

Detailed analysis of public RNAseq data and long non-coding RNA: a proposed enhancement to mesenchymal stem cell characterisation

Sebastien Riquier¹, Marc Mathieu¹, Anthony Boureux¹, Florence Ruffle¹, Jean-Marc Lemaitre¹, Farida Djouad¹, Nicolas Gilbert¹ and Therese Commes^{1,*}

February 27, 2020

Abstract

The development of RNA sequencing (RNAseq) and corresponding emergence of public datasets have created new avenues of transcriptional marker search. The long non-coding RNAs (lncRNAs) constitute an emerging class of transcripts with a potential for high tissue specificity and function. Using a dedicated bioinformatics pipeline, we propose to construct a cell-specific catalogue of unannotated lncRNAs and to identify the strongest cell markers. This pipeline uses *ab initio* transcript identification, pseudoalignment and new methodologies such as a specific k-mer approach for naive quantification of expression in numerous RNAseq data.

For an application model, we focused on Mesenchymal Stem Cells (MSCs), a type of adult multipotent stem-cells of diverse tissue origins. Frequently used in clinics, these cells lack extensive characterisation. Our pipeline was able to highlight different lncRNAs with high specificity for MSCs. *In silico* methodologies for functional prediction demonstrated that each candidate represents one specific state of MSCs biology. Together, these results suggest an approach that can be employed to harness lncRNA as cell marker, showing different candidates as potential actors in MSCs biology, while suggesting promising directions for future experimental investigations.

1 Introduction

¹

The increasing popularity of RNAseq and the ensuing aggregation of this data-type into public databases enable the search for new biomarkers across large cohorts of donors or cell types for the purpose pathological conditions or cellular lineages identification. As such, RNAseq has paved the way for the discovery of novel transcriptional biomarkers such as long noncoding RNAs (lncRNAs), that have emerged as a fundamental molecular class. A growing number of lncRNAs have been identified in the last decades, with their number approaching that of coding RNAs (17910 annotated human lncRNAs in the latest v32 version of the GENCODE versus 19 965 coding genes).

*To whom correspondence should be addressed: Tel: +33 046 7335711 ; Email: therese.commes@inserm.fr

¹IRMB, University of Montpellier, INSERM, Montpellier, France.

An increasing body of evidence has highlighted characteristics that define lncRNAs as therapeutic targets as well as potential tissue-specific markers [1]. Indeed, despite their non-coding nature, a large spectrum of functional mechanisms have been associated to lncRNAs [2, 3]. These include: endogenous competition (miRNA sponging for example), protein complex scaffolding and guides for active proteins with RNA-DNA homology interactions. These mechanisms occur in various physiological or pathological processes such as development, cancer and immunity [4, 5, 6]. To date, there is no finite list of long non-coding isoforms, making it difficult to construct a complete lncRNA catalogue due to the high number of transcripts and their tissue-specific expression [7, 8]. The absence of a complete catalogue makes it difficult to establish a comprehensive lncRNA expression profile. Thus, currently, the best strategy for the study of lncRNAs consists in the prediction of transcripts from a selection of RNAseq data in a tissue-specific condition. This strategy was successful in novel lncRNA biomarker discovery in pathological conditions [9, 10], but was poorly explored for cell lineage characterisation. Taking into account their functional importance and specificity, these RNAs should therefore not be ignored in establishing the molecular identity of a cell type.

Cell characterisation by specific markers bring different challenges such as the importance of probing the specificity of the marker and its limits in an extended number of various cell types, rather than using a control/patient experimental model. Moreover, the cells are not in a fixed state and display a variable transcriptional activity depending on cell status, environment, culture conditions and other parameters [1]. Furthermore, the lncRNAs' function is generally poorly assessed, except in the case of recurrent known transcripts (HOTAIR, H19). Thus, the *in silico* elaboration of a lncRNA catalogue that document the functional domains where the candidates could act, will be beneficial in the identification of lncRNAs' role and thus, in future experiments.

To this end, we have developed an integrated four-step procedure consisting of: i) an *ab initio* transcript reconstruction from RNAseq data and characterisation of novel transcripts, ii) a differential analysis using pseudoalignment coupled with a machine learning solution in order to extract the most cell-specific candidates, iii) an original step of tissue-expression validation with specific k-mers search in large and diversified transcriptomic datasets iv) an in-depth analysis to predict lncRNAs' functional potential from *in silico* prediction approaches. The notable advantage of introducing an *in silico* verification using k-mers is to allow a precise and in-depth determination of lncRNAs expression profile and to quickly interrogate their lineage specificity. In addition to that, validation of newly identified lncRNAs has been undertaken using RT-qPCR and long read sequencing (with Oxford Nanopore technology).

Mesenchymal Stem Cells (MSCs) are defined as multipotent adult stem cells, harvested from various tissues, including Bone Marrow (BM), Umbilical Cord (UC), and Adipose tissue (Ad). MSCs are an interesting cell type to explore since these cells lack the extended transcriptional characterisation that could highlight their lineage belonging and/or the possibility of distinguishing them from other mesodermal cell types such as fibroblasts and pericytes [11, 12]. The commonly admitted surface markers for MSCs, proposed by the International Society for Cellular Therapy (ISCT), and required to identify MSCs since 2006 are THY1 (CD90), NT5E (CD73), ENG (CD105) concerning the positive markers, and CD45, CD34, CD14 or CD11b, CD79alpha or CD19 and HLA-DR concerning the negative markers [13]. These markers are not distinctive and

may therefore not be sufficient for the definition of cellular or biological properties. Considering their different therapeutic properties (chondro and osteo differentiation potential, immunomodulation and production of trophic factors) [14] and given the increasing usage of these cells for academic and preclinical research [15], a detailed molecular characterisation of MSCs and predictive marker of functionality will constitute an important tool in regenerative medicine. lncRNAs have emerged as a class of transcripts with tissue-specific expression and important functions, such as the regulation of MSCs function [16, 17, 18], and remain largely unexplored in these cells.

To address this need, we performed a broad transcriptomic analysis of novel lncRNAs on human Mesenchymal Stem Cells (MSCs). We started from publicly available MSCs RNAseq, selecting ribodepleted datasets in order to enhance lncRNAs discovery and to explore the poly-(A)+ as poly(A)- lncRNAs. We restricted the differential expression analysis to a bone-marrow MSCs source compared to "nonMSC" cell counterparts. Once achieved, in depth *in silico* analysis was performed to check the lncRNA cell specific profiles with more and extensive datasets. To validate our approach, RNAseq data from eight publicly available libraries of normal MSCs containing a large diversity of noncancerous cell types were used for novel lncRNAs detection and tissue expression comparison. We initially reconstructed more than 70000 unannotated lncRNAs present in human bone-marrow MSCs. These lncRNAs were assigned depending on their position relative to annotated genes: MSC-related Long Intergenic Non Coding transcripts" named "Mlinc", and MSC-related Long Overlapping Antisense transcripts called "Mloanc". Among them, 35 Mlincs were specifically enriched in the cell lineage compared to the "non-MSC" counterpart group. Finally, after a further selection of the three most specific Mlincs, detailed *in vitro* and *in silico* functional exploration were performed.

2 Material and methods

2.1 Data collection and basic processing

The public RNAseq datasets (in fastq format) have been assessed using the ENCODE, EBI ArrayExpress service or SRA database at each step of the pipeline:

i) lncRNAs prediction and first differential analysis (Table S1), ii) k-mer search in ENCODE data to refine lncRNAs' specificity (Table S2), iii) k-mer search in FANTOM6 CAGE dataset and single cell analysis (scRNASeq) from Adipose MSCs by X. Liu et al raw data [19] for functional investigations (Table S3), iv) k-mer search in MSCs in different conditions (Table S4).

The reads quality were assessed with FastQC (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) to avoid the implementation of poor quality data in the analysis. Data from Peffers et al. [20], added to ENCODEs BM-MSCs RNAseq data, were selected for the MSCLinc and Mloanc characterisation and the differential step analysis considering the above-mentioned features: Ribo-zero technology, stranded and paired-ends RNAseq. Peffers data had a forward-reverse library orientation instead of a reverse-forward orientation of a classic Illumina sequencing, thereby the order of paired files was manually reversed. The fastq files used for lncRNAs prediction, referred as "MSC" group, were used for the differential analysis against the other cell types as "non-MSC" group, (Table S1) were mapped using CRAC v2.5.0 software [21] on the indexed GRCh38 human genome, including mitochondria, with stranded, -k 22 and rf options.

2.2 *Ab initio* assembly for transcripts prediction or unannotated transcripts prediction

The aligned reads of the "MSC" group were put through *ab initio* transcript assembly. Unannotated transcripts were predicted with the following procedure: i/ an *ab initio* reconstruction was performed on individual RNAseq with the StringTie [22] version 1.3.3b, with -c 5 -j 5 rf -f 0.1 (5 spliced reads are necessary to predict a junction and 5 reads at minimum are required to predict an expressed locus), ii/ the output individual gtf files obtained with the RNAseq of "MSC" group were then merged with the StringTie version 1.3.3b with -f 0.01 -m 200 and with a minimum TPM of 0.5, with the Ensembl human annotation (hg38) v90 used as guide for StringTie. The GTF was parsed with BEDTools [23] to dissociate new intergenic lncRNAs (lincRNAs) from annotated RNAs (coding or annotated lncRNAs), by applying filter criteria classically used in lncRNAs prediction [24], excluding transcript models overlapping (by 1 bp or more) any annotated coordinates. The resulting GTF of unannotated lincRNA from MSCs is referred as "Mlinc".

In parallel, the GTF was parsed with BEDTools to dissociate overlapping-antisens lncRNAs (named Mloanc), by applying filter criteria classically used in lncRNAs prediction, keeping transcript overlapping any annotated coordinates, then excluding transcript models overlapping these annotated coordinates on the same strand. The resulting GTF of MSCs overlapping-antisens lncRNAs is referred as "Mloanc" (Figure 1).

2.3 Long-read sequencing

The library was generated with 250 ng polyA+ mRNA purified from 50 μ g of human BM-MSCs total RNA. The polyA+ mRNA were treated according to the cDNA-PCR sequencing kit protocol (ref SQK-PCS108) as recommended by Oxford Nanopore. 3 254 396 sequences were obtained on the Oxford Nanopore Minion sequencer. The base calling was done with albacore version 2.2.7. 2 720 928 long-reads were successfully mapped using Minimap2 [25] version 2.10-r764 on GRCh38 human genome with default options used for Oxford Nanopore technology.

2.4 Quantification with pseudoalignment and feature selection

Kallisto v0.43.1 [26] was used directly on RNAseq raw fastq from the "MSC" and "non-MSC" groups. This pseudoalignment was performed with a number of bootstraps (-b) of 100, using a Kallisto index containing the sequences of all transcripts: the Ensembl coding and non-coding transcripts (v90) plus the predicted lincRNA and lncRNAs. Sleuth version 0.29.0 [27] was used with R for differential expression statistical analysis using the Walt test method, to compare the "MSC" group against the "non-MSC" group (including Lymphocytes, Macrophages, Hepatocytes, IPS, ESCs, HUVECs, Neurons, chondrocytes). Analysis was performed at the gene level for the annotated genes and at the transcript level for the predicted lincRNA and lncRNAs. Genes or lncRNAs having a log₂ fold-change between "MSC" and others greater than 0.5 and a p-value lower than or equal to 0.05 were selected. Finally, only transcripts/genes overexpressed in MSCs were selected. Each category (annotated transcripts, lincRNAs and lncRNAs) of potential candidates passing the first differentiation expression filter were separated for feature selection analysis. Boruta 6.0 [28] was used with 10000 maximum runs and a pvalue of 0.01 on each category, with multiple comparisons adjustment using the Bonferroni method (mcAdj = TRUE). Candidates passing the boruta test as "Confirmed" for each category were selected as reliable biomarkers.

2.5 Quantification by k-mers search

To quantify the expression of a transcript or a gene in available RNAseq with a rapid procedure, specific k-mers of 31nt length were extracted from the candidate sequence. A specific k-mer of an annotated candidate corresponds to a 31nt sequence that maps once on the genome and once on the reference transcriptome (Ensembl v90). In case of unannotated transcript (Mlinc, Mloanc) a specific k-mer maps once on the genome and is absent from reference transcriptome. The automated selection of specific k-mers is ensured by the Kmerator tool (in preparation) (<https://github.com/Transipedia/kmerator>). The k-mers were then quantified directly in raw fastq files using countTags (<https://github.com/Transipedia/countTags>). The quantification is expressed by the average count of all k-mers for one transcript, normalised by million of total k-mers in the raw file.

In FANTOM6 Dataset (<https://doi.org/10.1101/700864>, article not peer-reviewed) containing CAGE Analysis, to approach a "Transcript Per Millions" normalisation, the number of k-mers quantified was normalised by the total number of reads in million.

2.6 Genomic intervals assessment

DNase-seq intervals of enrichment were directly downloaded from ENCODE in bed format for BM-mesenchymal cells (ENCFF832FHZ) and hematopoietic progenitors (ENCFF378FCS). The H3K27ac (GSM3564514) and H3K4me3 (GSM3564510) ChIP results from undifferentiated BM-MSCs of the Agrawal Singh S. et al. study [29] were downloaded from GEO database in wig format, and remapped to the hg38 genome with CrossMap (<http://crossmap.sourceforge.net/>).

