species are the long non-coding RNAs (lncRNAs). The most basic definition of a lncRNA is a long RNA, at least 200 base pairs in length, that does not encode protein. They can be further classified by features such as their genomic location, structure, and expression [4]. Of the small number of lncRNAs that have been characterized, they have been shown to be functionally important for chromatin and other cellular processes that affect organismal development and cancer [5, 6]. However, there are 28,031 lncRNA transcripts annotated to date in GENCODEv19 and 91,000 in MiTranscriptome, and this number will continue to grow as deeper and more sensitive RNA sequencing data are generated. Recently, large-scale lncRNA analyses of published data (e.g. TCGA) have been conducted [7, 8], however the authors have not made their pipelines available. A number of databases and bioinformatics tools have been developed to annotate and catalog lncRNAs that are known or novel [9]. LncRNA2Function [10], LNCipedia [11], lncRNAdb [12] and lncRNAtor [13] provide comprehensive databases for known, annotated lncRNAs. iSeeRNA [14], CPC [15] and CPAT [16] introduced machine learning-based approaches only focusing on assessment of the coding probability of potential lncRNAs. lncRScan [17] and its new version lncRScan-SVM [18] are pipelines which provide novel multi-exonic lncRNA only discovery from RNA sequencing (RNA-seq), lacking the ability to integrate

* Correspondence: ioannis.aifantis@nyumc.org; aristotelis.tsirigos@nyumc.org
[1] Department of Pathology and Laura and Isaac Perlmutter Cancer Center, New York University School of Medicine, New York, NY 10016, USA
Full list of author information is available at the end of the article

other data types to further filter for interesting lncRNA candidates. It is known that lncRNAs, like coding genes, are enriched for histones that mark transcriptionally active sites such as histone H3K4me3 and H3K27ac [19, 20]. This approach has led to the successful identification of LUNAR1 and its function in regulating the IGF1R locus to sustain T-cell leukemia growth [21]. Thus integrating other genomic features allows for identification of lncRNAs with important biological functions. One of the fundamental issues in the field is to identify the function of the discovered lncRNAs. In this context, the ability to integrate a variety of genomic datasets to increase the probability of identifying lncRNAs that are functionally relevant will be of great value. Thus having an extensive bioinformatics pipeline that can quickly annotate and classify lncRNAs from RNA sequencing (RNA-seq) data will be valuable for identifying strong candidates for biological validation. To this end, we have developed an extensive computational pipeline to integrate different types of experimental data to annotate lncRNAs, which can be filtered by the user to identify specific lncRNAs of interest. The pipeline first aligns and assembles RNA-seq data to build a comprehensive transcriptome assembly for all the samples. Then, using a series of filtering criteria based on gene annotations, sequence length, expression level, coding potential and other features, a list of putative lncRNA candidates is defined containing basic information that includes transcript size, genomic location, and differential gene expression. Afterwards, depending on the raw data that the user may provide, our pipeline can process and annotate the putative lncRNAs with other processed information such as ChIP-seq data, copy number variation, Hi-C interaction etc. Gene tracks for UCSC genome browser and lncRNA local genomic snapshots are generated simultaneously in order to quickly visualize and assess each lncRNA by its features. The output of the pipeline is a comprehensive table and an interactive HTML report containing all the putative lncRNAs with their corresponding genomic features that the user has added into the pipeline. This report can then be filtered interactively by the user for specific lncRNAs of interest based on any combination of the genomic features included in the analysis.

Implementation

The lncRNA-screen workflow

lncRNA-screen is an extensive analysis pipeline providing various useful functions for lncRNA annotation and candidate selection. It encompasses multiple automated processes designed for lncRNA discovery and computational selection, including public data download, locally sequenced data integration (RNA-seq and ChIP-seq datasets), comprehensive transcriptome assembly, coding potential estimation, expression level quantification, differential expression comparisons and analysis, and lncRNA classification. Our pipeline enables fully customizable lncRNA discovery with insightful visualization embedded in each step of data processing. Most importantly, lncRNA-screen automatically generates an interactive lncRNA feature report that allows the user to conveniently search, filter, and rank by important features (e.g. expression level, presence of histone marks, etc.) extracted from the different input data types. Additionally, it provides a genome snapshot of each lncRNA locus to help user visually assess the relevance and quality of each candidate lncRNA. The main functionality of the lncRNA-screen pipeline compared to other published lncRNA analysis tools is summarized in Table 1. According to the table, lncRNA-screen provides the most extensive computational lncRNA discovery pipeline to date. The lncRNA-screen workflow can be divided into two

Table 1 The function comparison between lncRNA-screen and other publically available software

	lncRNA-screen	coRAL	RNA-CODE	lncRScan	iSeeRNA	CIRI	Annocript	LncRNA2Function
Stand-alone server	√	√	√	√	√	√	√	
Parallel computing	√							
Online server					√			√
Raw data support	√		√	√	√	√	√	√
Processed data support	√					√		
TCGA, GEO, SRA data automatic download	√							
Known lncRNA identification	√	√	√					
Novel lncRNA identification	√	√		√	√			
Differential expression analysis	√	√	√	√	√	√		
ChIP-seq histone mark integration	√							
Hi-C data integration	√							
lncRNA genome snapshots	√							
Interactive lncRNA report	√							

main phases (Fig. 1). In Phase I, the pipeline first conducts RNA-seq alignment and assembly individually for each sample and then merges them in order to construct a comprehensive transcriptome assembly. Next, a series of filtration steps are applied in order to detect lncRNAs based on reference annotations and genomic location. A putative lncRNA candidate list is generated in Phase I for further classification in Phase II using different data types. In Phase II, expression levels of the putative lncRNAs are quantified in all the samples and summarized at multiple user-defined group levels. Moreover, by integrating with ChIP-seq data, we identify typical histone mark profiles (H3K4me3, H3K27ac) around the TSS region of all lncRNAs. Pairwise differential expression analysis can also be performed between any two groups. At the same time, lncRNA-screen annotates all putative lncRNAs using the information obtained from the analysis and supplemented by additional types of "segmented" data (CNVs, SNPs, Hi-C, etc.) in order to generate a comprehensive interactive lncRNA feature report. The user can easily adjust parameters (see *README.md* file on GitHub repository) to find the most confident and functionally relevant lncRNAs candidates for further experimental validation. Furthermore, lncRNA-screen automatically produces different summary plots and a flowchart providing intuitive guidance to the user for adjusting the parameters. Finally, for every lncRNA, lncRNA-screen generates a genome snapshot which shows the gene structure, expression pattern and histone mark profile for the neighboring area. Additionally, if processed Hi-C data is provided, a local Hi-C interaction snapshot revealing potential looping events between each lncRNA and its neighboring genes is also generated.

Obtain the software and environment setup

lncRNA-screen is open-source software, free for academic use and available for download from GitHub repository: https://github.com/NYU-BFX/lncRNA-screen. Detailed instructions can be found in Additional file 1: Supplementary Material 2. It depends on a pipeline that

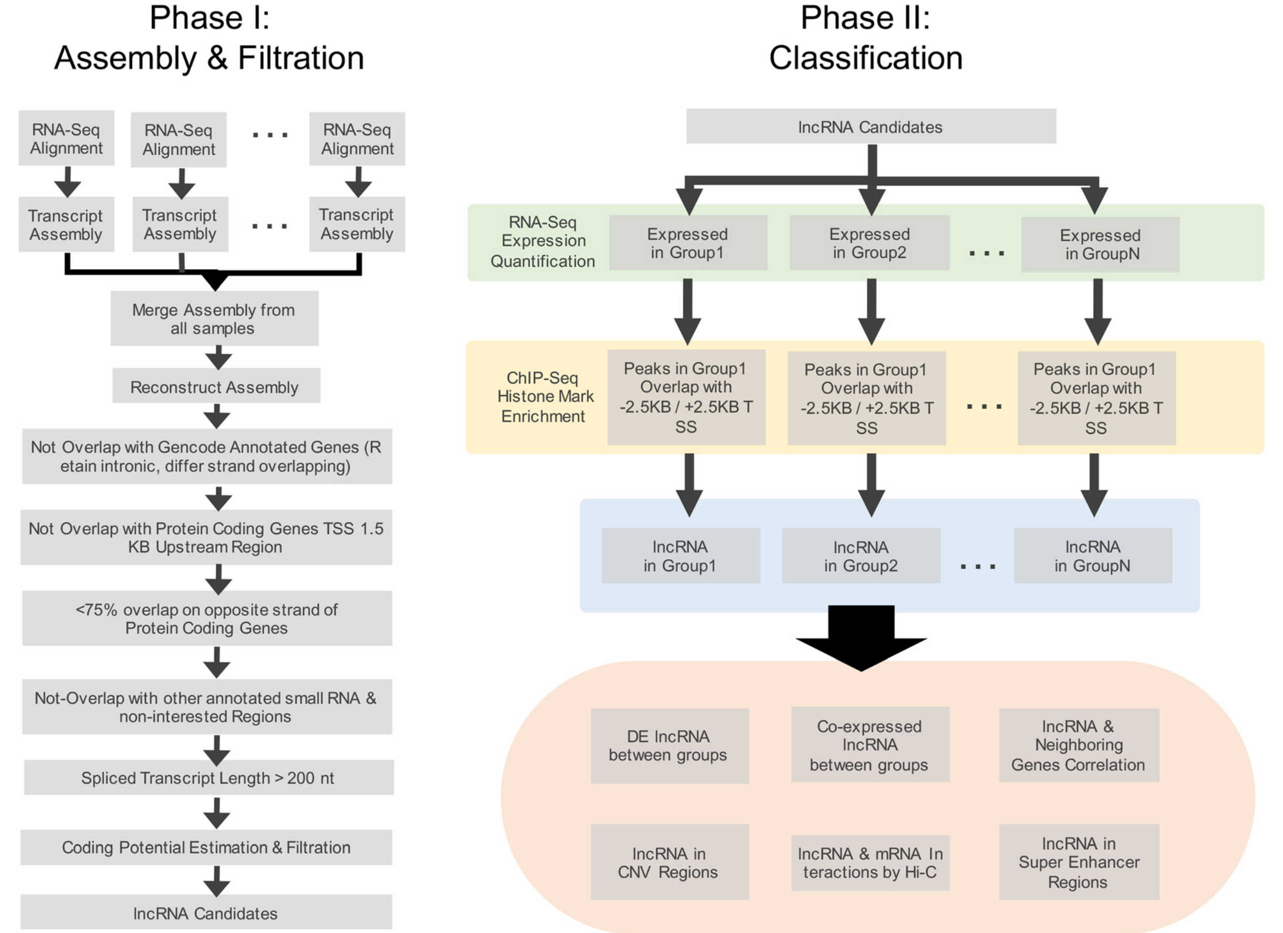

Phase I:
Assembly & Filtration
RNA-Seq Alignment
RNA-Seq Alignment
. . .
RNA-Seq Alignment
Transcript Assembly
Transcript Assembly
. . .
Transcript Assembly
Merge Assembly from all samples
Reconstruct Assembly
Not Overlap with Gencode Annotated Genes (Retain intronic, differ strand overlapping)
Not Overlap with Protein Coding Genes TSS 1.5 KB Upstream Region
<75% overlap on opposite strand of Protein Coding Genes
Not-Overlap with other annotated small RNA & non-interested Regions
Spliced Transcript Length > 200 nt
Coding Potential Estimation & Filtration
lncRNA Candidates
Phase II:
Classification
lncRNA Candidates
RNA-Seq Expression Quantification
Expressed in Group1
Expressed in Group2
. . .
Expressed in GroupN
ChIP-Seq Histone Mark Enrichment
Peaks in Group1 Overlap with -2.5KB / +2.5KB TSS
Peaks in Group1 Overlap with -2.5KB / +2.5KB TSS
. . .
Peaks in Group1 Overlap with -2.5KB / +2.5KB TSS
lncRNA in Group1
lncRNA in Group2
. . .
lncRNA in GroupN
DE lncRNA between groups
Co-expressed lncRNA between groups
lncRNA & Neighboring Genes Correlation
lncRNA in CNV Regions
lncRNA & mRNA Interactions by Hi-C
lncRNA in Super Enhancer Regions