2.7 *in silico* functional prediction

We used LncADeep [30] to identify particular correlations between candidates and proteins. Beginning with our selection of 3 candidates, we filtered shared predicted proteins and selected protein uniquely predicted as interacting uniquely with the concerned candidate. The pathways concerned with these unique protein were identified with reactome.

Tarpmir was used to identify possible target site of human miRNA from miRbase ($p = 0.5$) [31] and FEELnc [32] to decipher the coding potential of candidates, using the coding and non-coding par of Ensembl annotation sequences as model.

2.8 Single-cell analysis

Single-cell data were pseudoaligned with Kallisto, with the same index used for the initial bulk RNAseq analysis. Pseudoalignment of 10X genomics data, correction, sorting and counting was made by Kallisto "bus" functions. Count matrices were processed with Seurat R package [33, 34]. Empty droplets were estimated by barcode ranking knee and inflection points, only droplet with a minimal count of 10000 were kept. In the end, 26071 droplets remain.

After normalisation, Inter-donor batch effect was corrected with ComBat method in sva R package [35] (Combat function, prior.plots=FALSE, par.prior=TRUE). Cell cycle scoring was made by CellCycleScoring Seurat function, using gene set used by the initial authors [19]. Finally, other sources of unnecessary variability as percent of mitochondrial genes, cell cycle and number of UMIs were regressed using ScaleData Seurat function.

To decipher genes enriched in cells positive for our markers, cells with a scaled expression superior or equal to 0.1 were labelled as positives, whereas cells with an expression inferior to the level were labelled as negatives. Then, markers of these cell were deciphered using FindAllMarkers Seurat function with a minimum FC threshold of 0.15. Expression of our markers in the Ad-MSCs population was made by FeaturePlot Seurat function after UMAP dimensional reduction, the gene enrichments were represented with VlnPlot function.

2.9 Data visualisation

Genome browser-like figures were generated with Gviz R package [36]. Bam tracks were generated from merged BAMs used for transcript prediction. Heatmaps were generated using superHeat R package (<https://github.com/rbarter/superheat>).

2.10 Ethics approval and consents

Human primary MSCs was obtained from patients who had granted the authors written informed consent with approval of the General Direction for Research and Innovation, the department in responsible for questions of ethics within the French Ministry of Higher Education and Research (registration number: DC-2009-1052). Human primary myoblasts were collected from patients of the CHU of Montpellier, France (the Montpellier University Hospital) who had provided informed consent. All experiments were performed in accordance with the Declaration of Helsinki and approved by the ethical committee of the CHU of Montpellier (France). Samples were approved for storage by the French "Ministre de l'Enseignement et de la Recherche" (NDC-2008-594). Liver samples were obtained from the Biological Resource Center of Montpellier CHU (CRB-CHUM; <http://www.chu-montpellier.fr>; Biobank ID: BB-0033-00031). The procedure was approved by the French Ethics Committee and written or oral consent was obtained from the patients.

2.11 Cell preparation and culture conditions

MSCs were isolated from bone marrow aspirates of patients undergoing hip replacement surgery, as previously described [37]. Cell suspensions were plated in α -MEM supplemented with 10 % FCS, 1 ng/mL FGF2 (R&D Systems), 2 mM L-glutamine, 100 U/mL penicillin and 100 μ g/mL streptomycin. These were shown to be positive for CD44, CD73, CD90 and CD105 and negative for CD14, CD34 and CD45 and used at the third or fourth passage. Human skin fibroblasts were cultured in DMEM high glucose supplemented with 10 % FCS. For Ad-MSCs isolation, adipose tissue was digested with 250 U/mL collagenase type II for 1 h at 37 °C and centrifuged (300 g for 10 min) using routine laboratory practices. The stroma vascular fraction was collected and cells filtered successively through a 100 μ m, 70 μ m and 40 μ m porous membrane (Cell Strainer, BD-Biosciences, Le-Pont-de-Claix, France). Single cells were seeded at the initial density of 4000 cell/cm² in α MEM supplemented with 100 U/mL penicillin/streptomycin (PS), 2 mmol/mL glutamine (Glu) and 10% fetal calf serum (FCS). After 24 h, cultures were washed twice with PBS. After 1 week, cells were trypsinised and expanded at 2000 cells/cm² till day 14 (end of passage 1), where Ad-MSCs preparations were used.

Human umbilical vein endothelial cells (HUVEC) obtained from Clonetics (Lonza, Levallois Perret, France) were cultured in complete EGM-2MV (Lonza) supplemented with 3 % FCS (HyClone; Perbio Science, Brebières, France)

Primary human myoblasts were isolated and purified from skeletal muscles of donors, as described by Kitzmann et al [38]. Purified myoblasts were plated in Petri dishes and cultured in growth medium containing Dulbecco's Modified Eagle's Medium (Gibco) supplemented with 20 % foetal bovine serum (FBS) (GE Healthcare, PAA), 0.5 % Ultroser G serum substitute (PALL life sciences) and 50 µg/ml Gentamicin (Thermo Scientific, France) at 37°C in humidified atmosphere with 5 % CO₂. All experiments were carried out between passage 4 (P4) and P8 to avoid cell senescence.

IPSCs were maintained in mTeSR-1TM medium (STEMCELL Technologies), in Petri dishes with matrigel (Corning, France). For the passages, cells were incubated in Gentle Cell Dissociation Reagent (STEMCELL Technologies) at room temperature, dissociation medium was discarded and cells incubated in mTeSR medium. All cell cultures were performed at 37°C with 5% of O₂ and 10% of CO₂.

Primary human hepatocytes (PHHs) were isolated, as described previously [39], from liver resections performed in adult patients.

NSC derived from H9 or directly bought (StemPro) have been cultivated on laminine with StemPro NSC SFM medium.

H9 embryonic stem cells were cultivated in ESICO medium in a coculture H9/MEF (Mouse Embryonic Fibroblasts) at 37 °C with 5% of O₂ and 5% CO₂.

2.12 RNA preparation and reverse transcription

Total RNA was isolated using TRIzol reagent (Invitrogen) or RNeasy Mini Kit (Qiagen, France) according to the manufacturer protocol. RNA was quantified using a NanoDrop ND-1000 spectrophotometer (Thermo Fisher Scientific, France). RNA quality and quantity were further assessed using the 2100-Bioanalyzer (Agilent Technologies, Waldbronn, Germany). Only preparations with RNA integrity number (RIN) values above 7 were considered. Reverse-transcription was performed either with random hexamers using the GeneAmp Gold RNA PCR Core kit (Applied Biosystems) or with oligo(dT) using SuperScriptTM First-Strand Synthesis System for RT-qPCR (Invitrogen, France).

2.13 Real-time quantitative PCR

Primer pairs were designed with primer3 online software (<http://bioinfo.ut.ee/primer3-0.4.0/>) from the transcripts' sequences. Primer pairs with a perfect and unique match on the human genome were validated with ucsc blat software (<https://genome.ucsc.edu>). As a final verification, primers were visualised in parallel with the bam alignment using IGV (<http://software.broadinstitute.org/software>) to verify that the primers overlap zones with read coverage. If possible, primer-pairs were designed to span an intron when present in the genomic sequence. Primers were designed for a mean Tm of 60°C. Quantitative PCR (qPCR) were performed using LightCycler 480 SYBR Green I Master mix and real-time PCR instrument (Roche). PCR conditions were 95 °C for 5 min followed by 45 cycles of 15 s at 95 °C, 10 s at 60 °C and 20 s at 72 °C. For each reaction, a single amplicon with the expected melting temperature was obtained.

The gene encoding ribosomal protein S9 (RPS9) was used as house-keeping gene for normalisation. The threshold cycle (Ct) of each amplification curve was calculated by Roche's LightCycler 480 software using the second derivative maximum method. The relative amount of transcripts were calculated using the ddCt method [40].

Results

For the purpose of generating a catalogue of all transcripts in any particular cell type, we developed a pipeline for the characterisation of all RNAs and their expression profile in a large collection of RNAseq data. The procedure includes four steps: i) an *ab initio* transcripts reconstruction from RNAseq data and identification of unannotated transcripts, ii) a differential analysis using pseudoalignment coupled with a machine learning solution in order to extract the most cell-specific candidates, iii) an original step of tissue-expression validation with a kmer approach (comparing large transcriptomic datasets), iv) an in-depth analysis to predict lncRNAs functional potential from *in silico* prediction approaches (Figure 1). To illustrate the procedure, we produced a RNA catalogue from bone-marrow MSCs ("MSC" group).

2.14 General features of the predicted MSC catalogue of lncRNAs

As mentioned above, we started with the *ab initio* reconstruction of any transcript from bone-marrow RNAseq with Stringtie assembler after mapping the reads with the CRAC software (see Materials and Methods for parameters). New isoforms of annotated transcripts were ignored. Of the 200 243 transcripts present in Ensembl annotation (version 90), 105 511 (52.6%) were detected in MSCs (TPM >0.1 in pseudoalignment quantification).

73 463 new lncRNAs were reconstructed. This fraction of unannotated transcripts represent 41% of detected transcripts, so in our case the *ab initio* reconstruction made it possible to almost double the inventory of detectable signatures in MSCs (Figure 2A). Of these, 34 712 were found to be intergenic and were thus referred to as "Mlinc" RNAs (MSC-related long intergenic non-coding RNAs), and 38 751 were found to be overlapping coding regions but in anti-sense orientation and thus referred to as "Mloanc" RNAs (MSC-related long overlapping antisense non-coding RNAs, with criteria as described in materials and methods and Figure 2A).

The *ab initio* method by itself is not sufficient to efficiently determine the lncRNAs' full length sequences. Moreover, this step does not preclude the possibility of false positives and at this point of the analysis, all the different rebuilt transcripts are considered to be windows of RNA expression or possible artefacts. These candidates are filtered and, for most interesting candidates, their true form is to be refined through experimental methods. We also assessed the general characteristics of predicted *de novo* lncRNAs in MSCs. Globally, Mlincs and Mloancs are shorter transcripts with longer exons compared to coding genes and annotated lncRNAs. The large majority of predicted lncRNAs are mono exonic (99% for Mlincs, 79% for Mloancs), with a length close to 200nt (Figure 2B-C). A consequence of the abundance of mono-exonic lncRNAs is an infinitesimally small number of variant forms. Only 0.15% and 0.82% of Mlincs and Mloancs are not mono isoforms, respectively.

The GC content of reconstructed lncRNAs is lower than that of coding or non-coding annotated genes (Figure 2D). This low GC proportion of around 40% is a common feature in *ab initio* transcript prediction, observed in a majority of studies of different species, from mammals, insects, plants or prokaryotes [41, 42, 43, 44].

2.15 Enrichment of a restricted set of Mlincs and Mloancs

In this second step, our objective was to obtain a restricted set of potential transcripts, using successive filtering approaches that would reveal their cell specificity. We quantified annotated transcripts, Mlincs and Mloancs with kallisto pseudoalignment [26] in a cohort constituted of two groups : the "MSC" group contained the BM-MSCs initially used for *ab initio* reconstruction,

and the "non-MSC" group used for comparison, composed of a large panel of different cell types including hESC, hematopoietic precursors and stem cells, primary chondrocytes, IPS, hepatocytes, neurons, lymphocytes and macrophages (Table S1).

Only over-expressed transcripts in "MSC" group versus "nonMSC" group were selected. Differential statistical tests were made with Sleuth, a tool specially dedicated to Kallisto results [27] (see all selective parameters in Materials and Methods). We performed two differential expression analyses: one at the gene level for Ensembl annotation, and the other at the transcript level for unannotated transcripts, to give the most likely variant form of the predicted lncRNAs. After this differential analysis, 2801 annotated genes, 655 Mlincs and 3032 Mloancs are significantly overexpressed in BM-MSCs (Figure 2 E-F).

The lncRNAs are commonly known to be less expressed than coding genes and this was observed in our selected annotated genes and new lncRNAs (Figure 2G). As a validation of our procedure, we found the 3 positive MSCs markers of ISCT among the selected annotated genes: THY1 (CD90), ENG (CD105), and NT5E (CD73). We also retrieved some influencers of MSCs activity, for example WNT5A [45, 46], Lamin A/C [47], FAP [48] and others. The complete list of selected genes is provided in Table S5.

2.16 Feature selection for the most discriminating coding and non-coding markers

In an attempt to select the best candidates, we retained lncRNAs with the most discriminative profile between "MSC" and "non-MSC" groups. In our case, the limitation with a classical "top" ranking by Fold Change (FC) or p-value is the presence of subgroups of different types of cells inside the negative "non-MSC" group. The FC, estimated by the Beta value in Figure 2C, appears to be a biased indicator of differential expression as it can select strong but localised expressed lncRNAs in cells poorly represented in our negative group, leading to potential false positive results.

To avoid this problem, we used the Boruta feature selection (see material and method section), for selection of discriminating features based on "random forest" machine learning methodology. Boruta was used separately on each group of candidates (annotated genes, Mlincs and Mloancs) to extract a restricted representation of the most relevant MSCs signatures. The top 35 importance scores were selected for genes, Mlincs and Mloancs. We arbitrary chose to select the first 35 transcripts for each group based on our observation of a drop in the importance score in coding the gene series. Considering the expression profile of the top 35 coding genes and predicted Mlincs, BM-MSCs clusterised independently from other cells types (Figure 3A).