Fig. 1 The workflow of lncRNA-screen. Phase I conducts RNA-seq alignment and transcriptome assembly, performs transcript filtration and generates a putative lncRNA list. Phase II uses RNA-seq expression quantification, ChIP-seq histone marks and user-defined annotations (CNVs, Hi-C, etc.) to classify lncRNA into different groups

performs standard RNA-Seq analysis, also available on GitHub: https://github.com/NYU-BFX/RNA-Seq_Standard (instructions are provided in Additional file 2: Supplementary Material 3). Most of the dependencies used in this pipeline are integrated into our software packages and do not need to be re-installed or re-complied in Linux systems. All the R packages used will be downloaded automatically as the pipeline runs. A detailed description of all the dependencies and environment setup can be found in Additional file 3: Supplementary Material 4. Users can easily follow the instructions and we are continuously providing support of any questions regarding the lncRNA-screen pipeline installation.

Input data, sample sheet and group information setup

lncRNA-screen provides fully automated and parallel download of raw RNA-seq FASTQ files from different public data repositories including the Sequence Read Archive (SRA) from the National Center for Biotechnology Information (NCBI) and The Cancer Genome Atlas (TCGA). The user only needs to provide a list of SRA accession numbers or TCGA UUIDs matched with user-defined sample names to be used throughout the analysis in a sample sheet file. The pipeline automatically downloads the relevant files, while the various tools within our pipeline automatically identify them as the appropriate inputs for downstream analyses. Processed ChIP-seq histone mark data (H3K4me3, H3K27ac and H3K4me1) is required in BED4 format where the fourth column corresponds to the ChIP enrichment score, calculated by MACS2 or any other peak calling tool. Any type of data that can be represented as segmented data in BED4 format (e.g. CNV or SNP data) is also supported as input to lncRNA-screen. For example, copy-number variation segments can be integrated by lncRNA-screen to identify lncRNAs located in recurrently amplified or deleted regions in cancer samples. A group information sheet is also required, where rows correspond to sample names and different grouping strategies can be designated by adding an arbitrary number of columns. These groups might be different cell types, experimental conditions etc. Matched histone mark data for each group are also assigned using this file. lncRNA-screen will automatically perform the analysis for all user-defined grouping strategies.

Read quality assessment for RNA-seq data

FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) is a commonly used package for assessing the Next-generation sequencing read quality. lncRNA-screen utilizes FastQC as an automatic quality control (QC) procedure for each sample. It reports the distribution of average per-base and per-sequence quality, per-base and per-sequence GC content, sequence length distribution, sequence duplication level, etc. This information allows the user to quickly diagnose irregularities in their input samples and take appropriate action. Based on the results, the user can decide whether to perform pre-processing of the FASTQ files such as trimming during the next step.

Read alignment

Next, all sequences are aligned using the Spliced Transcripts Alignment to a Reference (STAR) software [22]. A pre-built STAR genome index is required in advance. For each sample, the pipeline identifies all the raw reads belonging to each sample, automatically determines whether they are single-end or paired-end and properly groups read pairs as well as different sequence files originating from different sequencing lanes. Then, STAR will automatically determine the strand specificity and read length. This is a significant advantage compared to TopHat2 [23] and other aligners, given that it is oftentimes a challenging task to retrieve this information from large public datasets such as TCGA or SRA. Determining this automatically using STAR is also useful in subsequent steps, for example during transcriptome assembly performed by Cufflinks. Trimming, soft clipping and other type of necessary pre-preprocessing of the reads can be done by STAR during the alignment process. Only uniquely mapped reads are retained in the final alignment. Raw read counts are generated at the same time by the STAR aligner for all the annotated features in the GENCODEv19 GTF as default annotation file (or other user supplied annotation files) provided when aligning the samples.

Post-alignment assessment and processing

After alignment, Picard-Tools (http://broadinstitute.github.io/picard) are used to assess the duplication rate and remove duplicate reads if necessary. We then generate read coverage signal track files in BIGWIG format with adjustable resolution compatible with IGV and UCSC genome browsers. The pipeline also provides a function which can merge and generate combined track files at the group level. An interactive HTML RNA-seq analysis report is generated automatically which incorporates an alignment quality report (example shown in Fig. 2; see Results for details) allowing the user to quickly inspect the alignment rates and the number of usable reads. A sample distance plot (example shown in Fig. 3; see Results for details) represents an unbiased clustering of the samples based on genes annotated by GENCODEv19. The pipeline also performs differential expression analysis for each pair of groups and reports all the GENCODEv19 annotated genes that pass a user-defined significance threshold if the number of groups provided by the user is sufficiently small (up to 10). The

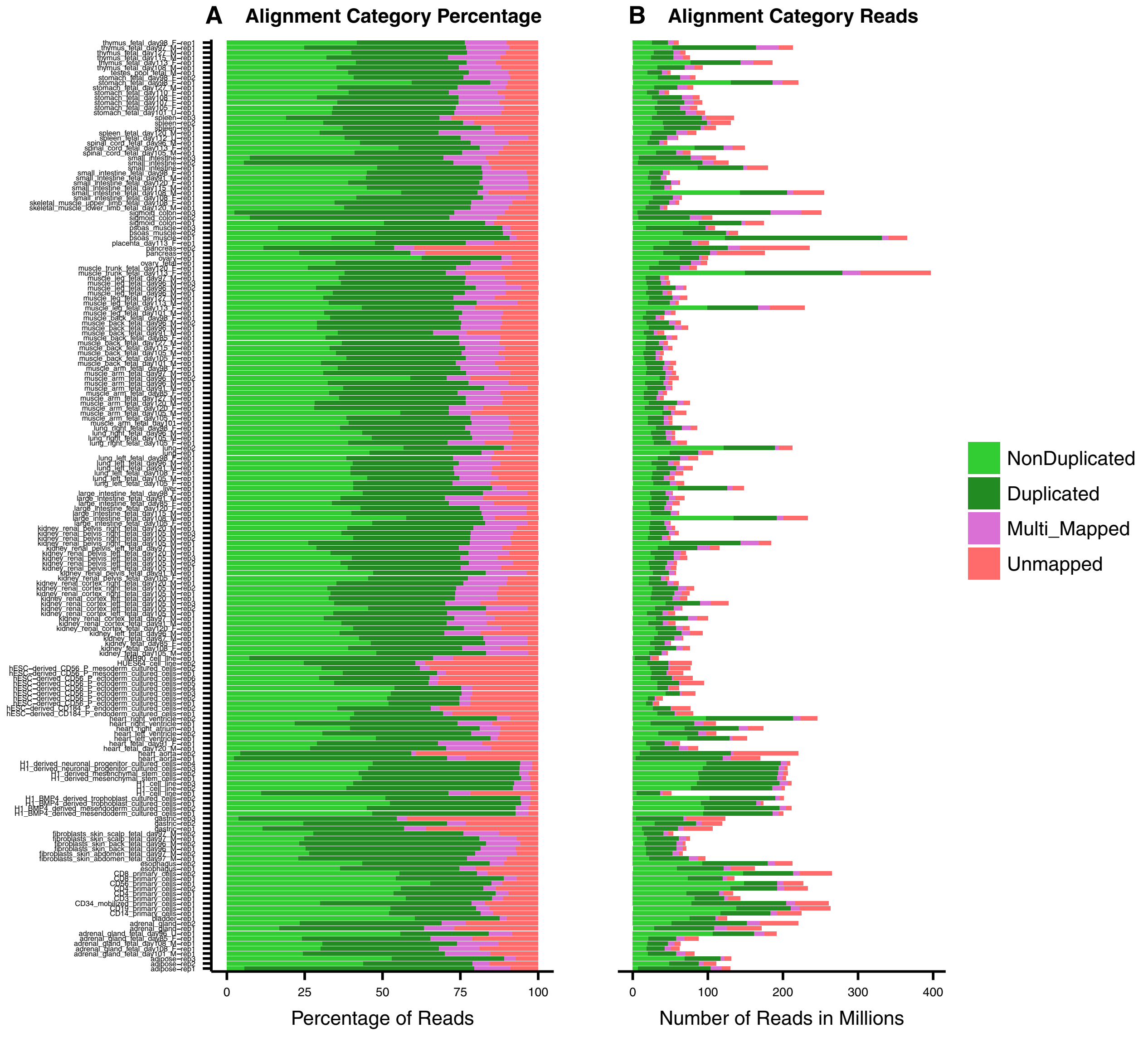


Fig. 2 RNA-seq alignment quality plot: (**a**) read percentages, (**b**) read counts

user can perform differential expression analysis semi-automatically for specific comparison groups if number of groups exceeded 10. This interactive RNA-seq report is designed for providing an overview of the sample differences and similarities between and within groups and for verifying that the user-defined list of genes follow the expected expression pattern across different groups.

Transcriptome assembly construction

Aligned sequences are assembled individually by Cufflinks 2.2.1 [24] using GENCODEv19 annotation as a guide transcriptome with default parameters. All ribosomal RNAs and snRNAs are masked. The strand specificity is automatically determined by the pipeline based on the result of STAR aligner. Finally, we use Cuffmerge to merge all the assemblies into a single comprehensive assembly from all assembled transcripts with GENCODEv19 annotation as a guided assembly.

Comprehensive identification of putative lncRNAs

To distinguish known transcripts from novel transcripts, we use the Cuffcompare result generated during Cuffmerge process, which compares the comprehensive assembly from all assembled samples and GENCODEv19 reference annotation. Based on the Cuffcompare classification result, all of the transcripts are categorized into 12 different classes (see http://cole-trapnell-lab.github.io/cufflinks/cuffcompare/#transfrag-class-codes). The user can define which categories to keep. By default, we keep the three following categories: a transfrag falling entirely

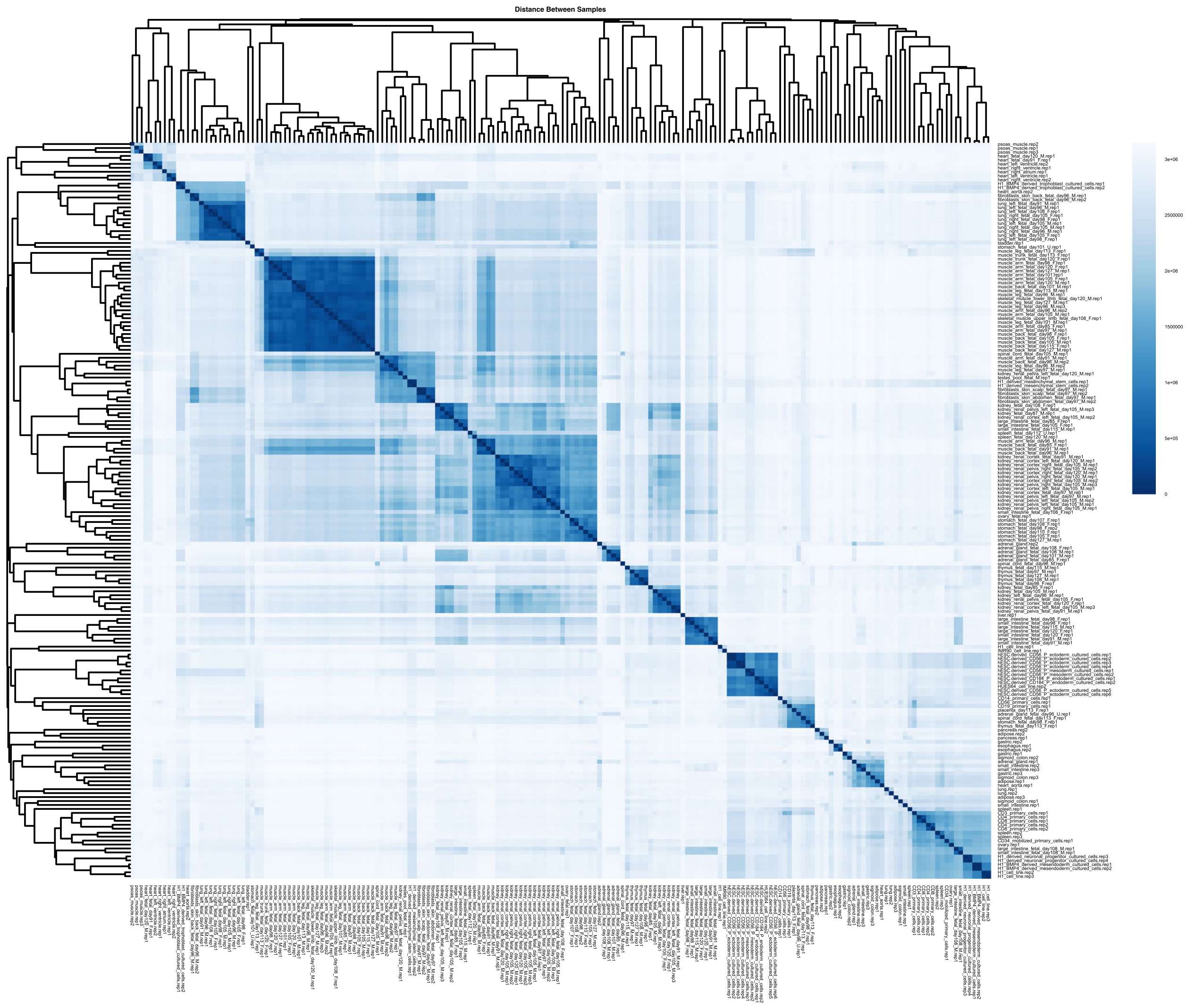


Fig. 3 Sample distance report shows unbiased clustering of the samples based on the ENSEMBL annotated genes

within a reference intron (class code "i"); unknown, intergenic transcript (class code "u"); an intron of the transfrag overlaps a reference intron on the opposite strand; exonic overlap with reference on the opposite strand (class code "x"). Annotated lncRNAs by GENCODEv19 will be added back to the filtered transcripts, forming a comprehensive annotated and novel lncRNA assembly. Then this lncRNA assembly is merged at gene level based on the unique gene ID given by Cuffmerge. Alternatively, spliced transcripts for each gene can be retrieved in the future if necessary by searching the entire transcript annotation by the Gene ID. Then, for each protein-coding gene in GENCODEv19, we extract its transcription start site (TSS) and extend it by 1.5 KB upstream and downstream (default value, modifiable by the user) to include potential alternative transcription start sites of protein-coding genes. Any putative lncRNA overlapping with these regions on the same strand will be excluded from the putative lncRNA list considering that alternative TSSs of the protein-coding genes may appear as novel transcripts. Annotated microRNAs, snRNAs, srpRNAs, tRNAs, scRNAs and antigen receptors datasets were obtained from UCSC and ENSEMBL databases. The last step of filtration is to exclude transcripts less than 200 nt (default value, modifiable by the use) in length, based on the lncRNA definition. This subset of putative lncRNA is considered to be the comprehensive putative lncRNA assembly and is used in all downstream analyses. Moreover, we annotate all the remaining putative lncRNAs into different categories based on their overlaps with RefSeq [25, 26], ENSEMBL and MiTranscriptome.

Estimation of lncRNAs abundances

We use featureCounts [27] to count the raw reads for all the genes included in the putative lncRNA assembly and

GENCODEv19 annotations. Strand-specificity and counting single-end versus paired-end reads is determined by the pipeline automatically in the previous step. The read abundance calculation is performed at the gene level and all the reads included are uniquely mapped. FPKM values are calculated accordingly. A summary bar chart is automatically generated after featureCounts is performed (example in Additional file 4: Figure S1), showing the number and percentage of reads/fragments that have been utilized by the assembly. Problematic samples can be identified in this step and should be excluded from the study if the percentage of reads/fragments assigned to the assembly is too low.

Coding potential estimation

We use Coding Potential Assessment Tool (CPAT) to estimate the coding potential of the filtered putative lncRNA. Using the pre-trained human (hg19) logistic regression model, CPAT reports putative ORF size and coding probability for each transcript. The optimum cutoff for human gene annotation is 0.364 as determined in the CPAT manuscript. The distribution of ORF sizes and coding potential scores for protein-coding and non-coding transcripts is produced automatically (example shown in Fig. 4; see results for details).

Integration of histone mark ChIP-seq data

ChIP-seq peaks for H3K4me3, H3K27ac and H3K4me1 can be used directly as input to lncRNA-screen. Alternatively, lncRNA-screen can perform its own ChIP-seq analysis starting from raw FASTQ files. Using Botwie2 [28] for alignment and MACS2 [29] for peak calling, histone mark peaks can be identified. Broad peak calling is used (q-value <0.05) and fold enrichment compared to the input is calculated. A user-defined fold change cutoff can be applied. Then we extend each putative lncRNA transcription start site by 1.5 KB upstream and downstream and assign the histone marks that have a peak overlapping this extended TSS region. The fold enrichment value of each overlapping peak is reported for each lncRNA.

Defining group-enriched lncRNAs

In order to determine all expressed lncRNAs in a specific group, the pipeline allows the user to set a FPKM cutoff based on the distribution of the lncRNA expression level. Then, we compute the average FPKM values for all the putative lncRNAs among different sample groups. Samples which have FPKM value below the cutoff are excluded before computing average FPKM values within groups. The number of samples that have an FPKM value above the cutoff will also be reported. Next, combining the ChIP-seq histone mark overlaps with the FPKM cutoff, we are able to define group-enriched expressed lncRNAs based on their expression value and the histone mark enrichment in the extended TSS regions. The user can specify which histone mark (or combination of histone marks) must be present. By default, we define a lncRNA as being expressed in a specific group by requiring the mean FPKM in this group to be greater than a user-defined cutoff (default 0.5) while overlapping with H3K4me3 histone marks in its extended TSS region. Additionally, we define an enhancer-RNA to be expressed in a specific group, if its mean FPKM in this group exceeds the cutoff value while overlapping with both H3K4me1 and H3K27ac but not with H3K4me3 histone marks in its extended TSS region. A "pie matrix" illustrating the number of lncRNAs in categories characterized by different histone marks, as well as the intersections between different sample groups is automatically generated (example shown in Fig. 5). The group information table, as described in a previous section, is essential for the pipeline to group the related samples or replicates under the same group name and to

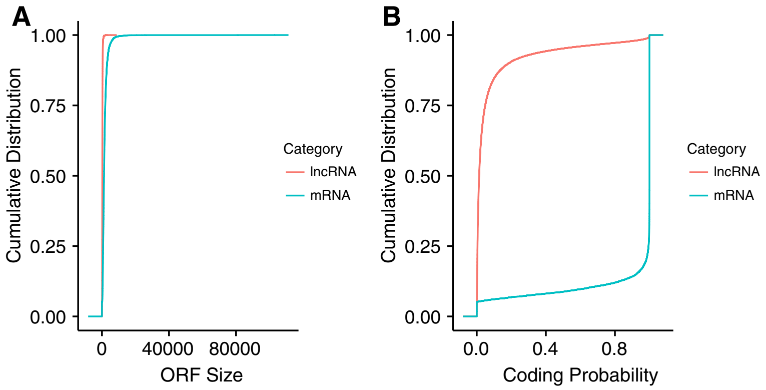


Fig. 4 Coding Potential Statistics. **a** Distribution of ORF size for novel lncRNAs, annotated lncRNAs and protein coding genes. **b** Distribution of coding potential score calculated by CPAT for novel lncRNAs, annotated lncRNA and protein coding genes