In contrast, the selection of Mloancs did not provide a satisfying clustering as they had similar expression profiles in MSCs and other closely related cell types, in particular in primary chondrocytes (Figure S1). For this reason, Mloancs were not retained for further analysis. Selected annotated genes showed a poor specificity, with only few candidates showing a clear difference of expression between MSC and others: APCDD1L, HOTAIR, KRTAP1-5 and SMILR. The 3 positive MSCs markers from the ISCT were absent in this selection. The novel top 35 Mlincs showed less expression overall but with a more distinctive profile and a higher number of possible MSCs markers with clear contrast of expression. The characteristics, genomic intervals and sequences of the 35 candidates are presented in Table S7.

To assess the potential of genes already proposed as potential MSCs biomarkers by ISCT (Figure S2) or other potential MSCs markers proposed by different authors [14] (Figure S3), we

made a separated expression heatmap, without filter. Among these previously proposed markers, THY1 (CD90) presented the most specific profile. However, each gene presented expression in distinct non-MSC types.

2.17 Validation of selected Mlincs with long reads sequencing

As mentioned above, classical annotation of lncRNAs with *ab initio* short read methods suffers from inaccuracies and biases. The Oxford Nanopore long read technology (ONT) can sequence entire cDNA, which constitutes a clear technological advantage, not only in confirming the existence of the transcripts but also as it makes it possible to precisely identify the genomic intervals of lncRNA candidates. We performed long-read sequencing of a poly(A) RNA library obtained from a BM-MSCs sample. Among the top 35 selected Mlincs, with around 3 million total reads, 4 transcripts are covered with the ONT sequencing.

These intergenic lncRNAs are named as Stringtie output (SetName. TranscriptNumber. VariantNumber): Mlinc.28428.2, MlincV4.128022.2, MlincV4.89912.1 and MlincV4.64225.1. To support the above transcriptional units, we compared them with our short read data and searched for epigenetic status at the locus of the Mlincs in bone-marrow stromal mesenchymal cells. We looked at DNase sensitive site, H3K27 acetylation, H3K4 trimethylation that commonly corresponds to active regulatory regions (Figure 3B). We globally observed a DNA accessibility enrichment and acetylation of Histone 3 at the promotor region of our candidates, correlating with DNase sensitivity hotspots in BM mesenchymal cells that reinforce the prediction of the expression windows. In particular, for Mlinc.28428.2, the transcript observed with long reads sequencing corresponded to the prediction made with short reads. It was also supported by Mlinc.28428.1, a variant that differ by the absence of the second exon. Similar characteristics were observed for Mlinc.128022, which also produced two variants with a different organisation of 5 exons. The two other candidates, Mlinc.89912.1 and Mlinc.64225.1 are mono-exonic. Mlinc.89912.1 occurs at the close proximity of FGF5 3' end, in reverse orientation at this locus. For this reason, the different epigenomic features could not be attributed with certainty to the Mlinc. For Mlinc.64225.1, the sequence is longer than the *ab initio* short read prediction. KRTAP1-5, HOTAIR and SMILR, selected for their good expression profile, were also covered by long reads (Data not shown).

2.18 High-throughput investigation of a marker's specificity by specific k-mers search

A marker can only be considered specific within the limits of the diversity of samples used for its study. Considering the growing number of cell/tissues and transcriptional profiles, it is essential to probe the limits of a chosen biomarker against these various cell types. Most of published analyses highlighting new potential biomarkers of MSCs or fibroblasts have been restricted to a comparison between only few cell types and, as discussed, commonly described markers are not strictly distinctive. In order to assess the expression of Mlincs candidates in a large number of samples, we extracted specific 31nt k-mers from each of their sequences, as previously described [49]. These simplified but candidate-specific (oligonucleotide-like) probes allow a simple and fast presence/absence search to be made on large-scale cohorts and a direct quantification in raw fastq data. The k-mers were quantified in ENCODE human RNAseq database, including "primary cells" and "*in vitro* differentiated cell" categories (Table S2). Particularly, as the bibliography suggests that MSCs can also express phenotypic characteristics of endothelial, neural, smooth

muscle cells, skeletal myoblasts and cardiac myocytes, RNAseq samples from this mesodermal origin were tested.

With ISCT positive markers, we observed an expected expression profile that recapitulate previous biological studies, particularly the high expression of Endoglin (ENG, CD105) in endothelial cells (Figure S4) and the overexpression of NT5E (CD73) in epithelial and endothelial cells (Figure S5). Interestingly, their expression varied among MSCs cell sources : NT5E (CD73) was strongly enriched in adipose and BM derived MSCs, and THY1 (CD90) in umbilical cord derived MSCs (UC-MSCs) (Figure S6). We next analysed the expression profile using our candidate annotated genes Mlinc specific k-mers (Figure 4). The specific k-mers search supported the stated expression profile of Mlincs previously shown : our Mlinc candidates were positive in MSCs and displayed weak or absent expression in cells of ectodermal lineage, hematopoietic or endothelial origins.

However the high throughput and naive quantification in the ENCODE cohort made it possible to extend the observation of this absence of expression into cell types not previously studied. Moreover, this diversity showed that the expression of most of the candidates, contrarily to positive markers of the ISCT, were exclusive of cells with mesodermal origin. All candidates were expressed in at least one type of fibroblasts and differentially present in other mesodermal cell types. For the 4 selected Mlincs, they shared: (i) a systematic and strong expression in cell types like skin fibroblasts, and cells derived from reservoir of mesenchymal progenitors (muscle satellite cells or dermis papilla cells), (ii) a regular over-expression in regular cardiac myocytes, (iii) a significant and irregular expression in smooth muscle cells. The ENCODE cohort containing MSCs of different origins, we can therefore observe that the Mlincs show differences of expression depending of the tissular origin, these candidates being mainly expressed in two MSCs types. The results permitted the classification of our Mlincs according to observed specificity, from the most promising to the least restricted profile : Mlinc.28428.2 is expressed in Ad and BM derived MSCs. It is the candidate with the clearest absence of expression in non-mesodermal cells, and with the poorest relative expression in Smooth Muscle Cells (SMC). Mlinc.128022.2 is expressed in Adipose and Bone Marrow MSCs, and particularly in preadipocytes and muscle cells : (myoblasts, myocytes and myotubes). Mlinc.89912.1 is principally expressed BM-MSCs and less in UC-MSCs and AD-MSCs, but shows expression in epithelial and endothelial cell. Finally, Mlinc.64225.1 differs from other Mlincs as it is also strongly expressed in keratinocytes, HSC, and epithelial cells (Figure S7). Its expression in critical non-MSC types has led us to retain the three other Mlincs for further investigations.

2.19 RT-qPCR mimics the *in silico* prediction and deciphers multiple transcript variants

To confirm the specificity of selected Mlincs' expression experimentally, we performed RT-qPCR on a set of 80 RNA preparations from different primary cell (Figure 4C). These includes MSCs from Bone Marrow (BM), Adipose (Ad) and Umbilical Cord (UC), fibroblasts of different tissue origins, IPS cells, neural stem cells, myoblasts, HUVECs and hepatocytes (Table S6). RT-qPCR and amplicon sequencing using sets of specific primers (Table S7) confirmed different predicted forms of the Mlincs candidates in BM-MSCs (Figure S8). We designed two primer pairs for both Mlinc.128022 variants to validate the existence of first splice, and two pairs for Mlinc.28428 variants, one overlapping the second exon and another corresponding to a splice between first and third exon. All variations captured by the primer designs were quantified, suggesting that all these different variations predicted *in silico* exist biologically in MSCs. We confirmed most of the

expression profiles obtained by k-mers quantification using RT-qPCR, notably the specificity of expression dependency on the MSCs tissular origin: over expression of Mlinc.28428 and 128022 in BM- and Ad-MSCs. Nevertheless, few exceptions, such as Mlinc.89912.1, present an enrichment in UC-MSCs not found in k-mers quantification. Moreover, the restricted expression to cells of mesodermal origin is replicated in our RT-qPCR results. We obtained similar observations with annotated candidates : overexpression of KRTAP1-5 and SMILR in BM-MSCs specifically, and of HOTAIR in UC- and BM-MSCs.

2.20 *In silico* prediction of lncRNAs interactions and functions

The relative specificity of selected Mlincs for mesenchymal cells could be an indication of their roles in MSCs function. The prediction of their possible function could therefore suggest their suitability as markers of MSCs' function potential. To this end, we explored assumptions on the function of Mlinc.28428.2, Mlinc.128022.2 and Mlinc.89912.1 candidates using different published methods. We first used bioinformatic tools based on machine learning and deep learning to decipher general characteristics of our candidates : FEELnc to assess coding potential, tarpMir to decipher "miRNA sponge" function, and LncADeep to analyse potential interactions with proteins. Only two of the 35 selected Mlincs and none of the 3 selected Mlincs with validated specificity were revealed as potentially coding RNAs with the majority being predicted as non-coding by FEELnc (33/35). None candidate had more than five target sites for a given miRNA, indicating a low probability of a "miRNA sponge" activity (Table S7). For the 3 retained Mlincs, predicted interacting proteins by LncADeep were submitted to Reactome (Table S8).

We noted a predicted interaction between Mlinc.28428.2 and Betacatenin (CTNNB1) as part of apoptosis-linked modules, 5'-3' Exoribonuclease 1, component of the CCR4-NOT complex, mRNA Decapping Enzyme 1B as part of the mRNA decapping and decay pathways. The interaction was also predicted with different mediators of RNA polymerase II transcription subunits (MED), ATP Binding Cassette Subfamily B Members as part of the PPARA activity linked to ER-stress [50], and Proteasome subunits for intracellular transport, response to hypoxia and cell cycle modules. Mlinc.128022 could interact with important genes like THY1 (CD90), NRF1 (mitochondria metabolism) with no module clearly highlighted. Mlinc.89912 could interact with tubulins, UBB (ubiquitin B), SMG6 nonsense mediated mRNA decay factor and ribosomes subunits (RPSX) proteins, (RPL24) for nonsense Mediated Decay (NMD), PINK1 (mitophagy) and finally MGMT, as part of the MGMT mediated DNA damage reversal module.

We further quantified expression of candidates by counting their specific k-mers in the entire FANTOM6 set of 154 Known-Downed (KD) annotated lncRNAs in human dermal fibroblasts (<https://doi.org/10.1101/700864>, not peer-reviewed). We selected the KD experiments where expression of the Mlincs was statistically differential when compared with controls. Particular attention was paid to KD lncRNAs with reported function(s) in bibliography, and to KD lncRNAs overlapping a gene with reported functions. Mlinc.28428.2 is down-regulated when JPX, SERTAD4-AS1, BOLA3-AS1, and SNRPD3 are KD, and over-expressed with the KD of PTCHD3P1, ERVK3.1, MEG3, among other lncRNAs without reported function (Figure 5A). Interestingly, interactions between p53 pathway and JPX [51], SNRPD3 [52] and MEG3 [53, 54] respectively, have been previously reported. All these features converge on the hypothesis of a link between the function of Mlinc.28428, stress response, senescence and cellular maintenance. The implications of BOLA3 [55, 56] and PTCHD3P1 [57] in mitochondria homeostasis and glycolysis, the role of BOLA3 in stress response [58], the status of SERTAD4 as target of the YAP/TAZ pathway [59], vital pathway of stress response [60], and the role of MEG3 in aging [61], all reinforce

this hypothesis.

Mlinc.128022.2 is down-regulated with the KD of FOXN3-AS1, A1BG-AS1, CD27-AS1, and FLVCR1-AS1 (Figure 5B). FOXN3 seems to be more than a regulator of cell cycle, it is also described as a regulator of osteogenesis in different cases of defective craniofacial development [62, 63]. Moreover, the reported over-expression of FOXN3 during the early stages of MSCs osteodifferentiation [64], and downregulation of CD27-AS1 in MSCs of donors with bone fracture [65], allow us to hypothesise a possible function of Mlinc.128022 in bone remodelling and osteogenesis. In addition, both A1BG-AS1 and FLVCR1-AS have an influence in osteogenesis and cell differentiation. A recent study showed that A1BG-AS1 interacts with miR-216a and SMAD7 in suppressing hepatocellular carcinoma proliferation [66], both partners having an important role in the positive regulation of osteoblastic differentiation in mice [67] [68]. FLVR1 participates to resistance of oxydative stress by heme exportation in mouse MSCs [69], iron metabolism being closely linked with bone homeostasis, formation [70] and cell differentiation [71].

Finally, Mlinc.89912.1 is down-regulated after the KD of NEAT1-1 and PCAT6, and over-expressed when MFI2.AS1, CDKN2B.AS1 (or ANRIL) and MKLN1.AS2 are KD (Figure 5C). The manifest relations between cell proliferation and CDKN2B-AS1 [72, 73], MFI2 [74], MFI.AS1 [75], PCAT6 [76] and NEAT1 [77, 78], an combination with the between ones and DNA damage repair response, [79, 80] reinforces the prediction of a role of Mlinc.89912 in these mechanisms. Moreover, we explored RNAseq from chromatin, nucleus and cytoplasm subcellular compartments of fibroblastic cells in the FANTOM6 Dataset. Mlinc.28428 and Mlinc.128022 are enriched in at least cytoplasm, whereas Mlinc.89912 is enriched in free nucleus fraction suggesting interaction with nuclear component (Figure 5C).

2.21 The single cell RNAseq: an emergent level of completion in marker search

We analysed the single cell RNAseq data from 26 071 adipose MSCs (Ad-MSCs) to assess the heterogeneity of the 3 Mlincs and to explore their expression at the single-cell level. No clear correlation between cell cycle and expression of our Mlincs was identified (Figure S9). We observed a high variability of the number of cells expressing the markers (Threshold ≥ 0.1). 11 927/26 071 were Mlinc.28428-positives, 4944 were Mlinc.128022-positives, and 404 were Mlinc.89912-positives. For each Mlinc, we performed a differential test to decipher genes differentially expressed in Ad-MSCs Mlinc-positive and Mlinc-negative cells.