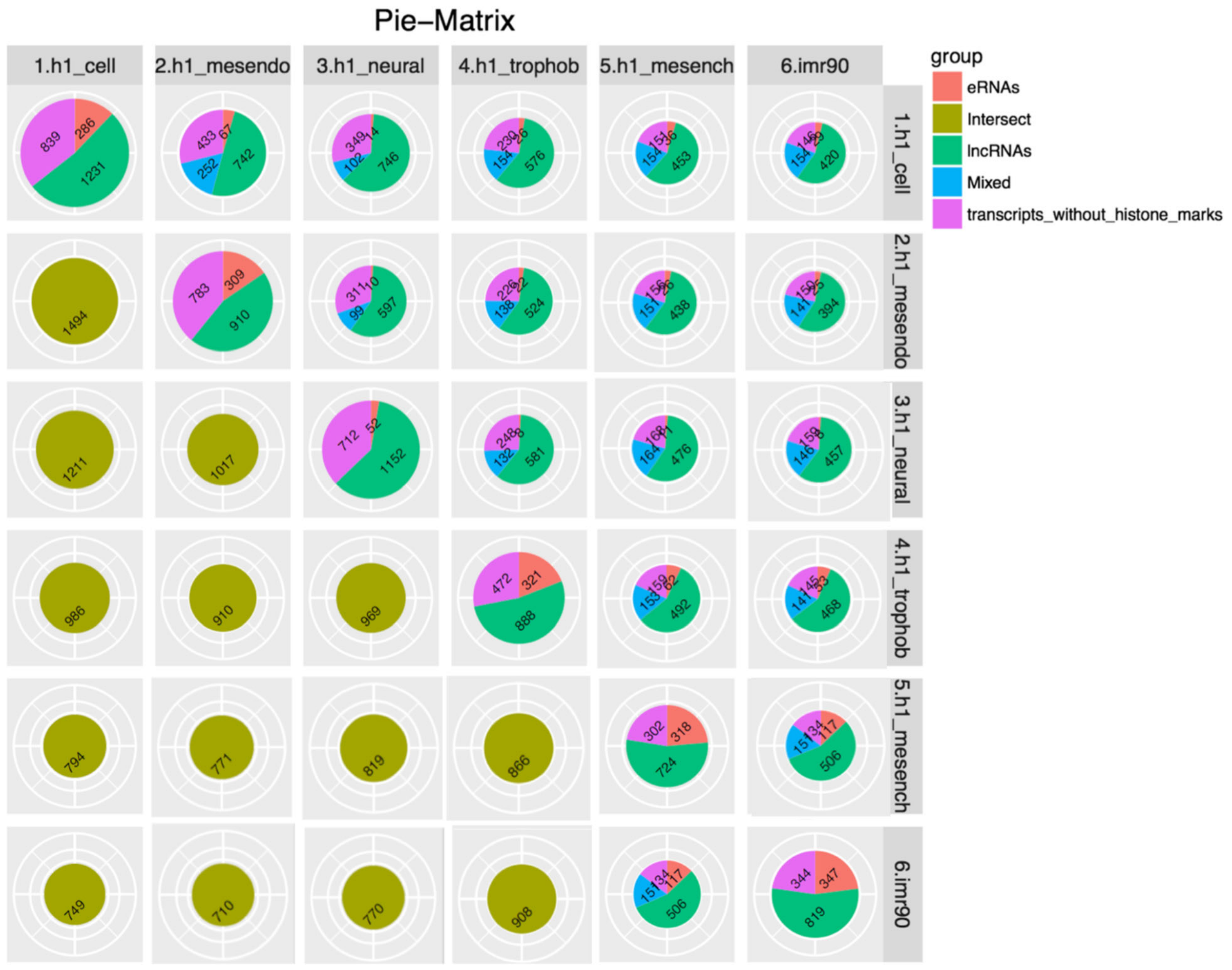


Fig. 5 Pie-matrix showing the number of lncRNAs in categories defined by different histone mark enrichment profiles and pairwise overlap of lncRNAs between different groups in different categories. The pie charts in the diagonal from top left to bottom right show total number of lncRNAs in each group and categorize based on the histone mark enrichment profile. Other pie charts above this diagonal show the overlap of lncRNAs between the column and row names of the groups in different histone mark enrichment categories. Remaining pie charts below the diagonal show only the total number of overlapping lncRNAs between the column and row names of the groups. The sizes of the pie charts represent the total number of lncRNAs in each group

match the RNA-seq data and ChIP-seq data. Different levels of grouping can be achieved by adding an additional group column so that the user can explore the similarity and difference between samples in different ways.

Pairwise differential expression analysis between designated groups

Our pipeline also integrates pairwise differential expression analysis using DESeq2 [30]. If a list of pairwise comparison groups of interest is provided, the pipeline performs all the differential expression analyses provided in the list. Otherwise, by default the program performs differential expression for up to 10 comparison groups. *P*-values, FDR, and log2 fold changes are provided for each lncRNA.

lncRNA selection and the comprehensive putative lncRNA feature report

The default criteria of lncRNA selection are illustrated in the example flowchart of Fig. 6. The flowchart is generated automatically and shows in detail the filtering criteria applied in each step as well as the number of putative lncRNA selected (and excluded). This allows the user to inspect the breakdown of the impact of each filtration step. The user can then modify the parameter file and re-run all the steps after the "expression estimation" step, thus obtaining an updated lncRNA list and the corresponding flowchart using their custom criteria.

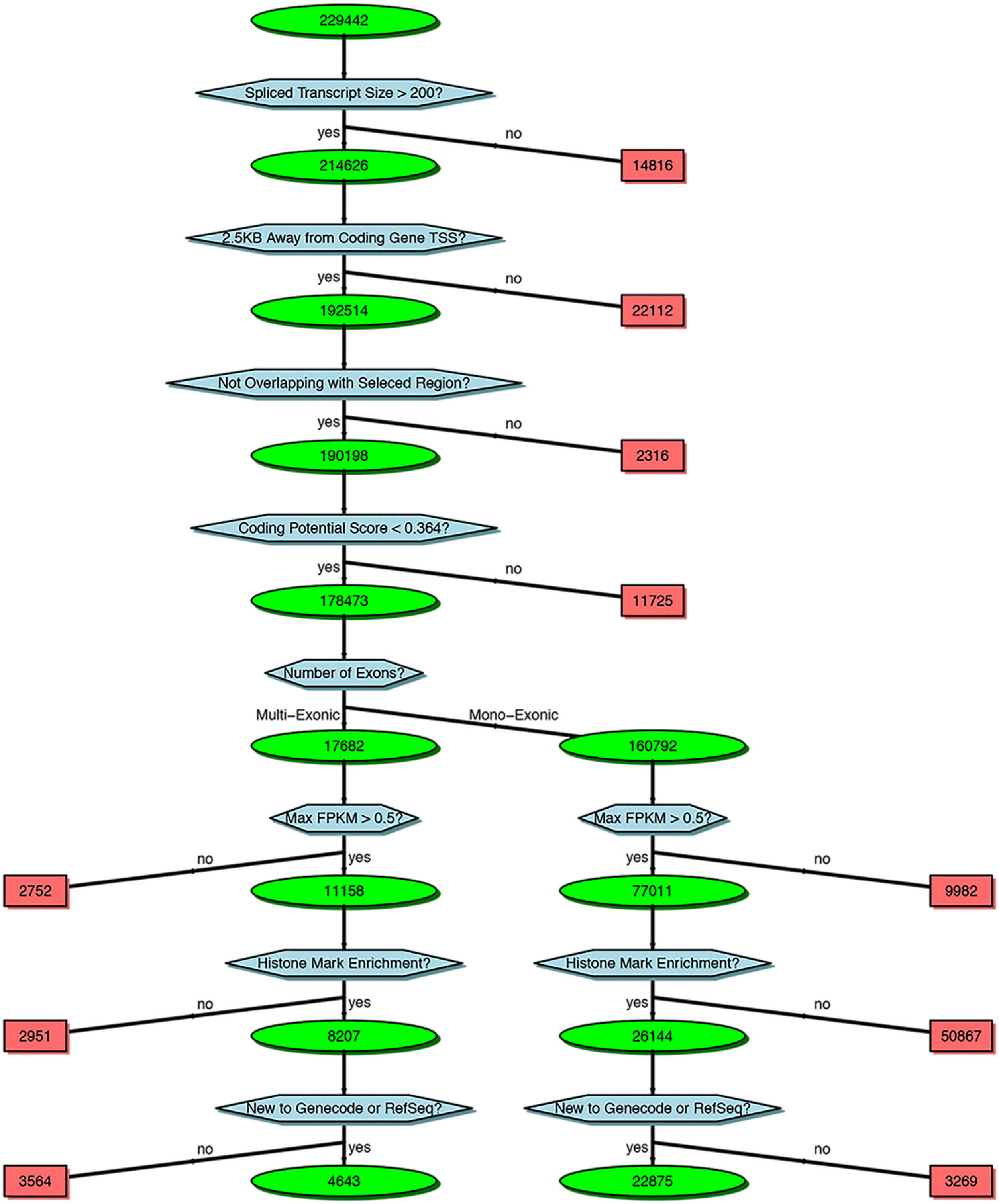
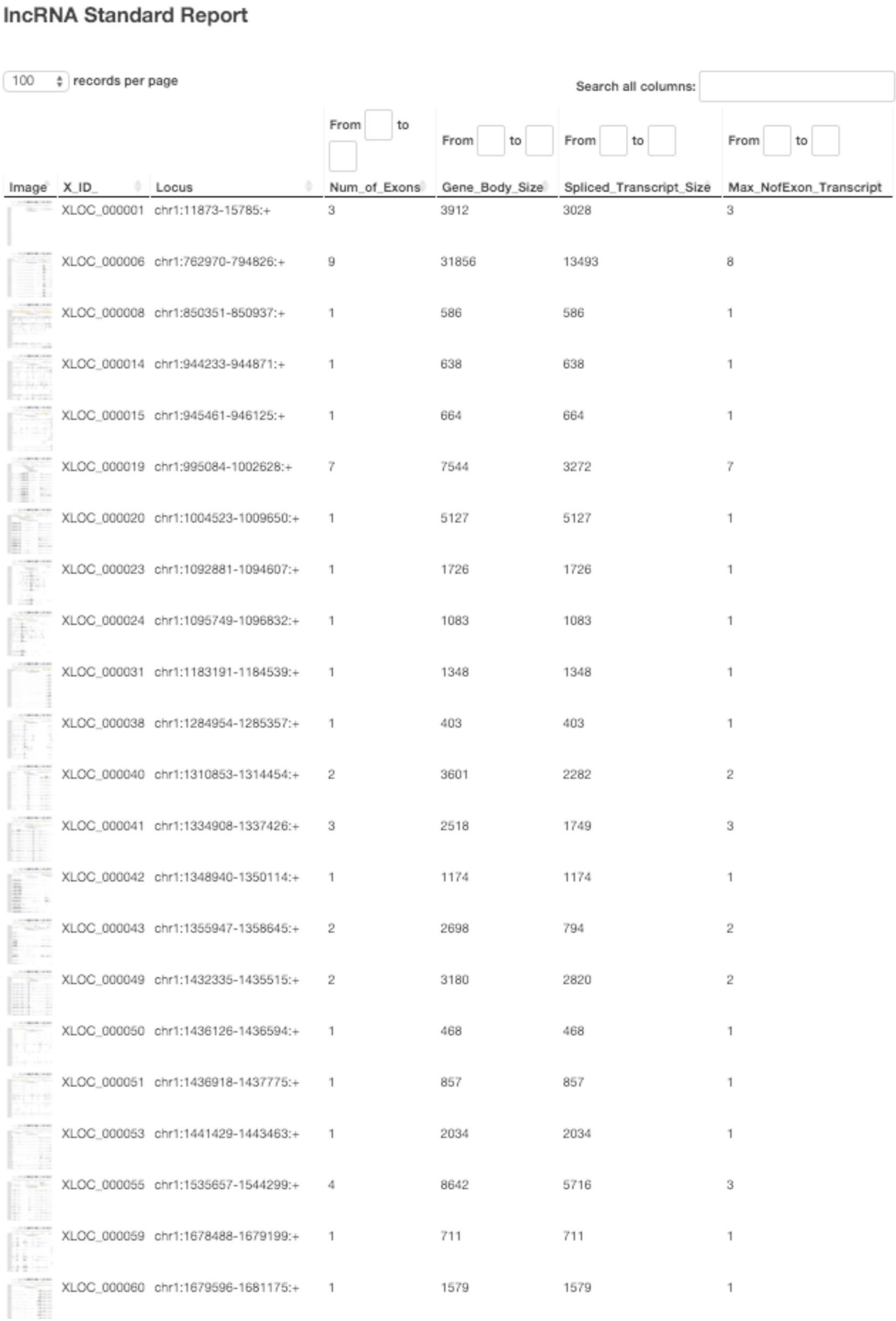


Fig. 6 Flowchart of lncRNA selection criteria and number of lncRNAs retained in each step

Finally, we collect all the results generated by our pipeline into a comprehensive putative lncRNA feature report which includes the columns shown in Table 3. The comprehensive putative lncRNA feature report is provided to the user in both HTML (example snapshot shown in Fig. 7) and Excel formats. Both versions allow



lncRNA Standard Report

100 records per page

Search all columns:

Image	X_ID_	Locus	From □ to □ Num_of_Exons	From □ to □ Gene_Body_Size	From □ to □ Spliced_Transcript_Size	From □ to □ Max_NofExon_Transcript
	XLOC_000001	chr1:11873-15785:+	3	3912	3028	3
	XLOC_000006	chr1:762970-794826:+	9	31856	13493	8
	XLOC_000008	chr1:850351-850937:+	1	586	586	1
	XLOC_000014	chr1:944233-944871:+	1	638	638	1
	XLOC_000015	chr1:945461-946125:+	1	664	664	1
	XLOC_000019	chr1:995084-1002628:+	7	7544	3272	7
	XLOC_000020	chr1:1004523-1009650:+	1	5127	5127	1
	XLOC_000023	chr1:1092881-1094607:+	1	1726	1726	1
	XLOC_000024	chr1:1095749-1096832:+	1	1083	1083	1
	XLOC_000031	chr1:1183191-1184539:+	1	1348	1348	1
	XLOC_000038	chr1:1284954-1285357:+	1	403	403	1
	XLOC_000040	chr1:1310853-1314454:+	2	3601	2282	2
	XLOC_000041	chr1:1334908-1337426:+	3	2518	1749	3
	XLOC_000042	chr1:1348940-1350114:+	1	1174	1174	1
	XLOC_000043	chr1:1355947-1358645:+	2	2698	794	2
	XLOC_000049	chr1:1432335-1435515:+	2	3180	2820	2
	XLOC_000050	chr1:1436126-1436594:+	1	468	468	1
	XLOC_000051	chr1:1436918-1437775:+	1	857	857	1
	XLOC_000053	chr1:1441429-1443463:+	1	2034	2034	1
	XLOC_000055	chr1:1535657-1544299:+	4	8642	5716	3
	XLOC_000059	chr1:1678488-1679199:+	1	711	711	1
	XLOC_000060	chr1:1679596-1681175:+	1	1579	1579	1

Fig. 7 A screenshot of the interactive HTML report generated by lncRNA-screen, which enables sorting, filtering, searching based on different lncRNA features and a preview of the lncRNA genome snapshot and local heatmap

the user to conveniently filter and search based on user-defined criteria.

lncRNA heatmaps

Two types of heatmaps are generated for selected lncRNAs which pass all the filtration criteria. First, we generated a supervised clustering heatmap (example shown in Fig. 8), designed to show the group-enriched lncRNAs. Because a lncRNA may be expressed in different groups at the same time, in this type of heatmap, the lncRNAs may appear multiple times. The order of the sample groups is the same as the order of the lncRNAs discovered in each group, which ensures that the group-enriched lncRNAs are always located near the diagonal of the heatmap. Second, we provide an unsupervised hierarchical clustering heatmap (example shown in Fig. 9) for samples (columns) as well as lncRNAs (rows). This heatmap lets the user inspect the

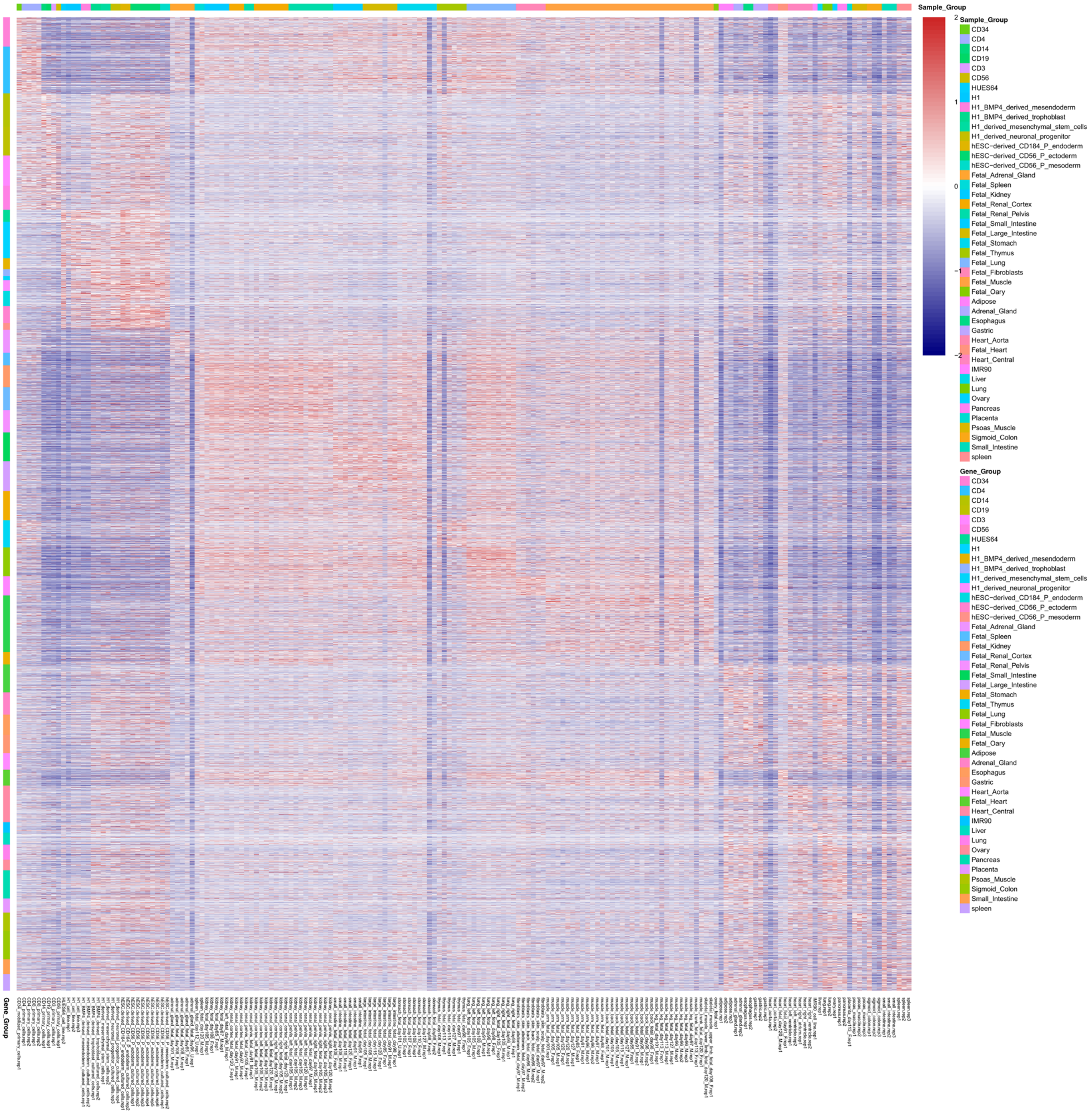


Fig. 8 Supervised lncRNA expression heatmap. It plots all group-enriched lncRNAs selected by user-defined filtration criteria. lncRNAs may appear multiple times due to co-expression in multiple groups. The order of the samples is the same as the order of the lncRNA discovered in each group

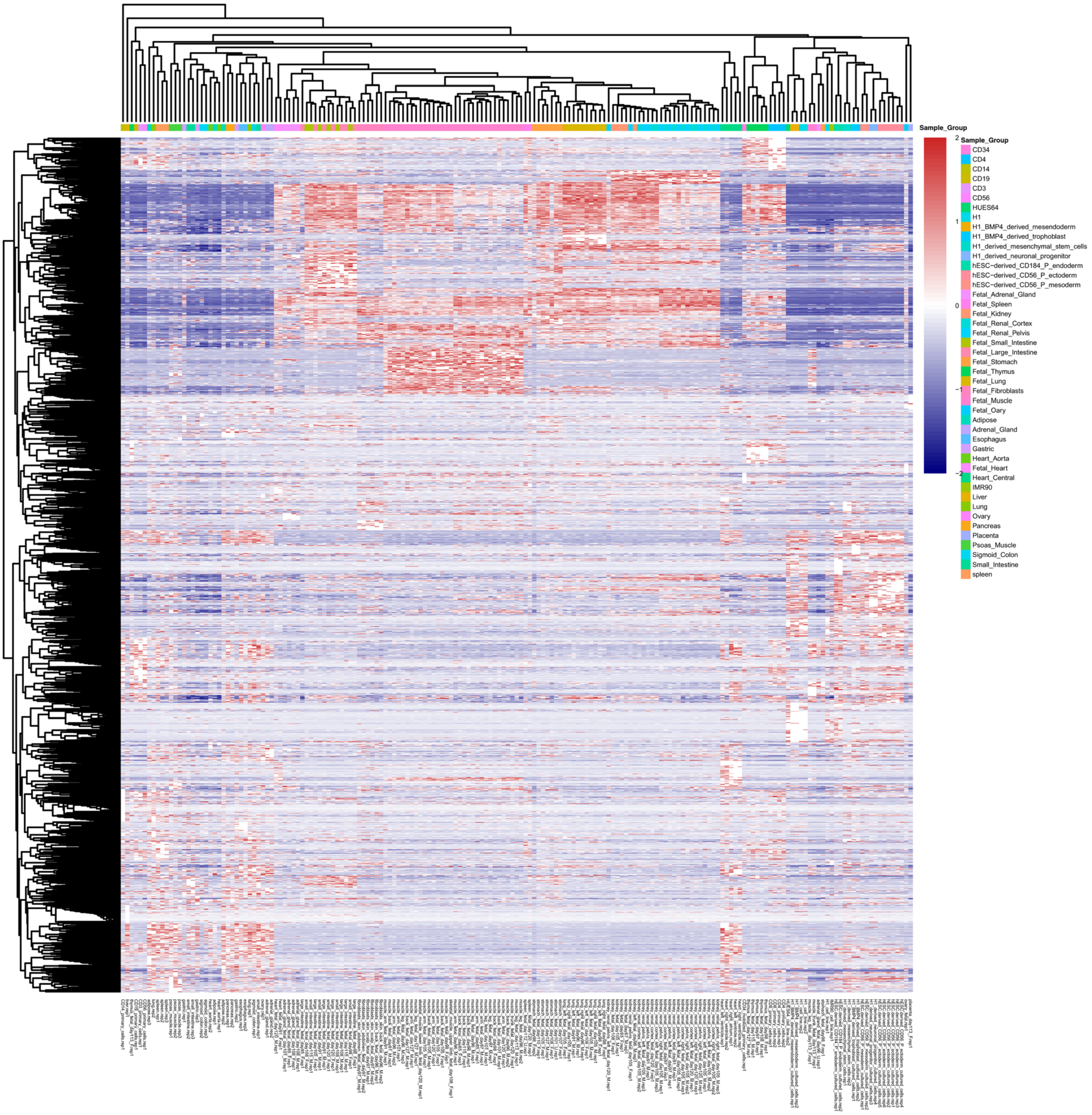


Fig. 9 Unsupervised lncRNA expression heatmap. Hierarchical clustering is performed in both rows and columns which categorized lncRNAs into different groups and identify sample similarity and difference

similarity and specificity between sample groups in the filtered lncRNA level, and also allows the user to search for lncRNAs co-expressed or co-differentially expressed between groups.