We found that Mlinc.28428-positive cells under-expressed H19 and PI16 (Figure 5A). These genes, that present a diversity of functions, are involved in stress mechanisms (oxydative response and shear stress), inflammation in fibroblasts and MSCs, and senescence pathways [81, 82, 83, 84]. Despite the low number of differentially expressed genes in Mlinc.28428-positive cells, their functional behaviour and their known targets suggest a pathway linked to stress response and senescence establishment that reinforce our previous assumptions on Mlinc.28428 function.

Mlinc.128022-positive cells are enriched in FTH1, TPM2, FTL and CD24 and present a lower expression in HMGN2, HMGB1, ODC1, PTTG1, BIRC5, EIF5A, MKI67, UBE2S, FGF5, HAS2-AS1 (Figure 5B). A significant portion of these genes are linked to osteogenic properties of MSCs as previously observed with FANTOM analysis. The Mlinc.128022-positive cells have an increased expression of ferritin (light and heavy chains), major actor in iron metabolism in osteoblastic cell line [85], that is also involved in osteogenic differentiation [86] and osteogenic calcification [87]. Two genes, enriched in Mlinc.128022-positive cells are positively linked to the osteogenic

differentiation potential of MSCs : the tropomyosin 2 (TPM2), downregulated when hMSCs were cultured in OS medium for the induction of osteoblasts at the calcification phase [88], and CD24 a membrane antigen recently proposed as a new marker for the sub-fraction of notochordal cells with increased differentiation capabilities [89]. In addition ODC1, under-represented in Mlinc.128022-positive cells, inhibited the MSCs osteogenic differentiation [90, 91]. Finally, the decrease of FGF5, MKI67, BIRC5 (Survivin) and PTTG1 (securin) expressions, all linked to proliferation active phases of cell cycle, tend to show cell with arrested cell cycle. These data suggest that the expression profile of Mlinc.128022 positive cells indicate a subpopulation of undifferentiated osteogenic progenitors, probably in senescence or quiescence.

Mlinc.89912-positive cells are enriched in FGF5 and HIST1H4C (Figure 5C). FGF5 is a protein with mitogenic properties, identified as an oncogene, that facilitates cell proliferation in both autocrine [92] and paracrine manner [93]. HIST1H4C, the Histone Core number 4, is a cell cycle-related gene. Modification of histone H4 (post-transcriptional or mutation) has been highlighted as important for Non-Homologous End-Joining (NHEJ) in yeast [94]. Its mutation cause genomic instability, resulting in increased apoptosis and cell cycle progression anomalies in zebrafish development. It reinforces our assumptions that Mlinc.89912 has a role in cell proliferation and DNA damage repair. In conclusion, the single cell RNAseq analysis enabled the observation of different features that characterise the phenotype of Mlincs positive cells and reinforced hypotheses on their functions previously observed through k-mers quantification.

2.22 K-mers analysis of markers in functional cell situation

Previously, we have presented a number of strategies to formulate hypotheses on the functions of an unannotated lncRNAs, suggesting directions of future experimental investigations. To evaluate the relevance of these strategies, we sought to quantify with specific k-mers search the expression of our Mlincs in MSCs in different conditions, linked to above mentioned findings: stress and senescence for Mlinc.28428.2, osteodifferentiation for Mlinc.128022.2 and cell cycle/proliferation for Mlinc.89912. We downloaded RNAseq data corresponding to the above-mentioned focus, described in Table S4.

As shown in Figure 6, we observed a statistically relevant increase of Mlinc.28428 expression in MSCs under replicative stress and in MSCs with CrisprCas9 depletion of genes with important role against senescence. In the Wang et al. study [95], MSCs senescence was observed with the KO of ATF6 and the stress induced with tunicamycin (endoplasmic reticulum stress) and late passage (replicative stress). Mlinc.28428 expression increased with tunicamycin treatment, late passage and ATF6 KO. The highest increase is observed in ATF6 KO MSCs associated with late passage condition.

In Fu et al. study [96], YAP, but not TAZ, was found to safeguard MSCs from cellular senescence as shown by KO experiments. Interestingly, YAP KO, but not TAZ, significantly increases the expression of Mlinc.28428.2. This would lead us to conclude that Mlinc.28428 is overexpressed in senescence and stress conditions, suggesting a role in one or both of these phenomena.

The change in Mlinc.128022 expression is strictly linked to osteodifferentiation conditions. Mlinc.128022 expression shows a relevant increase in MSCs exposed to fungal metabolite Cytochalasin D (CytoD). The cytoD is reported as a osteogenic stimulant in the concerned study [97]. Moreover, no expression variation was observed between MSCs and MSC-derived adipocytes from Wang et al. study, implying a role in adipodifferentiation. Agrawal Singh et al. have studied osteogenic MSCs differentiation [29], with a similar increase of Mlinc.128022 being observed after ten days.

We then quantified the expression of Mlinc.89912 in a study that compare proliferating MSCs versus confluent MSCs [98, 99]. Our candidate was clearly over-expressed in proliferating cells, validating its capacity to mark the MSCs in proliferation. Moreover, its expression was not statistically modified when MSCs were exposed to EGF with pro-mitotic capabilities [100]. However Mlinc.89912 expression was reduced when IWR-1, an inhibitor of Beta-catenin nuclear translocation, that reduced the proliferation of MSCs, was added to the medium. The functional domains of these genes is summarised in table 1 and confirm the potential functional role suggested from FANTOM data: stress-related pathways for Mlinc.28428, MSCs differentiation with a presumed orientation in osteo-progenitors for Mlinc.128022 and a more restricted role in proliferation and DNA repair for Mlinc.89912.

3 Discussion

With recent evolution of omics analysis, the landscape of biomarkers has been extended beyond known genes to the exploration of the unexplored transcriptome. This potential has been assessed in pathological conditions but to a lesser extent in cell-specific conditions, where this new pool of potential markers could be used to identify less well-characterised cells and hence predict their function. In this article, we propose an integrated procedure and strategies to identify the best markers (annotated or not) in a cell-specific condition, and predict their potential functions, primarily from RNA sequencing data (Figure 1). RNAseq facilitates the creation of large lncRNA catalogues [8, 101] through the total catalogue of lncRNAs remains incomplete given the diversity of biological entities and lncRNAs specific expression in non-pathological, cell-specific conditions. The creation of a "home-made" catalogue associated with a specific condition remains the best way to assess the full diversity of potential biomarkers in a cell, rather than resorting to a global catalogue made from diverse samples. To give an idea of the completeness of such a focused lncRNA catalogue when compared to a global one, Jiang et al. recently published "an expanded landscape of human long non-coding RNA" with 25 000 new lncRNAs from normal and tumor tissues, whereas in our focused analysis only 50% of our 35 selected MSCLinc can be found in this collection [101].

Futhermore, providing new candidates of good quality to improve lncRNA collection remains a complex task. As could be expected, the raw catalogue in our study contains predictions of disparate quality observed with a large number of mono-exonic transcripts. Without any filter, *ab initio* methods are insufficient to adequately reconstruct full length transcripts. The usage of long-read sequencing has been particularly effective in helping to validate our predictions. Given the benefits of full-length RNA sequence, long-read RNAseq should become the standard for lncRNA validation. A specific lncRNA can be the one presenting the most relevant properties after *in silico* analysis. The first task remain the identification of the more specific markers for a given cell type, task that present differences from classic comparative analysis. The MSCs markers proposed in the past were determined through a simple comparison between MSCs of a certain origin with negative cell whose types are either unique or few in number.

Historically, MSCs have been compared to bone marrow haematopoietic stem cells. However, our initial RNAseq analysis revealed that all potential MSCs markers proposed in the past are expressed in at least one other non-mesenchymatous cells type, and so do not constitute exclusive MSCs markers at the transcriptome level. Even if all cell types cannot be investigated, the diversity of the negative cell set is a critical criterion in selecting the most specific transcripts. In keeping with this idea, we then restricted the list of potential biomarkers with an enrichment step based

Table 1: Summary of functional investigation results

Mlinc	Predicted RNA-Protein interactions (lncADeep)	subcompartment enrichment	FANTOM6 expr. changes	Diff. genes in positive cells	K-mers investigations
Mlinc.28428	apoptosis, mRNA decay, PPARA activity, intracellular transport, response to hypoxia and cell cycle	Chromatin, cytoplasm	BOLA3-AS1, JPX, SERTAD4-AS1, PTCHD3P1, ERVK3.1, SNRPD3, MEG3	H19 - PI16 -	Stress, senescence
Mlinc.128022	THY1, NRF1	Chromatin, cytoplasm	FOXN3-AS1, A1BG-AS1, CD27-AS1, FLVCR1-AS1	FTH1, TPM2, FTL, CD24, HMGN2, HMGB1, ODC1, PTTG1, BIRC5, EIF5A, MKI67, UBE2S, FGF5, HAS2-AS1	Osteodiff., stress
Mlinc.89912	MGMT-mediated DNA damage reversal, Nonsense Mediated Decay, Tubulin metabolism	nucleus (free), cytoplasm	NEAT1_1, PCAT6, MFI2.AS1, MKLN1.AS2, CDKN2B.AS1	FGF5 and HIST1H4C	Proliferation

on a differential expression comparing BM-MSCs to other cells including mainly stem cells, as well as differentiated cells of various lineages (lymphocytes, macrophages, primary chondrocytes, hepatocytes and neurons). In the enriched list, the overexpressed annotated genes contained members of MSC-related pathways as well as the ISCT markers. This result supported the MSCs characterisation made by the original authors [13], thus validating the identity of MSCs used for this RNAseq analysis with the currently defined criteria. The problem with classical differential

analysis used on diverse "non-MSC" group is that all the "non-MSC" group is considered to be homogeneous. As a result, candidates with positive expression in small cell groups could pass statistical test, creating false positives. For this kind of differential analysis, we propose to select the most discriminating transcripts by feature selection, a machine learning methodology, that reduces the number of non-discriminating candidates after selection. We used feature selection through Boruta, a method based on random forest, to retain the top 35 of the most relevant MSCs signature for annotated genes, Mlincs and Mloancs separately. Putting aside our initial focus on unannotated lncRNAs, different annotated lncRNAs or coding genes with interesting profiles were also selected by feature selection : among them, KRTAP1-5 have been exclusively studied in BM-MSCs [102], where its preferential expression was validated by our results. These discoveries can bring new features concerning these genes and suggest directions for future investigations concerning their impact on the MSCs.

However, a marker is classically considered as specific on condition that its positive expression cannot be observed in any other cell type. Therefore, the expression of these potential markers should be explored in an entire RNAseq database to further validate its specificity. The exploration of a wide set of RNAseq data as proposed by ENCODE including a diversified set of primary and stem cells could support or invalidate the specificity of potential markers. In order to assess the expression of Mlincs candidates in a large number of samples, we used a signature for each candidate, extracting specific 31nt k-mers from their sequences. The specific k-mers extraction was made using Kmerator software. These k-mers were then quantified in the ENCODE human RNAseq database. The new and simplified procedure based on k-mers counting and large scale RNAseq exploration has the following advantages: i) a direct textual search that requires less time and CPU resources than classical methods, ii) a restricted set of lncRNAs supported by different results in the biological (wet) and *in silico* level (RNAseq data). The counterpart of the extensive vision of marker expression is that we observe a limit of specificity among our best candidates. We observed expression in fibroblasts, in close primary cells of common embryonic origin like smooth muscle cells (SMC) and other tissue-specific fibroblastic cells. Other tissue resident fibroblastic cells like skeletal muscle satellite cells, pre-adipocytes, and fibroblasts from different sources, especially dermis, express our selected MSClincs markers. The question of the differences between MSCs and related cell types is crucial to the issue. Specifically, the differences between MSCs and fibroblasts remain a subject of debate [103, 12]. According to the ISCT statement, no phenotypical differences have been reported between fibroblasts of different sources and adult MSCs [104], suggesting a hypothesis of a uniform cell type that show functional variation depending on the tissue source. Our results support this idea: distinguishing MSCs from fibroblast with only few positive markers remains a complicated task.

Moreover we observe low to medium expression of our candidates in close cell type from the same embryonic origin such as muscular cells and smooth muscle cells (SMC). This could be due to a shared phenotype between cells with close embryonic origin. Common markers between MSCs and SMCs have already been described. Notably, MSCs can express similar levels of SMC markers such as alpha-actin [105, 106]. Moreover Kumar et al. [107] determined that MSCs, pericytes, and SMCs could have the same mesenchymo-angioblast progenitor and that SMCs share a certain plasticity with MSCs as they can be differentiated in chondrocyte-like and beige adipocytes or myofibroblasts. However, a lot of cell types in ENCODE have not been actively sorted by expression of their respective surface markers, and fibroblast contamination is a classical feature in primary cell culture. We should not therefore exclude the possibility of fibroblast contamination when investigating marker for MSCs by bulk omics technology. Given this, single-cell RNAseq could be the best solution to identify the source of marker expression in counterpart cells.

To conclude, our extensive cell type comparison shows that the discovery of a marker of MSCs as distinct cell type is not plausible. After deepening our own research on MSCs biomarkers at the annotated and unannotated level, we were unable to find a marker that could simultaneously i) distinguish MSCs to close or homologous cell types (fibroblasts, satellite cells, SMCs) ii) be present in all MSCs types iii) distinguish MSCs from more characterised cell types (Hematopoietic lineage, neurones etc). Our results suggest, like other studies, a strong proximity between MSCs, fibroblast and mesodermal cell types.

More than a marker of MSCs, candidates extracted by our method could be used to explore important features in MSCs biology and therefore warrant investigation into their function, assuming that the specificity of RNA for a cell type can highlight its importance in cell activity. Even if the functional invalidation stands as the principal method to efficiently determine the function of a lncRNA, its expression and co-expression with known genes can potentially characterise a function or an intrinsic state of a cell type, particularly for MSCs with reported diversity of states and function (ex : differentiation, immunomodulation, senescence, proliferation...). In our opinion, it is vital that during the creation of a catalogue of lncRNAs, a restricted set of selected biomarkers should be studied more intensively, both in term of specificity and functions. Assumptions on functional domains, where lncRNAs could act, could increase the relevance and visibility of discovered lncRNAs, and far from the bioinformatics implications, encourage future biological investigations. We decided to investigate the three selected MSCLincs, validated by k-mers search, RT-qPCR and long-read sequencing, in term of biological impact with complementary *in silico* experimental approaches. We propose to use different *in silico* strategies, depending on the amount and diversity of the available data. The analysis confirms the non-coding potential of candidates and indicates a low probability of "miRNA sponge" activity. However, protein potential interaction results give interesting paths that were then investigated by complementary exploration. The k-mers quantification permits a naive high throughput exploration of numerous RNAseq data, simultaneously exploring potential functions and specificity to assess their potential. Instead of different cells, each candidate's expression was quantified in MSCs in different experimental conditions. FANTOM6 data recently offered a pilot about lncRNAs functional investigation, with a high-throughput invalidation of 154 lncRNAs and coding genes in fibroblasts and their RNAseq counterpart added to phenotypical observations. The utilisation of co-expressions between knock-out genes and candidates lncRNAs remains an efficient way to decipher lncRNAs function, provided number of KD genes is high. Moreover, the availability of recent single cell data of MSCs have been a good complement to lncRNAs functional investigation.

Using scRNAseq from Ad-MSCs [19], we observed that our markers are not expressed in all cells but constitute different subpopulations with different levels of rarity in Ad-MSCs. FANTOM6 and single-cell analysis could permit tracing three components of these states : stress inducible cells, lineage committed osteogenic progenitors and proliferating cells. Globally, we observed a global concordance of the results between the different strategies used for functional prediction. Mlinc.28428 has concomitant expression with genes related to the stress response pathway. Mlinc.28428 could be a good target for treatment to study the senescence process, age pathologies or stress response. Mlinc.128022 potentially interacts with THY1 (CD90) and has co-occurrences with genes linked to osteoprogenitors and cell differentiation. The k-mers search highlights its participation in MSCs' osteo-differentiation. Finally, Mlinc.89912 potentially interacts with damage repair and RNA decay, and tubulin metabolism, all linked to cell proliferation and cell cycle. Moreover, the subcompartment enrichment corresponds to this prediction: Mlinc.89912.1 is the only candidate to have possible interactions with DNA-repair system, a hypothesis corresponding to his observed enrichment in the nucleus. A final selection of bulk RNAseq of MSCs in specific biological con-

ditions allowed confirmation of our initial assumptions, showing that the different strategies we propose could be used to give relevant indication of the lncRNAs' functions. These results show that a lncRNA selected by its expression specificity has a high probability of being part of a functional mechanism.

In conclusion, we have predicted genes and lncRNAs enriched in MSCs and proposed several selection steps including feature selection (machine learning), large scale signature search, RT-qPCR validation, *in silico* tools and single cell analysis. We present the application of a new way of quantification in RNAseq : The specific k-mers search could be used as a naive information in lncRNA catalogue creation. The strategies presented here are transferable to other cell types and different studies while the specificity and functional assumption present a significant potential in long non-coding transcriptome exploration. We present three lncRNAs markers of bone marrow and adipose MSCs that passed all selection steps and present interesting features: Mlinc.28428.2, Mlinc.128022.2 and Mlinc.89912.1. These markers could be used by the scientific community as potential targets for functional biological experiments on MSCs, with pre-indications of potential functions to orientate the experiments, and finally initiate the objective of transition between informational problematics and cell biology.

4 Funding

Grant information: this work was supported by the Agence Nationale de la recherche for the projects "Computational Biology Institute" and "Transipedia" '[grant numbers 18-CE45-0020-02, ANR-10-INBS-09]' and the Canceropole Grand-Sud-Ouest Trans-kmer" project '[grant number 2017-EM24]'

Acknowledgements

We thank for their generous gifts, G.Carnac for myoblasts, M.Le Quintrec-Donnette for HUVECs, E. Sanchez for dermal fibroblasts, D. Noel and ML. Vignais for mesenchymal stromal cells, C. Crozet for IPS, S. Gerbal and M. Daujat for hepatocytes. We thank Philippe Clair for his advice on qPCR, the qPHD plateform, Montpellier GenomiX and Jean-Marc Holder (SeqOne) for text corrections.

4.0.1 Conflict of interest statement.

The authors declare that they have no competing interests.

References

- [1] Gloss, B. S. and Dinger, M. E. (January, 2016) The specificity of long noncoding RNA expression. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, **1859**(1), 16–22.
- [2] Meseure, D., Drak Alsibai, K., Nicolas, A., Bieche, I., and Morillon, A. (2015) Long Noncoding RNAs as New Architects in Cancer Epigenetics, Prognostic Biomarkers, and Potential Therapeutic Targets. *BioMed Research International*, **2015**.

- [3] Bouckenheimer, J., Assou, S., Riquier, S., Hou, C., Philippe, N., Sansac, C., Lavabre-Bertrand, T., Commes, T., Lematre, J.-M., Boureux, A., and Vos, J. D. (September, 2016) Long non-coding RNAs in human early embryonic development and their potential in ART. *Human Reproduction Update*.
- [4] Li, L. and Chang, H. Y. (October, 2014) Physiological roles of long noncoding RNAs: Insights from knockout mice. *Trends in cell biology*, **24**(10), 594–602.
- [5] Dhamija, S. and Diederichs, S. (2016) From junk to master regulators of invasion: lncRNA functions in migration, EMT and metastasis. *International Journal of Cancer*, **139**(2), 269–280.
- [6] Li, X. and Li, N. (December, 2018) LncRNAs on guard. *International Immunopharmacology*, **65**, 60–63.
- [7] Morillon, A. and Gautheret, D. (June, 2019) Bridging the gap between reference and real transcriptomes. *Genome Biology*, **20**.
- [8] Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guig, R., and Johnson, R. (September, 2018) Towards a complete map of the human long non-coding RNA transcriptome. *Nature reviews. Genetics*, **19**(9), 535–548.
- [9] James, A. R., Schroeder, M. P., Neumann, M., Bastian, L., Eckert, C., Gkbuget, N., Tanchez, J. O., Schlee, C., Isaakidis, K., Schwartz, S., Burmeister, T., von Stackelberg, A., Rieger, M. A., Gllner, S., Horstman, M., Schrappe, M., Kirschner-Schwabe, R., Brggemann, M., Mller-Tidow, C., Serve, H., Akalin, A., and Baldus, C. D. (January, 2019) Long non-coding RNAs defining major subtypes of B cell precursor acute lymphoblastic leukemia. *Journal of Hematology & Oncology*, **12**.
- [10] Liu, X., Ma, Y., Yin, K., Li, W., Chen, W., Zhang, Y., Zhu, C., Li, T., Han, B., Liu, X., Wang, S., and Zhou, Z. (June, 2019) Long non-coding and coding RNA profiling using strand-specific RNA-seq in human hypertrophic cardiomyopathy. *Scientific Data*, **6**(1), 1–7.
- [11] Lv, F.-J., Tuan, R. S., Cheung, K. M., and Leung, V. Y. (June, 2014) Concise Review: The Surface Markers and Identity of Human Mesenchymal Stem Cells. *STEM CELLS*, **32**(6), 1408–1419.
- [12] Soundararajan, M. and Kannan, S. (December, 2018) Fibroblasts and mesenchymal stem cells: Two sides of the same coin?. *Journal of Cellular Physiology*, **233**(12), 9099–9109.
- [13] Dominici, M., Le Blanc, K., Mueller, I., Slaper-Cortenbach, I., Marini, F., Krause, D., Deans, R., Keating, A., Prockop, D., and Horwitz, E. (2006) Minimal criteria for defining multipotent mesenchymal stromal cells. The International Society for Cellular Therapy position statement. *Cytotherapy*, **8**(4), 315–317.
- [14] Fitzsimmons, R. E. B., Mazurek, M. S., Soos, A., and Simmons, C. A. (August, 2018) Mesenchymal Stromal/Stem Cells in Regenerative Medicine and Tissue Engineering. *Stem Cells International*, **2018**.
- [15] Olsen, T. R., Ng, K. S., Lock, L. T., Ahsan, T., and Rowley, J. A. (June, 2018) Peak MSCAre We There Yet?. *Frontiers in Medicine*, **5**.

- [16] Tye, C. E., Gordon, J. A. R., Martin-Buley, L. A., Stein, J. L., Lian, J. B., and Stein, G. S. (March, 2015) Could lncRNAs be the missing links in control of mesenchymal stem cell differentiation?. *Journal of Cellular Physiology*, **230**(3), 526–534.
- [17] Kalwa, M., Hnzelmann, S., Otto, S., Kuo, C.-C., Franzen, J., Joussen, S., Fernandez-Rebollo, E., Rath, B., Koch, C., Hofmann, A., Lee, S.-H., Teschendorff, A. E., Denecke, B., Lin, Q., Widschwendter, M., Weinhold, E., Costa, I. G., and Wagner, W. (December, 2016) The lncRNA HOTAIR impacts on mesenchymal stem cells via triple helix formation. *Nucleic Acids Research*, **44**(22), 10631–10643.
- [18] Song, W., Gu, W., Qian, Y., Ma, X., Mao, Y., and Liu, W. (2015) Identification of long non-coding RNA involved in osteogenic differentiation from mesenchymal stem cells using RNA-Seq data. *Genetics and Molecular Research*, **14**(4), 18268–18279.
- [19] Liu, X., Xiang, Q., Xu, F., Huang, J., Yu, N., Zhang, Q., Long, X., and Zhou, Z. (February, 2019) Single-cell RNA-seq of cultured human adipose-derived mesenchymal stem cells. *Scientific Data*, **6**, 190031.
- [20] Peffers, M. J., Collins, J., Fang, Y., Goljanek-Whysall, K., Rushton, M., Loughlin, J., Proctor, C., and Clegg, P. D. (2016) Age-related changes in mesenchymal stem cells identified using a multi-omics approach. *European Cells & Materials*, **31**, 136–159.
- [21] Philippe, N., Salson, M., Commes, T., and Rivals, E. (2013) CRAC: an integrated approach to the analysis of RNA-seq reads. *Genome Biology*, **14**, R30.
- [22] Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L. (March, 2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, **33**(3), 290–295.
- [23] Quinlan, A. R. and Hall, I. M. (March, 2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**(6), 841–842.
- [24] Ilott, N. E. and Ponting, C. P. (September, 2013) Predicting long non-coding RNAs using RNA sequencing. *Methods (San Diego, Calif.)*, **63**(1), 50–59.
- [25] Li, H. (July, 2016) Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*, **32**(14), 2103–2110.
- [26] Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (May, 2016) Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, **34**(5), 525–527.
- [27] Pimentel, H., Bray, N. L., Puente, S., Melsted, P., and Pachter, L. (July, 2017) Differential analysis of RNA-seq incorporating quantification uncertainty. *Nature Methods*, **14**(7), 687–690.
- [28] Kursa, M. B., Jankowski, A., and Rudnicki, W. R. (December, 2010) Boruta - A System for Feature Selection. *Fundam. Inf.*, **101**(4), 271–285.
- [29] Agrawal Singh, S., Lerdrup, M., Gomes, A.-L. R., van de Werken, H. J., Vilstrup Johansen, J., Andersson, R., Sandelin, A., Helin, K., and Hansen, K. PLZF targets developmental enhancers for activation during osteogenic differentiation of human mesenchymal stem cells. *eLife*, **8**.