lncRNA genome snapshots

The pipeline generates a genome snapshot (example shown in Fig. 10) for each lncRNA, centered around the lncRNA locus and zooming out 10 times from the lncRNA length. In this snapshot, the user can choose to display ENSEMBL, RefSeq, GENCODEv19, or other user-defined GTF or BED format annotations. The comprehensive merged assembly is also shown in this snapshot, containing all the transcripts assembled without any filtration criteria steps. Moreover, the user can choose to plot bigwig RNA-seq signal tracks either at the merged group level or at the sample level. The HTML report (Fig. 7) also includes a preview of the lncRNA genome snapshot for all lncRNAs.

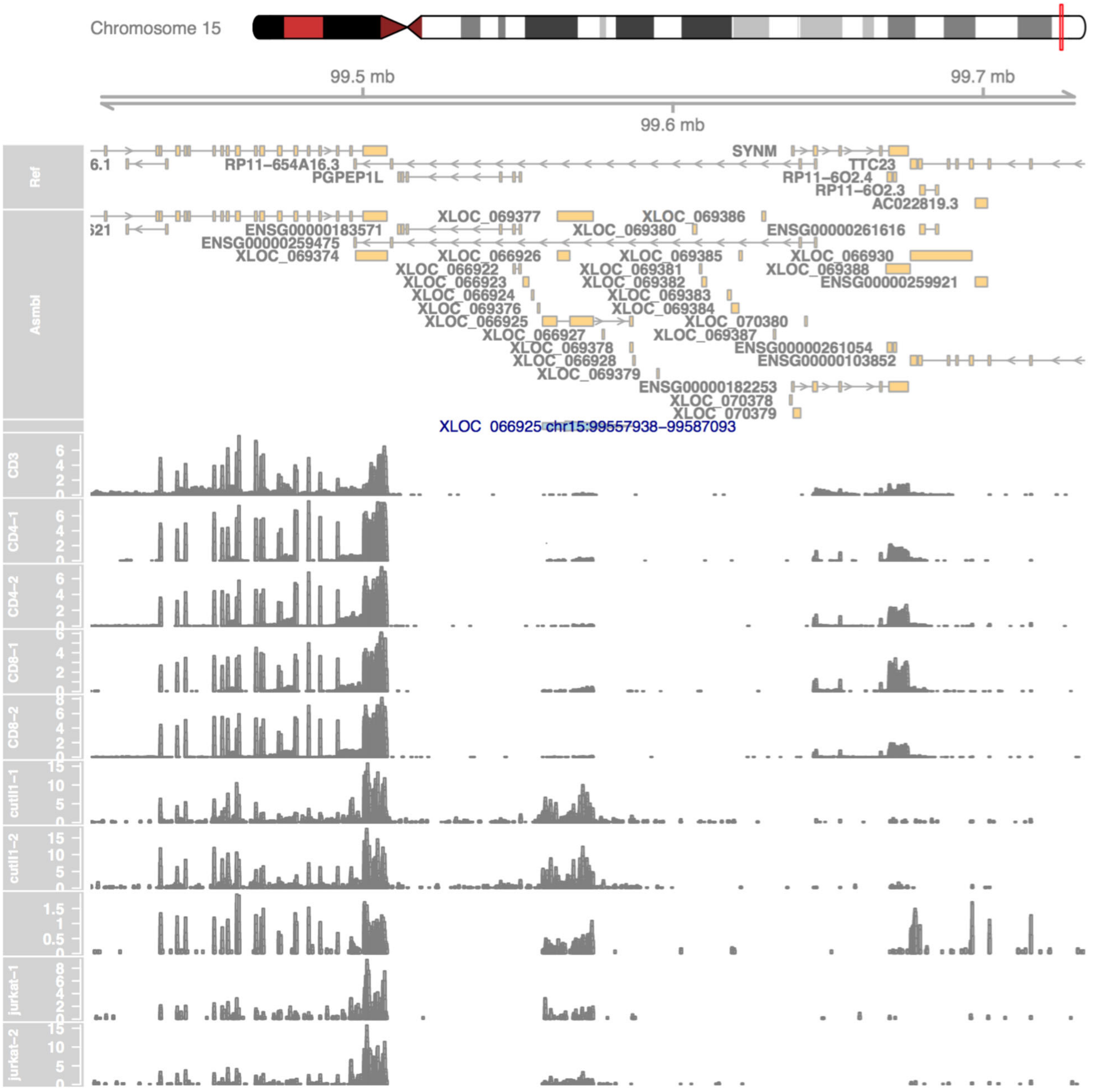


Fig. 10 lncRNA genome snapshot. Snapshot of the area surrounding each lncRNA, showing the RefSeq annotations, lncRNA-screen assembly and selected RNA-seq or ChIP-seq signal tracks

Results

Quick installation, setup, execution and interactive browsing of the results

lncRNA-screen can be downloaded as a single zip file from GitHub through the link below. The reference files setup script step will automatically download the necessary references and dependencies. Following instructions in the "how to run" link below, the user can easily set up the preferred parameters and run the entire pipeline using a single command. After the run is completed, the user can interactively browse the results using the automatically generated HTML report (see example below) or import the automatically generated table into Excel to do more complex filtering.

Setup and excute: https://github.com/NYU-BFX/lncRNA-screen/blob/master/README.md

Browse the results: http://www.hpc.med.nyu.edu/~gongy05/lncRNA-screen/H1_Cells/lncRNA_report.html

lncRNAs in Roadmap Epigenomics

To demonstrate the usage and performance of lncRNA-screen in big datasets, we used data from the Roadmap Epigenomics project [31], which contains RNA-seq and

Table 2 List of features included in the final lncRNA feature report

Single Column	One column per group
lncRNA ID	Mean FPKM above user-defined threshold
Locus	Percentage of samples above FPKM threshold
Number of Exons	H3K27ac peak enrichment score
Gene Body Size	H3K4me3 peak enrichment score
ORF Size	H3K4me1 peak enrichment score
Coding Probability	Histone marks enrichment and FPKM cutoff combination
Gencode Annotation	Differential expression analysis between groups
RefSeq Annotation	Hi-C interaction between lncRNA and neighboring coding genes
Ensembl Annotation	
MiTranscriptome Annotation	
SNPs annotation	
Copy number gain/loss value	

ChIP-seq data representing a collection of human stem cells and tissue type. The raw FastQ reads of total 198 RNA-seq samples from Roadmap Epigenomics project were downloaded from SRA using SRA-toolkit and aligned to the GENCODEv19 reference genome by STAR (version 2.4.2a) with default parameters (see methods). After quality control (Fig. 2 and Additional file 4: Figure S1), 187 samples (see Additional file 4: Table S1) were successfully processed and classified into 40 groups based on cell type. Accepted reads for each sample were assembled individually using Cufflinks (version 2.2.1) providing the guide reference (RefSeq Flat Table GTF file). Cuffmerge was employed to merge all the assemblies into a comprehensive transcriptome assembly, yielding 491,218 transcripts, forming 229,442 genes in total. By comparing the merged comprehensive transcriptome assembly with GENCODEv19 annotated genes using Cuffcompare, transcripts are classified into different categories based on their structure compatibility of the GENCODEv19 reference annotation. All filtering steps in Phase I of lncRNA-screen, including the number of lncRNAs selected and discarded by various filtering criteria are shown in Fig. 6. The total number of putative lncRNAs identified are 178,473. Both our coding genes, annotated and novel lncRNA candidates were tested by CPAT and the coding potential distribution comparison are in Fig. 4. Novel lncRNA and annotated lncRNA discovered in this pipeline showed similar distribution in ORF and coding potential, and significantly different than coding genes. We use the recommended coding potential cutoff 0.364 for human genome, excluding 11,725 genes from our putative lncRNA list, which is 6% of our total candidates. We also included the ChIP-seq histone marks broad peak calling result from MACS2 (H3K4me3, H3K27ac and H3K4me1) from Roadmap Epigenomics project for all of the 40 groups and matched them with corresponding RNA-seq sample groups. The lncRNA feature report in both HTML (see the link in GitHub) and Excel format (see the link in GitHub) included all 178,473 putative lncRNAs and their features shown in Table 2. The lncRNA feature report (Fig. 9) not only enables sorting, filtering, searching for all lncRNA features extracted from

Table 3 Number of putative lncRNAs identified in each group

Group	Number of lncRNAs	Group	Number of lncRNAs
Adipose	8709	Fetal Thymus	5763
Adrenal Gland	3950	Gastric	3978
Bladder	695	H1 BMP4 derived mesendoderm	1711
CD14	3982	H1 BMP4 derived trophoblast	844
CD19	6188	H1 derived mesenchymal stem cells	443
CD34	3661	H1 derived neuronal progenitor	1067
CD3	6113	H1	8271
CD4	5232	Heart Aorta	17,092
CD56	9596	Heart Central	6126
Esophagus	3443	hESC-derived CD184 P endoderm	1367
Fetal Adrenal Gland	12,008	hESC-derived CD56 P ectoderm	2086
Fetal Fibroblasts	1568	hESC-derived CD56 P mesoderm	900
Fetal Heart	2081	HUES64	1257
Fetal Kidney	4429	IMR90	2152
Fetal Large Intestine	5824	Liver	785
Fetal Lung	4930	Lung	2359
Fetal Muscle	10,477	Ovary	1056
Fetal Oary	2472	Pancreas	5204
Fetal Renal Cortex	4578	Placenta	9030
Fetal Renal Pelvis	4621	Psoas Muscle	2223
Fetal Small Intestine	4508	Sigmoid Colon	18,827
Fetal Spinal Cord	8113	Small Intestine	4760
Fetal Spleen	2448	spleen	2953
Fetal Stomach	10,330	Testes Pool	1266