- [30] Yang, C., Yang, L., Zhou, M., Xie, H., Zhang, C., Wang, M. D., and Zhu, H. (November, 2018) LncADeep: an ab initio lncRNA identification and functional annotation tool based on deep learning. *Bioinformatics*, **34**(22), 3825–3834.
- [31] Ding, J., Li, X., and Hu, H. (September, 2016) TarPmiR: a new approach for microRNA target site prediction. *Bioinformatics*, **32**(18), 2768–2775.
- [32] Wucher, V., Legeai, F., Hdan, B., Rizk, G., Lagoutte, L., Leeb, T., Jagannathan, V., Cadieu, E., David, A., Lohi, H., Cirera, S., Fredholm, M., Botherel, N., Leegwater, P. A. J., Le Bguec, C., Fieten, H., Johnson, J., Alfldi, J., Andr, C., Lindblad-Toh, K., Hitte, C., and Derrien, T. (May, 2017) FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Research*, **45**(8), e57–e57.
- [33] Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (June, 2019) Comprehensive Integration of Single-Cell Data. *Cell*, **177**(7), 1888–1902.e21.
- [34] Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (May, 2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, **36**(5), 411–420.
- [35] Johnson, W. E., Li, C., and Rabinovic, A. (January, 2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**(1), 118–127.
- [36] Hahne, F. and Ivanek, R. (2016) Visualizing Genomic Data Using Gviz and Bioconductor. In Math, E. and Davis, S., (eds.), *Statistical Genomics: Methods and Protocols*, Methods in Molecular Biology pp. 335–351 Springer New York New York, NY.
- [37] Djouad, F., Bony, C., Hupl, T., Uz, G., Lahlou, N., Louis-Prence, P., Apparailly, F., Canovas, F., Rme, T., Sany, J., Jorgensen, C., and Nol, D. (2005) Transcriptional profiles discriminate bone marrow-derived and synovium-derived mesenchymal stem cells. *Arthritis Research & Therapy*, **7**(6), R1304–R1315.
- [38] Kitzmann, M., Bonnieu, A., Duret, C., Vernus, B., Barro, M., LaoudjChenivesse, D., Verdi, J. M., and Carnac, G. (2006) Inhibition of Notch signaling induces myotube hypertrophy by recruiting a subpopulation of reserve cells. *Journal of Cellular Physiology*, **208**(3), 538–548.
- [39] Pichard, L., Raulet, E., Fabre, G., Ferrini, J. B., Ourlin, J.-C., and Maurel, P. (2006) Human Hepatocyte Culture. In Phillips, I. R. and Shephard, E. A., (eds.), *Cytochrome P450 Protocols*, Methods in Molecular Biology pp. 283–293 Humana Press Totowa, NJ.
- [40] Livak, K. J. and Schmittgen, T. D. (December, 2001) Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2^{CT} Method. *Methods*, **25**(4), 402–408.
- [41] Niazi, F. and Valadkhan, S. (April, 2012) Computational analysis of functional long noncoding RNAs reveals lack of peptide-coding capacity and parallels with 3 UTRs. *RNA*, **18**(4), 825–843.
- [42] Wang, Y., Xu, T., He, W., Shen, X., Zhao, Q., Bai, J., and You, M. (January, 2018) Genome-wide identification and characterization of putative lncRNAs in the diamondback moth, *Plutella xylostella* (L.). *Genomics*, **110**(1), 35–42.

- [43] Cagirici, H. B., Alptekin, B., and Budak, H. (September, 2017) RNA Sequencing and Co-expressed Long Non-coding RNA in Modern and Wild Wheats. *Scientific Reports*, **7**.
- [44] Salari, R., Aksay, C., Karakoc, E., Unrau, P. J., Hajirasouliha, I., and Sahinalp, S. C. (May, 2009) smyRNA: A Novel Ab Initio ncRNA Gene Finder. *PLoS ONE*, **4**(5).
- [45] Gu, Q., Tian, H., Zhang, K., Chen, D., Chen, D., Wang, X., and Zhao, J. (2018) Wnt5a/FZD4 Mediates the Mechanical Stretch-Induced Osteogenic Differentiation of Bone Mesenchymal Stem Cells. *Cellular Physiology and Biochemistry*, **48**(1), 215–226.
- [46] Diederichs, S., Tonnier, V., Mrz, M., Dreher, S. I., Geisbsch, A., and Richter, W. (October, 2019) Regulation of WNT5A and WNT11 during MSC in vitro chondrogenesis: WNT inhibition lowers BMP and hedgehog activity, and reduces hypertrophy. *Cellular and molecular life sciences: CMLS*, **76**(19), 3875–3889.
- [47] Bermeo, S., Vidal, C., Zhou, H., and Duque, G. (2015) Lamin A/C Acts as an Essential Factor in Mesenchymal Stem Cell Differentiation Through the Regulation of the Dynamics of the Wnt/-Catenin Pathway. *Journal of Cellular Biochemistry*, **116**(10), 2344–2353.
- [48] Chung, K.-M., Hsu, S.-C., Chu, Y.-R., Lin, M.-Y., Jiaang, W.-T., Chen, R.-H., and Chen, X. (February, 2014) Fibroblast Activation Protein (FAP) Is Essential for the Migration of Bone Marrow Mesenchymal Stem Cells through RhoA Activation. *PLoS ONE*, **9**(2).
- [49] Ruffl, F., Audoux, J., Boureux, A., Beaumeunier, S., Gaillard, J.-B., Bou Samra, E., Megarbane, A., Cassinat, B., Chomienne, C., Alves, R., Riquier, S., Gilbert, N., Lemaitre, J.-M., Bacq-Daian, D., Boug, A. L., Philippe, N., and Commes, T. (December, 2017) New chimeric RNAs in acute myeloid leukemia. *F1000Research*, **6**, 1302.
- [50] Krieken, S. E. v. d., Popeijus, H. E., Mensink, R. P., and Plat, J. (2017) Link Between ER-Stress, PPAR-Alpha Activation, and BET Inhibition in Relation to Apolipoprotein A-I Transcription in HepG2 Cells. *Journal of Cellular Biochemistry*, **118**(8), 2161–2167.
- [51] Delbridge, A. R. D., Kueh, A. J., Ke, F., Zamudio, N. M., El-Saafin, F., Jansz, N., Wang, G.-Y., Iminitoff, M., Beck, T., Haupt, S., Hu, Y., May, R. E., Whitehead, L., Tai, L., Chiang, W., Herold, M. J., Haupt, Y., Smyth, G. K., Thomas, T., Blewitt, M. E., Strasser, A., and Voss, A. K. (April, 2019) Loss of p53 Causes Stochastic Aberrant X-Chromosome Inactivation and Female-Specific Neural Tube Defects. *Cell Reports*, **27**(2), 442–454.e5.
- [52] Siebringván Olst, E., Blijlevens, M., de Menezes, R. X., van der MeulenMuileman, I. H., Smit, E. F., and van Beusechem, V. W. (May, 2017) A genomewide siRNA screen for regulators of tumor suppressor p53 activity in human nonsmall cell lung cancer cells identifies components of the RNA splicing machinery as targets for anticancer treatment. *Molecular Oncology*, **11**(5), 534–551.
- [53] Zhou, Y., Zhong, Y., Wang, Y., Zhang, X., Batista, D. L., Gejman, R., Ansell, P. J., Zhao, J., Weng, C., and Klibanski, A. (August, 2007) Activation of p53 by MEG3 Non-coding RNA. *Journal of Biological Chemistry*, **282**(34), 24731–24742.
- [54] Uroda, T., Anastasakou, E., Rossi, A., Teulon, J.-M., Pellequer, J.-L., Annibale, P., Pessey, O., Inga, A., Chilln, I., and Marcia, M. (September, 2019) Conserved Pseudoknots in lncRNA MEG3 Are Essential for Stimulation of the p53 Pathway. *Molecular Cell*, **75**(5), 982–995.e9.

- [55] Haack, T. B., Rolinski, B., Haberberger, B., Zimmermann, F., Schum, J., Strecker, V., Graf, E., Athing, U., Hoppen, T., Wittig, I., Sperl, W., Freisinger, P., Mayr, J. A., Strom, T. M., Meitinger, T., and Prokisch, H. (2013) Homozygous missense mutation in BOLA3 causes multiple mitochondrial dysfunctions syndrome in two siblings. *Journal of Inherited Metabolic Disease*, **36**(1), 55–62.
- [56] Yu Qiujun, Tai Yi-Yin, Tang Ying, Zhao Jingsi, Negi Vinny, Culley Miranda K., Pilli Jyotsna, Sun Wei, Brugger Karin, Mayr Johannes, Saggar Rajeev, Saggar Rajan, Wallace W. Dean, Ross David J., Waxman Aaron B., Wendell Stacy G., Mullett Steven J., Sembrat John, Rojas Mauricio, Khan Omar F., Dahlman James E., Sugahara Masataka, Kagiyama Nobuyuki, Satoh Taiju, Zhang Manling, Feng Ning, Gorcsan John, Vargas Sara O., Harley Kathleen J., Kumar Rahul, Graham Brian B., Langer Robert, Anderson Daniel G., Wang Bing, Shiva Sruti, Bertero Thomas, and Chan Stephen Y. (May, 2019) BOLA (BolA Family Member 3) Deficiency Controls Endothelial Metabolism and Glycine Homeostasis in Pulmonary Hypertension. *Circulation*, **139**(19), 2238–2255.
- [57] Wang, J. and Li, K. (April, 2018) AB042. P013. LncRNAPTCHD3P1 enhances chemosensitivity of gemcitabine in pancreatic cancer and inhibits cancer cell proliferation and metastasis via inhibiting Warburg effect. *Annals of Pancreatic Cancer*, **1**(4).
- [58] Qin, L., Wang, M., Zuo, J., Feng, X., Liang, X., Wu, Z., and Ye, H. (April, 2015) Cytosolic BolA Plays a Repressive Role in the Tolerance against Excess Iron and MV-Induced Oxidative Stress in Plants. *PLoS ONE*, **10**(4).
- [59] Kitajima, S., Asahina, H., Chen, T., Guo, S., Quiceno, L. G., Cavanaugh, J. D., Merlino, A. A., Tange, S., Terai, H., Kim, J. W., Wang, X., Zhou, S., Xu, M., Wang, S., Zhu, Z., Thai, T. C., Takahashi, C., Wang, Y., Neve, R., Stinson, S., Tamayo, P., Watanabe, H., Kirschmeier, P. T., Wong, K.-K., and Barbie, D. A. (September, 2018) Overcoming Resistance to Dual Innate Immune and MEK Inhibition Downstream of KRAS. *Cancer cell*, **34**(3), 439–452.e6.
- [60] Raj, N. and Bam, R. (August, 2019) Reciprocal Crosstalk Between YAP1/Hippo Pathway and the p53 Family Proteins: Mechanisms and Outcomes in Cancer. *Frontiers in Cell and Developmental Biology*, **7**.
- [61] He, J., Tu, C., and Liu, Y. (2018) Role of lncRNAs in aging and age-related diseases. *AGING MEDICINE*, **1**(2), 158–175.
- [62] Schuff, M., Rssner, A., Wacker, S. A., Donow, C., Gessert, S., and Knchel, W. (2007) FoxN3 is required for craniofacial and eye development of *Xenopus laevis*. *Developmental Dynamics*, **236**(1), 226–239.
- [63] Samaan, G., Yugo, D., Rajagopalan, S., Wall, J., Donnell, R., Goldowitz, D., Gopalakrishnan, R., and Venkatachalam, S. (September, 2010) Foxn3 is essential for craniofacial development in mice and a putative candidate involved in human congenital craniofacial defects. *Biochemical and Biophysical Research Communications*, **400**(1), 60–65.
- [64] Brum, A. M., van de Peppel, J., van der Leije, C. S., Schreuders-Koedam, M., Eijken, M., van der Eerden, B. C. J., and van Leeuwen, J. P. T. M. (October, 2015) Connectivity Map-based discovery of parbendazole reveals targetable human osteogenic pathway. *Proceedings of the National Academy of Sciences of the United States of America*, **112**(41), 12711–12716.

- [65] del Real, A., Prez-Campo, F. M., Fernndez, A. F., Saudo, C., Ibarbia, C. G., Prez-Nez, M. I., Crieking, W. V., Braspenning, M., Alonso, M. A., Fraga, M. F., and Riancho, J. A. (December, 2016) Differential analysis of genome-wide methylation and gene expression in mesenchymal stem cells of patients with fractures and osteoarthritis. *Epigenetics*, **12**(2), 113–122.
- [66] Bai, J., Yao, B., Wang, L., Sun, L., Chen, T., Liu, R., Yin, G., Xu, Q., and Yang, W. (2019) lncRNA A1BG-AS1 suppresses proliferation and invasion of hepatocellular carcinoma cells by targeting miR-216a-5p. *Journal of Cellular Biochemistry*, **120**(6), 10310–10322.
- [67] Li, N., Lee, W. Y.-W., Lin, S.-E., Ni, M., Zhang, T., Huang, X.-R., Lan, H.-Y., and Li, G. (October, 2014) Partial loss of Smad7 function impairs bone remodeling, osteogenesis and enhances osteoclastogenesis in mice. *Bone*, **67**, 46–55.
- [68] Vishal, M., Vimalraj, S., Ajeetha, R., Gokulnath, M., Keerthana, R., He, Z., Partridge, N. C., and Selvamurugan, N. (2017) MicroRNA-590-5p Stabilizes Runx2 by Targeting Smad7 During Osteoblast Differentiation. *Journal of Cellular Physiology*, **232**(2), 371–380.
- [69] Nowak, W. N., Taha, H., Kachamakova-Trojanowska, N., Stpniewski, J., Markiewicz, J. A., Kusienicka, A., Szade, K., Szade, A., Bukowska-Strakova, K., Hajduk, K., Klska, D., Kopacz, A., Grochot-Przczek, A., Barthenheier, K., Cauvin, C., Dulak, J., and Jzkowicz, A. (October, 2017) Murine Bone Marrow Mesenchymal Stromal Cells Respond Efficiently to Oxidative Stress Despite the Low Level of Heme Oxygenases 1 and 2. *Antioxidants & Redox Signaling*, **29**(2), 111–127.
- [70] Balogh, E., Paragh, G., and Jeney, V. (October, 2018) Influence of Iron on Bone Homeostasis. *Pharmaceuticals*, **11**(4).
- [71] Puri, N., Sodhi, K., Haarstad, M., Kim, D. H., Bohinc, S., Foglio, E., Favero, G., and Abraham, N. G. (June, 2012) Heme Induced Oxidative Stress Attenuates Sirtuin1 and Enhances Adipogenesis in Mesenchymal Stem Cells and Mouse Pre-Adipocytes. *Journal of Cellular Biochemistry*, **113**(6), 1926–1935.
- [72] Luo, Y., Tao, H., Jin, L., Xiang, W., and Guo, W. (November, 2019) CDKN2B-AS1 Exerts Oncogenic Role in Osteosarcoma by Promoting Cell Proliferation and Epithelial to Mesenchymal Transition. *Cancer Biotherapy and Radiopharmaceuticals*.
- [73] Congrains, A., Kamide, K., Ohishi, M., and Rakugi, H. (January, 2013) ANRIL: Molecular Mechanisms and Implications in Human Health. *International Journal of Molecular Sciences*, **14**(1), 1278–1292.
- [74] Yin, Z., Ding, H., He, E., Chen, J., and Li, M. (October, 2016) Overexpression of long non-coding RNA MFI2 promotes cell proliferation and suppresses apoptosis in human osteosarcoma. *Oncology Reports*, **36**(4), 2033–2040.
- [75] Li, C., Tan, F., Pei, Q., Zhou, Z., Zhou, Y., Zhang, L., Wang, D., and Pei, H. (2019) Non-coding RNA MFI2-AS1 promotes colorectal cancer cell proliferation, migration and invasion through miR-574-5p/MYCBP axis. *Cell Proliferation*, **52**(4), e12632.

- [76] Zhu, C., Huang, L., Xu, F., Li, P., Li, P., and Hu, F. (October, 2019) LncRNA PCAT6 promotes tumor progression in osteosarcoma via activation of TGF- pathway by sponging miR-185-5p. *Biochemical and Biophysical Research Communications*,.
- [77] Dong, P., Xiong, Y., Yue, J., Hanley, S. J. B., Kobayashi, N., Todo, Y., and Watari, H. (October, 2018) Long Non-coding RNA NEAT1: A Novel Target for Diagnosis and Therapy in Human Tumors. *Frontiers in Genetics*, **9**.
- [78] Ahmed, A. S. I., Dong, K., Liu, J., Wen, T., Yu, L., Xu, F., Kang, X., Osman, I., Hu, G., Bunting, K. M., Crethers, D., Gao, H., Zhang, W., Liu, Y., Wen, K., Agarwal, G., Hirose, T., Nakagawa, S., Vazdarjanova, A., and Zhou, J. (September, 2018) Long noncoding RNA NEAT1 (nuclear paraspeckle assembly transcript 1) is critical for phenotypic switching of vascular smooth muscle cells. *Proceedings of the National Academy of Sciences*, **115**(37), E8660–E8667.
- [79] Taiana, E., Favasuli, V., Ronchetti, D., Todoerti, K., Pelizzoni, F., Manzoni, M., Barbieri, M., Fabris, S., Silvestris, I., Cantafio, M. E. G., Platonova, N., Zuccal, V., Maltese, L., Soncini, D., Ruberti, S., Cea, M., Chiaramonte, R., Amodio, N., Tassone, P., Agnelli, L., and Neri, A. (August, 2019) Long non-coding RNA NEAT1 targeting impairs the DNA repair machinery and triggers anti-tumor activity in multiple myeloma. *Leukemia*, pp. 1–11.
- [80] Wan, G., Mathur, R., Hu, X., Liu, Y., Zhang, X., Peng, G., and Lu, X. (May, 2013) Long non-coding RNA ANRIL (CDKN2B-AS) is induced by the ATM-E2F1 signaling pathway. *Cellular signalling*, **25**(5), 1086–1095.
- [81] Ding, K., Liao, Y., Gong, D., Zhao, X., and Ji, W. (July, 2018) Effect of long non-coding RNA H19 on oxidative stress and chemotherapy resistance of CD133+ cancer stem cells via the MAPK/ERK signaling pathway in hepatocellular carcinoma. *Biochemical and Biophysical Research Communications*, **502**(2), 194–201.
- [82] Yu, J.-L., Li, C., Che, L.-H., Zhao, Y.-H., and Guo, Y.-B. (2019) Downregulation of long noncoding RNA H19 rescues hippocampal neurons from apoptosis and oxidative stress by inhibiting IGF2 methylation in mice with streptozotocin-induced diabetes mellitus. *Journal of Cellular Physiology*, **234**(7), 10655–10670.
- [83] Hazell, G. G. J., Peachey, A. M. G., Teasdale, J. E., Sala-Newby, G. B., Angelini, G. D., Newby, A. C., and White, S. J. (December, 2016) PI16 is a shear stress and inflammation-regulated inhibitor of MMP2. *Scientific Reports*, **6**.
- [84] Puvvula, P. K. (May, 2019) LncRNAs Regulatory Networks in Cellular Senescence. *International Journal of Molecular Sciences*, **20**(11).
- [85] Spanner, M., Weber, K., Lanske, B., Ihbe, A., Siggelkow, H., Schtze, H., and Atkinson, M. J. (August, 1995) The iron-binding protein ferritin is expressed in cells of the osteoblastic lineage in vitro and in vivo. *Bone*, **17**(2), 161–165.
- [86] Balogh, E., Tolnai, E., Nagy, B., Nagy, B., Balla, G., Balla, J., and Jeney, V. (September, 2016) Iron overload inhibits osteogenic commitment and differentiation of mesenchymal stem cells via the induction of ferritin. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, **1862**(9), 1640–1649.

- [87] Zarjou, A., Jeney, V., Arosio, P., Poli, M., Antal-Szalms, P., Agarwal, A., Balla, G., and Balla, J. (June, 2009) Ferritin Prevents Calcification and Osteoblastic Differentiation of Vascular Smooth Muscle Cells. *Journal of the American Society of Nephrology*, **20**(6), 1254–1263.
- [88] Doi, M., Nagano, A., and Nakamura, Y. (January, 2002) Genome-wide Screening by cDNA Microarray of Genes Associated with Matrix Mineralization by Human Mesenchymal Stem Cells in Vitro. *Biochemical and Biophysical Research Communications*, **290**(1), 381–390.
- [89] Liu, Z., Zheng, Z., Qi, J., Wang, J., Zhou, Q., Hu, F., Liang, J., Li, C., Zhang, W., and Zhang, X. (December, 2018) CD24 identifies nucleus pulposus progenitors/notochordal cells for disc regeneration. *Journal of Biological Engineering*, **12**(1), 35.
- [90] Tsai, Y.-H., Lin, K.-L., Huang, Y.-P., Hsu, Y.-C., Chen, C.-H., Chen, Y., Sie, M.-H., Wang, G.-J., and Lee, M.-J. (2015) Suppression of ornithine decarboxylase promotes osteogenic differentiation of human bone marrow-derived mesenchymal stem cells. *FEBS Letters*, **589**(16), 2058–2065.
- [91] Chang, C.-F., Hsu, K.-H., Shen, C.-N., Li, C.-L., and Lu, J. (2014) Enrichment and Characterization of Two Subgroups of Committed Osteogenic Cells in the Mouse Endosteal Bone Marrow with Expression Levels of CD24. *Journal of Bone Research*, **2**(2), 1–9.
- [92] Park, G. C., Song, J. S., Park, H.-Y., Shin, S.-C., Jang, J. Y., Lee, J.-C., Wang, S.-G., Lee, B.-J., and Jung, J.-S. (May, 2016) Role of Fibroblast Growth Factor-5 on the Proliferation of Human Tonsil-Derived Mesenchymal Stem Cells. *Stem Cells and Development*, **25**(15), 1149–1160.
- [93] Kornmann, M., Ishiwata, T., Beger, H. G., and Korc, M. (September, 1997) Fibroblast growth factor-5 stimulates mitogenic signaling and is overexpressed in human pancreatic cancer: evidence for autocrine and paracrine actions. *Oncogene*, **15**(12), 1417–1424.
- [94] Williamson, E. A., Wray, J. W., Bansal, P., and Hromas, R. (2012) Overview for the Histone Codes for DNA Repair. *Progress in molecular biology and translational science*, **110**, 207–227.
- [95] Wang, S., Hu, B., Ding, Z., Dang, Y., Wu, J., Li, D., Liu, X., Xiao, B., Zhang, W., Ren, R., Lei, J., Hu, H., Chen, C., Chan, P., Li, D., Qu, J., Tang, F., and Liu, G.-H. (January, 2018) ATF6 safeguards organelle homeostasis and cellular aging in human mesenchymal stem cells. *Cell Discovery*, **4**.
- [96] Fu, L., Hu, Y., Song, M., Liu, Z., Zhang, W., Yu, F.-X., Wu, J., Wang, S., Izpisua Belmonte, J. C., Chan, P., Qu, J., Tang, F., and Liu, G.-H. (April, 2019) Up-regulation of FOXD1 by YAP alleviates senescence and osteoarthritis. *PLoS Biology*, **17**(4).
- [97] Samsonraj, R. M., Dudakovic, A., Manzar, B., Sen, B., Dietz, A. B., Cool, S. M., Rubin, J., and van Wijnen, A. J. (December, 2017) Osteogenic Stimulation of Human Adipose-Derived Mesenchymal Stem Cells Using a Fungal Metabolite That Suppresses the Polycomb Group Protein EZH2. *Stem Cells Translational Medicine*, **7**(2), 197–209.

- [98] Dudakovic, A., Gluscevic, M., Paradise, C. R., Dudakovic, H., Khani, F., Thaler, R., Ahmed, F. S., Li, X., Dietz, A. B., Stein, G. S., Montecino, M. A., Deyle, D. R., Westendorf, J. J., and van Wijnen, A. J. (April, 2017) Profiling of human epigenetic regulators using a semi-automated real-time qPCR platform validated by next generation sequencing. *Gene*, **609**, 28–37.
- [99] Camilleri, E. T., Gustafson, M. P., Dudakovic, A., Riester, S. M., Garces, C. G., Paradise, C. R., Takai, H., Karperien, M., Cool, S., Sampen, H.-J. I., Larson, A. N., Qu, W., Smith, J., Dietz, A. B., and van Wijnen, A. J. (August, 2016) Identification and validation of multiple cell surface markers of clinical-grade adipose-derived mesenchymal stromal cells as novel release criteria for good manufacturing practice-compliant production. *Stem Cell Research & Therapy*, **7**.
- [100] Knight, C., James, S., Kuntin, D., Fox, J., Newling, K., Hollings, S., Pennock, R., and Genever, P. (January, 2019) Epidermal growth factor can signal via -catenin to control proliferation of mesenchymal stem cells independently of canonical Wnt signalling. *Cellular Signalling*, **53**, 256–268.
- [101] Jiang, S., Cheng, S.-J., Ren, L.-C., Wang, Q., Kang, Y.-J., Ding, Y., Hou, M., Yang, X.-X., Lin, Y., Liang, N., and Gao, G. (September, 2019) An expanded landscape of human long noncoding RNA. *Nucleic Acids Research*, **47**(15), 7842–7856.
- [102] Chang, T.-H., Huang, H.-D., Ong, W.-K., Fu, Y.-J., Lee, O. K., Chien, S., and Ho, J. H. (April, 2014) The effects of actin cytoskeleton perturbation on keratin intermediate filament formation in mesenchymal stem/stromal cells. *Biomaterials*, **35**(13), 3934–3944.
- [103] Chang, Y., Li, H., and Guo, Z. (2014) Mesenchymal stem cell-like properties in fibroblasts. *Cellular Physiology and Biochemistry: International Journal of Experimental Cellular Physiology, Biochemistry, and Pharmacology*, **34**(3), 703–714.
- [104] Denu, R. A., Nemcek, S., Bloom, D. D., Goodrich, A. D., Kim, J., Mosher, D. F., and Hematti, P. (August, 2016) Fibroblasts and Mesenchymal Stromal/Stem Cells Are Phenotypically Indistinguishable. *Acta Haematologica*, **136**(2), 85–97.
- [105] Ball, S. G., Shuttleworth, A. C., and Kiely, C. M. (April, 2004) Direct cell contact influences bone marrow mesenchymal stem cell fate. *The International Journal of Biochemistry & Cell Biology*, **36**(4), 714–727.
- [106] Tamama, K., Sen, C. K., and Wells, A. (October, 2008) Differentiation of Bone Marrow Mesenchymal Stem Cells into the Smooth Muscle Lineage by Blocking ERK/MAPK Signaling Pathway. *Stem Cells and Development*, **17**(5), 897–908.
- [107] Kumar, A., DSouza, S. S., Moskvin, O. V., Toh, H., Wang, B., Zhang, J., Swanson, S., Guo, L.-W., Thomson, J. A., and Slukvin, I. I. (May, 2017) Specification and Diversification of Pericytes and Smooth Muscle Cells from Mesenchymoangioblasts. *Cell Reports*, **19**(9), 1902–1916.

5 Figure Legends

5.1 Figure 1: Flowchart representation of the Pipeline used for this in the study.

The 4 steps of the flowchart are described A) Ab initio reconstruction of transcript expressed in MSC from SRA dataset and creation of a reference (gtf+fasta) for quantification of Ensembl annotated genes, unannotated intergenic (Mlincs) and unannotated overlapping antisens (Mloanc). The results are shown in Fig2.

B) Differential Analysis for the selection of MSC markers (restrained candidate set) with i/ kallisto pseudoalignment and Sleuth differential test followed by feature selection by random forest with Boruta package. Long read sequencing and active transcription in MSC by epigenetic marks information completed the selection step (see figures 2 and 3).

C) Validation of cell expression specificity of candidates by kmer quantification in ENCODE RNAseq datasets (see table S2 for list of data) and qPCR validation. The results are presented in figure 4.