Table 4 Number of transcripts by category for each multi-lineage differentiated embryonic stem cells

Cell Type	enhancers	eRNA	lncRNA	mRNA
hESC	6323	286	1231	8605
Mesench	16,736	318	724	8660
Mesendo	2371	309	910	8396
Neupro	1487	52	1152	8652
Trophect	12,663	321	888	8766

different data resources, but also includes a preview of the lncRNA genome snapshot. Examples of high confident lncRNA local snapshot is shown in Fig. 10. The report also includes a pre-set UCSC Genome browser session link to provide advanced genome browse function. With default filtration parameters, pie-matrix (Fig. 5) is showing all the lncRNAs expressed (mean FPKM >0.5) in at least one group by categories defined by 3 histone mark overlaps in the extended TSS regions (TSS flanked by +/− 1.5 KB). It also shows pairwise overlap of lncRNAs between different groups in different categories. A total of 8207 unique lncRNAs were identified as expressed in at least one group and have all H3K4me3, H3K4me1 and H3K27ac histone marks enrichment in its matched cell type group. And a total of 26,144 unique enhancer-RNAs were identified as expressed in at least one group and having both H3K4me1 and H3K27ac but do not have H3K4me3 histone marks enrichment in the matched group. A breakdown list of number of lncRNAs identified in each group is in Table 3. For these lncRNAs identified in each group, we generated a supervised heatmap (Fig. 7) ensuring the order of the groups and the order of lncRNAs of each groups to be identical for rows and columns. Therefore, the diagonal position of the heatmap shows relatively higher FPKM values across all the samples which proved that lncRNAs identified in a specific group have the relatively higher expression values comparing to other groups. The unsupervised automatic hierarchical clustered heatmap (Fig. 8) for all the 8207 unique group-enriched lncRNAs revealed some clusters of lncRNAs which are differentially expressed in TALL cell lines comparing with other cell types.

Integration with Hi-C data

To demonstrate the flexibility of our pipeline we integrated RNA-seq/ChIP-seq data from Roadmap Epigenomics with matched Hi-C data from a previous study [32, 33]. First, we rerun the lncRNA-screen pipeline focusing only on the samples that have matched Hi-C data (the report is included in GitHub). For each cell type, we defined expressed mRNAs, lncRNAs including annotated and novel, enhancer-RNAs and enhancer regions using expression profile and H3K4me3, H3K27ac and H3K4me1 histone mark occupancy. The numbers of elements in each category for each cell type are included in Table 4. Hi-C analysis was performed using our HiC-bench pipeline [34]. HiC-bench automatically produces various plots to help the user assess the quality of the data as well as compare different samples. Paired-end reads were mapped to the reference genome (hg19 or mm10) using Bowtie2 [28]. Local alignments of input read pairs were performed as they consist of chimeric reads between two (non-consecutive) interacting fragments. This approach yielded a high percentage of mappable reads (>25%) for all datasets. Mapped read pairs were subsequently filtered for known artifacts of the Hi-C protocol such as self-ligation, mapping too far from the enzyme's known cutting sites etc. (Fig. 11). Samples clustered as expected by Principal Component Analysis

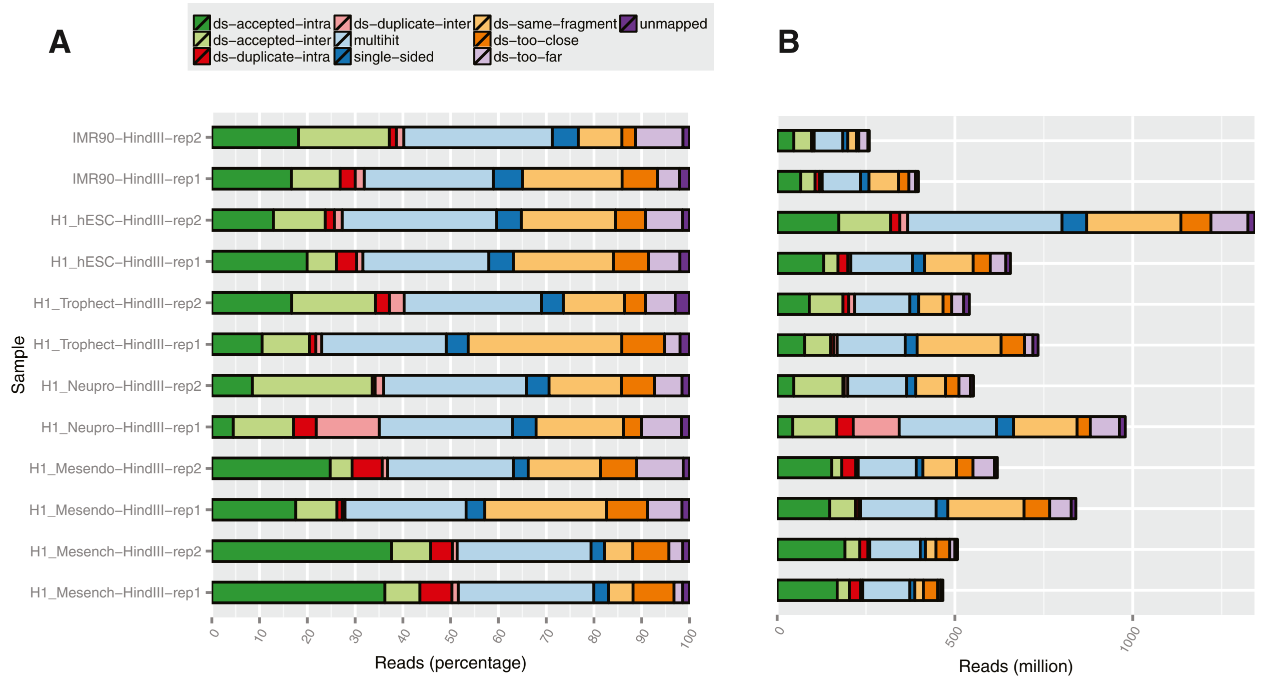


Fig. 11 Hi-C quality assessment plots. **a** Alignment and filtering statistics (read pair counts and percentages). **b** Average normalized Hi-C counts are presented as a function of the distance between the interacting partners

(Additional file 4: Figure S2A), and the average Hi-C count showed the characteristic dependency on the distance between the interacting fragments as demonstrated in previous studies (Additional file 4: Figure S2B). Additional file 4: Figure S3A, B shows the sizes and the number of detected topologically-associated domains (TADs) detected in each cell type and each replicate, and Additional file 4: Figure S3C shows the pairwise overlaps of boundaries between all pairs of samples and replicates. HiC-bench also generates a table of all the interacting loci annotated with genes, ChIP-seq peaks and any other region file that the user provides. Using this feature, we compiled a comprehensive report of all interactions that involve the lncRNAs discovered by our lncRNA-screen pipeline in the matched RNA-seq/ChIP-seq datasets. Overall, we found that 268 lncRNAs in hESC, 239 lncRNAs in mesendoderm cells, 9 lncRNAs in neural progenitor cells and 254 lncRNAs in mesenchymal cells interacting with at least one mRNA in cis within the context of topological domains. Most importantly, mRNA expression appears to be sensitive to changes in looping with their lncRNA interacting partners. We used HiCPlotter [35] to generate the Hi-C maps mRNA-lncRNA interaction plots. As an example we show the putative lncRNA CTD-2006C1.2 on chromosome 19 in Fig. 12. This lncRNA interacts with multiple protein-coding genes, mainly zinc-finger proteins within a single topological domain. When we compared both the expression and the Hi-C interaction intensity of the lncRNA to its neighboring genes, we observed an interesting contrasting pattern. Upstream, the lncRNA interacts with two protein-coding genes, ZNF441 and ZNF491: both genes are up-regulated in Mesenchymal cells compared to hESCs and show a

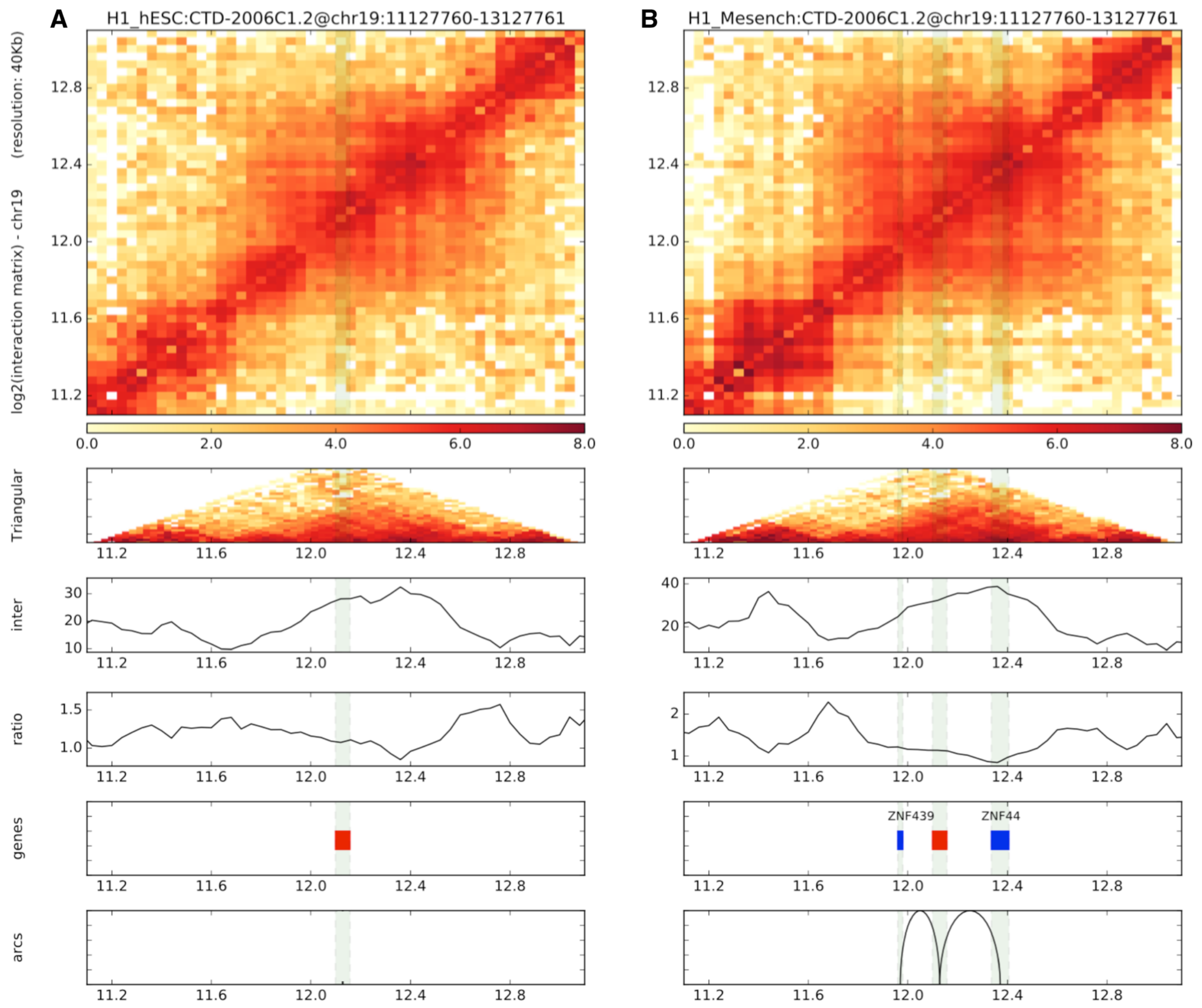


Fig. 12 Example of lncRNA-mRNA interactions. Putative lncRNA CTD-2006C1.2 is shown to differentially interact with two upstream and two downstream zinc-finger genes: hESCs (*left panel*) and Mesenchymal (*right panel*)

concomitant increase in Hi-C looping. In contrast, downstream, the lncRNA interacts with ZNF878 and ZNF625, both down-regulated in Mesenchymal cells compared to hESCs with concomitant decrease in Hi-C looping.

Conclusions

We developed lncRNA-screen, an easy-to-use integrative lncRNA discovery platform for comprehensive mapping and characterization of lncRNAs using a variety of genomics datasets. The main objective of this work was to facilitate the computational discovery of lncRNA candidates to be further examined by experimental screening as well as functional experiments. More specifically, our goal was to enable experimental laboratories with limited genomics expertise to quickly and comprehensively characterize lncRNAs in their particular field of study (e.g. cancer, stem cells, development). Our pipeline can be installed using one self-contained installation package available on GitHub, and, importantly, is designed to *enable execution of an entire analysis using a single command*. Initializing a new analysis is also simplified into setting up a trivial sample sheet to describe the datasets involved in the study. Additionally, lncRNA-screen generates an interactive lncRNA report which allows the user to explore the results of the analysis, define their own custom criteria for selecting lncRNAs, and interactively visualize the filtered results using the UCSC Genome Browser and pre-built genome snapshots. The pipeline is compatible with both stand-alone server environments and high-performance computing clusters. In summary, our pipeline provides a comprehensive solution for lncRNA discovery and an intuitive interactive report for identifying promising lncRNA candidates. lncRNA-screen is available as free open-source software on GitHub and our bioinformatics team offers installation and usage support.

Additional files

Additional file 1: Supplementary material 2. (DOCX 128 kb)

Additional file 2: Supplementary material 3. (DOCX 158 kb)

Additional file 3: Supplementary material 4. (DOCX 100 kb)

Additional file 4: Supplementary material 1. **Table S1.** Datasets included in this study. **Figure S1.** featureCounts assigned reads plot: (A) percentages, (B) counts. **Figure S2.** (A) Principal Component Analysis of Hi-C contact matrices. (B) Average Hi-C count as a function of distance between interacting loci. **Figure S3.** (A) Distribution of TAD sizes across all the samples. (B) Number of domains across all samples. (C) Pairwise overlaps of TAD boundaries across samples. (DOCX 1349 kb)

Abbreviations

CNVs: Copy number variations; FDR: False discovery rate; FPKM: Fragments per kilobase of transcript per million mapped reads; lncRNAs: Long Non-Coding RNAs; ORF: Open reading frame; SNPs: Single nucleotide polymorphisms; TSS: Transcription start site

Acknowledgements

We thank Igor Dolgalev and Stephen Kelly from the NYU Applied Bioinformatics Laboratories and all the Aifantis Lab members for inspiring discussions. We also thank Charalampos Lazaris for advice on how to use HiCPlotter. This work has used computing resources at the NYU High-Performance Computing Facility (HPCF), so we thank Eric Peskin, Ali Siavosh-Haghighi and Loren Koenig for their technical support.

Funding

The study was supported by a Research Scholar Grant, RSG-15-189-01 - RMC from the American Cancer Society to Aristotelis Tsirigos (AT) and the reseach program R01CA194923 and R01CA169784 from National Institute of Health (National Cancer Institute) to Iannis Aifantis (IA). Aristotelis Tsirigos was supported by a Research Scholar Grant, RSG-15-189-01 - RMC from the American Cancer Society.

Availability of data and materials

Published RNA-seq and ChIP-seq data are from Roadmap Epigenomic Project website (http://egg2.wustl.edu/roadmap/web_portal) [31]. Published Hi-C data were downloaded from Gene Expression Omnibus, using the accession number GSE52457 [32, 33].
Project name: lncRNA-screen.
Project home page: https://github.com/NYU-BFX/lncRNA-screen.
Archived version: https://github.com/NYU-BFX/lncRNA-screen/releases/tag/v.02.
Operating system: Redhat Linux GNU (64 bit).
Programming language: R, C++, Python, Unix shell scripts, Perl.
Other requirements: Python 2.7+, R 3.2.0+, Perl 5.0 + .
License: MIT.
Any restrictions to use by non-academics: None.

Authors' contributions

YG designed and implemented the pipeline, performed computational analyses, generated figures and wrote the user manual. HH and TT offered expertise on lncRNA biology. TT designed an early prototype of the pipeline. AT analyzed the Hi-C data. YG and AT wrote the manuscript. AT and IA designed and supervised this research project. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and Institutional affiliations.

Author details

[1]Department of Pathology and Laura and Isaac Perlmutter Cancer Center, New York University School of Medicine, New York, NY 10016, USA. [2]NYU Cancer Institute and Helen L. and Martin S. Kimmel Center for Stem Cell Biology, New York University School of Medicine, New York, NY 10016, USA. [3]Department of Human Genetics, Miller School of Medicine, University of Miami, Coral Gables, FL 33136, USA. [4]Department of Population Health, New York University School of Medicine, New York, NY 10016, USA. [5]Regeneron Pharmaceuticals, Inc., Tarrytown, NY 10591, USA. [6]Applied Bioinformatics Laboratories, New York University School of Medicine, New York, NY 10016, USA.

Received: 23 January 2017 Accepted: 24 May 2017
Published online: 05 June 2017

References

1. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. Nature. 2012;489:101–8.
2. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, et al. The transcriptional landscape of the mammalian genome. Science. 2005; 309:1559–63.

3. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, et al. RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science. 2007;316:1484–8.
4. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, Lagarde J, Veeravalli L, Ruan X, Ruan Y, Lassmann T, Carninci P, Brown JB, Lipovich L, Gonzalez JM, Thomas M, Davis CA, Shiekhattar R, Gingeras TR, Hubbard TJ, Notredame C, Harrow J, Guigó R: The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. 2012.
5. Sahu A, Singhal U, Chinnaiyan AM. Long noncoding RNAs in cancer: from function to translation. Trends in Cancer. 2015;1:93–109.
6. Kung JT, Colognori D, Lee JT. Long noncoding RNAs: past, present, and future. Genetics. 2013;193:651–69.
7. Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, et al. The landscape of long noncoding RNAs in the human transcriptome. Nat Genet. 2015;47:199–208.
8. Yan X, Hu Z, Feng Y, Hu X, Yuan J, Zhao SD, et al. Comprehensive genomic characterization of long non-coding RNAs across human cancers. Cancer Cell. 2015;28:529–40.
9. Yotsukura S, Du Verle D, Hancock T, Natsume-Kitatani Y, Mamitsuka H. Computational recognition for long non-coding RNA (lncRNA): software and databases. Brief Bioinformatics. 2016;18(1):9–27.
10. Jiang Q, Ma R, Wang J, Wu X, Jin S, Peng J, Tan R, Zhang T, Li Y, Wang Y: LncRNA2Function: a comprehensive resource for functional investigation of human lncRNAs based on RNA-seq data. 2015.
11. Volders PJ, Helsens K, Wang X, Menten B, Martens L, Gevaert K, et al. LNCipedia: a database for annotated human lncRNA transcript sequences and structures. Nucleic Acids res. 2013;41(Database issue):D246–51.
12. Amaral PP, Clark MB, Gascoigne DK, Dinger ME, Mattick JS. lncRNAdb: a reference database for long noncoding RNAs. Nucleic Acids res. 2011;39(Database issue):D146–51.
13. Park C, Yu N, Choi I, Kim W, Lee S. lncRNAtor: a comprehensive resource for functional investigation of long noncoding RNAs. Bioinformatics. 2014;30(17):2480–5.
14. Sun K, Chen X, Jiang P, Song X, Wang H, Sun H: iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. 2013.
15. Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res. 2007;35(Web Server issue):W345–9.
16. Wang L, Park HJ, Dasari S, Wang S, Kocher JP, Li W. CPAT: coding-potential assessment tool using an alignment-free logistic regression model. Nucleic Acids res. 2013;41:e74.
17. Sun L, Zhang Z, Bailey TL, Perkins AC, Tallack MR, Xu Z, et al. Prediction of novel long non-coding RNAs based on RNA-Seq data of mouse Klf1 knockout study. BMC Bioinformatics. 2012;13:331.
18. Sun L, Liu H, Zhang L, Meng J. lncRScan-SVM: a tool for predicting long non-coding RNAs using support vector machine. PLoS One. 2015;10:e0139654.
19. Sati S, Ghosh S, Jain V, Scaria V, Sengupta S. Genome-wide analysis reveals distinct patterns of epigenetic features in long non-coding RNA loci. Nucleic Acids res. 2012;40:10018–31.
20. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 2009;458:223–7.
21. Trimarchi T, Bilal E, Ntziachristos P, Fabbri G, Dalla-Favera R, Tsirigos A, et al. Genome-wide mapping and characterization of notch-regulated long noncoding RNAs in acute leukemia. Cell. 2014;158:593–606.
22. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
23. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14:R36.
24. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28:511–5.
25. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, et al. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. 2014;42(Database issue):D756–63.
26. Tatusova T, Ciufo S, Fedorov B, O'Neill K, Tolstoy I. RefSeq microbial genomes database: new representation and annotation strategy. Nucleic Acids res. 2014;42(Database issue):D553–9.
27. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923–30.
28. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. Nat Methods. 2012;9:357–9.
29. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9:R137.
30. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
31. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, et al. The NIH Roadmap Epigenomics mapping consortium. Nat Biotechnol. 2010;28:1045–8.
32. Xie W, Schultz MD, Lister R, Hou Z, Rajagopal N, Ray P, et al. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. Cell. 2013;153:1134–48.
33. Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, et al. Chromatin architecture reorganization during stem cell differentiation. Nature. 2015;518:331–6.
34. Lazaris C, Kelly S, Ntziachristos P, Aifantis I, Tsirigos A. HiC-bench: comprehensive and reproducible Hi-C data analysis designed for parameter exploration and benchmarking. BMC Genomics. 2017;18:22.
35. Akdemir KC, Chin L. HiCPlotter integrates genomic data with interaction matrices. Genome Biol. 2015;16:198.