D) Functional investigations were performed with in silico prediction methods from the sequence of candidates, followed by k-mer quantification with FANTOM6 data set, single cell RNAseq and selected MSC conditions. Kmer quantification phases are shown by corresponding icons (figures 5 and 6).

5.2 Figure 2: Overview of annotated genes and unannotated transcripts enriched in BM-MSC.

A) Left pannels represented: i/EnsemblV90 transcript categories and distribution, ii/ transcripts distribution expressed in MSCs, showing unnotated transcripts obtained with Ab initio reconstruction by StringTie vs annotated transcripts (expression > 0.1 TPM) iii/ Predicted long non-coding RNA(lncRNA) from unannotated reconstructed transcripts include new long non coding RNA with intergenic (Mlinc) and antisens (Mlncoa) RNA categories.

B-C-D) Distribution of transcript length, exon length and GC percentage across different categories respectively with the same colors as in A pannel : coding transcripts (blue), annotated lincRNA (pink), annotated overlapping antisens lncRNA (purple), novel lincRNA (Mlincs, yellow), novel overlapping antisens RNA (Mloanc, red).

E) Representation of annotated genes (top pannel) and unannotated transcripts (bottom pannel) overexpressed in MSC versus non-MSC types ($\log_{2}FC > 0.5$ and $padj < 0.05$), separately showed in MA plot.

F) Total number of transcripts transcript by category. The colored bar indicated the number of differential expression of annotated genes (Ensemblv90) and unannotated transcripts (Mlinc and Mloanc).

G) Global expression in BM-MSC (with Sleuth normalization) of the same categories as in F for annotated genes and unannotated transcripts.

5.3 Figure 3: Selection of a refined set of best candidates by random forest (top35), long read sequencing and epigenetic features.

- A) Expression of the best MSC-specific candidates selected by boruta machine learning along MSC group and not MSC cohorts. Left pannel : top35 most relevant annotated genes (non-coding included); Right pannel: unannotated intergenic lncRNAs (Mlincs) and their average importance scores determined by Boruta method displayed in upside line plot.
- B) Genomic visualisation of Mlincs predictions 28428, 128022, 89912, and 64225 (MSClinc orange) from short reads alignment of all MSC group files (blue/magenta and bam visualisation), compared with long read alignments (long reads grey). Additional epigenomic features are shown to reveal active transcriptional activity from trimethylation of Histone H3 (H3K4me3), acetylation of Histone H3 H3K27 in MSCs (H3K4me3 and H3K27ac, green), and Dnase sensitivity hotspots of MSC (MSC DNase, red).

5.4 Figure 4: High throughput exploration of selected candidate across strong variety of samples by k-mer quantification in RNAseq and biological validation by RT-qPCR.

- A : List of tissues for the cell specific expression exploration (samples with ID numbers are listed in table S1)
- B: Relative expression of Mlinc.28428.2, Mlinc.128022.2, and Mlinc.89912.1 across ENCODEs ribodepleted RNAseq datas, made by k-mer quantification, normalized by k-mer by million.
- C: qPCR relative quantification was performed on the selected 3 Mlincs in MSC of different origins (BM-MSC, Ad-MSC, Umbilical cord msc) and other indicated cell types. Relative quantification (Log induction) was quantified by ddCt method using non MSC types as calibrator (mean of triplicates). Student tests have been made between triplicates, each test use BM-MSCs as reference group (ns : $P > 0.05$, *: $P \leq 0.05$, **: $P \leq 0.01$, ***: $P \leq 0.001$, ****: $P \leq 0.0001$).

5.5 Figure 5: Prediction of potential functions of candidates with k-mer quantification and single-cell.

For each Mlinc (Mlinc.28428 (A), Mlinc.128022 (B) and Mlinc.89912 (C) respectively) 3 steps of prediction were performed. a. Enrichment in the different subcompartments of fibroblasts from FANTOM6 dataset: Free nuclear fraction (Nuc), chromatin (Chr) and cytoplasm (Cyt); b. Expression of marker in FANTOM6 data depending of the KnockDown (KD) of an annotated lncRNA. Normalised count of all specific k-mers is averaged by sample (zeros values deleted) and t-tests are made between control and KD fibroblasts. c. General expression of Mlincs inside Ad-MSC population, dimensionnal reduction made with UMAP method, made from batch corrected counts. Expression of differentially expressed annotated genes between positive and negative cells for Mlinc.28428, Mlinc.128022 and Mlinc.89912 respectively.

5.6 Figure 6: Expression of markers in different datasets from SRA in cell conditions related to previous findings

- A) Expression of Mlinc.28428.1 in the context of oxydative, replicative, or KO-driven, stress and senescence (PRJNA396193, PRJNA433339). Relevant changes of expressions are showed with t-test results (ns : $P > 0.05$, *: $P \leq 0.05$, **: $P \leq 0.01$, ***: $P \leq 0.001$, ****: $P \leq 0.0001$).
- B) Expression of Mlinc.128022 in osteodifferentiation conditions (PRJNA515466) or osteodifferentiation potential (PRJNA379707). Relevant changes of expressions are showed with t-test results (ns : $P > 0.05$, *: $P \leq 0.05$, **: $P \leq 0.01$, ***: $P \leq 0.001$, ****: $P \leq 0.0001$).
- C) Expression of Mlinc.89912 in the context of proliferation (PRJNA328824 and PRJNA498109). Relevant changes of expressions are showed with t-test results (ns : $P > 0.05$, *: $P \leq 0.05$, **: $P \leq 0.01$, ***: $P \leq 0.001$, ****: $P \leq 0.0001$). The detailed list of datasets is provided in table S4.

Figure 1

bioRxiv preprint doi: <https://doi.org/10.1101/2020.03.09.976001>; this version posted March 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

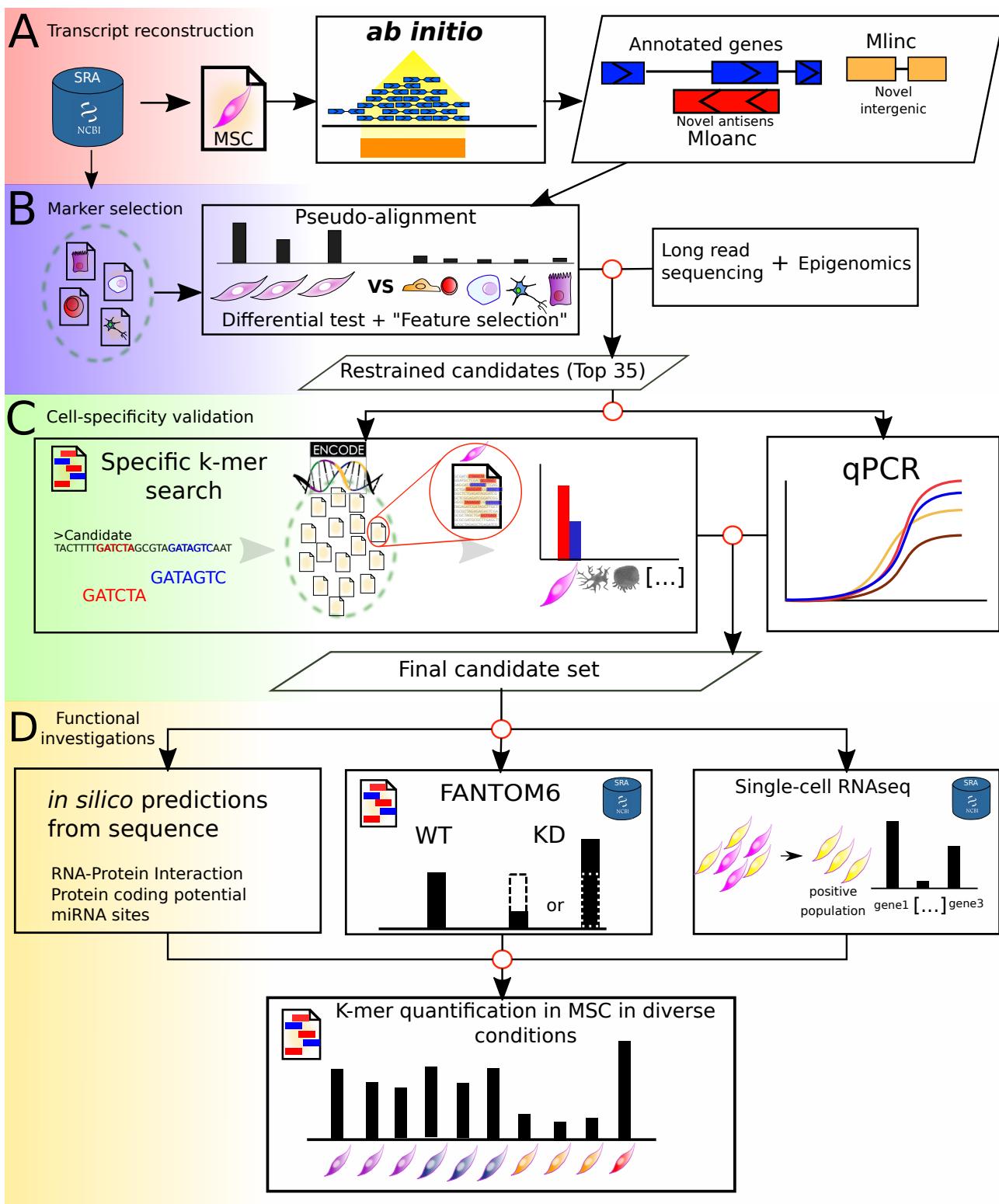


Figure 2

bioRxiv preprint doi: <https://doi.org/10.1101/2020.03.09.976001>; this version posted March 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

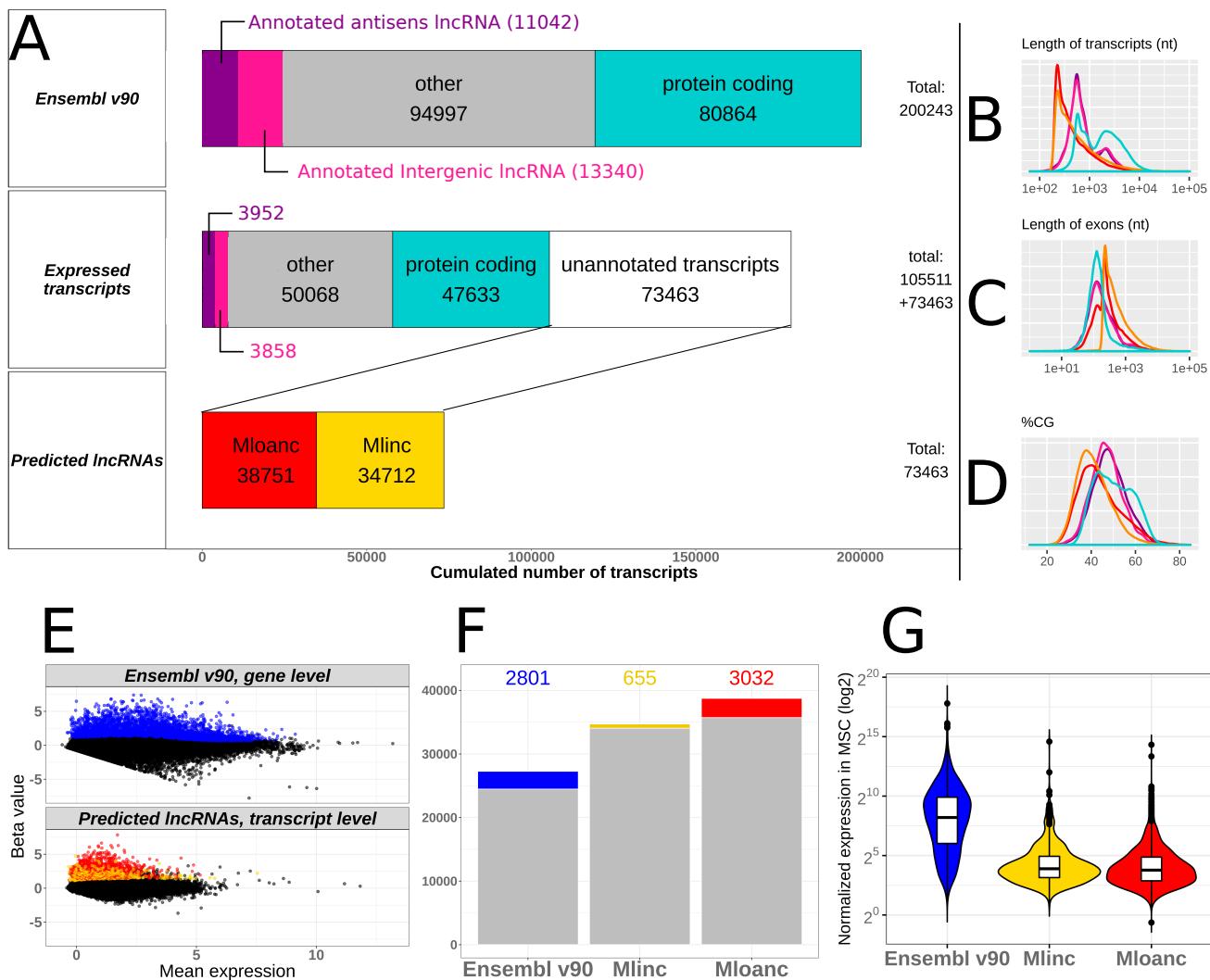


Figure 3

bioRxiv preprint doi: <https://doi.org/10.1101/2020.03.09.976001>; this version posted March 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

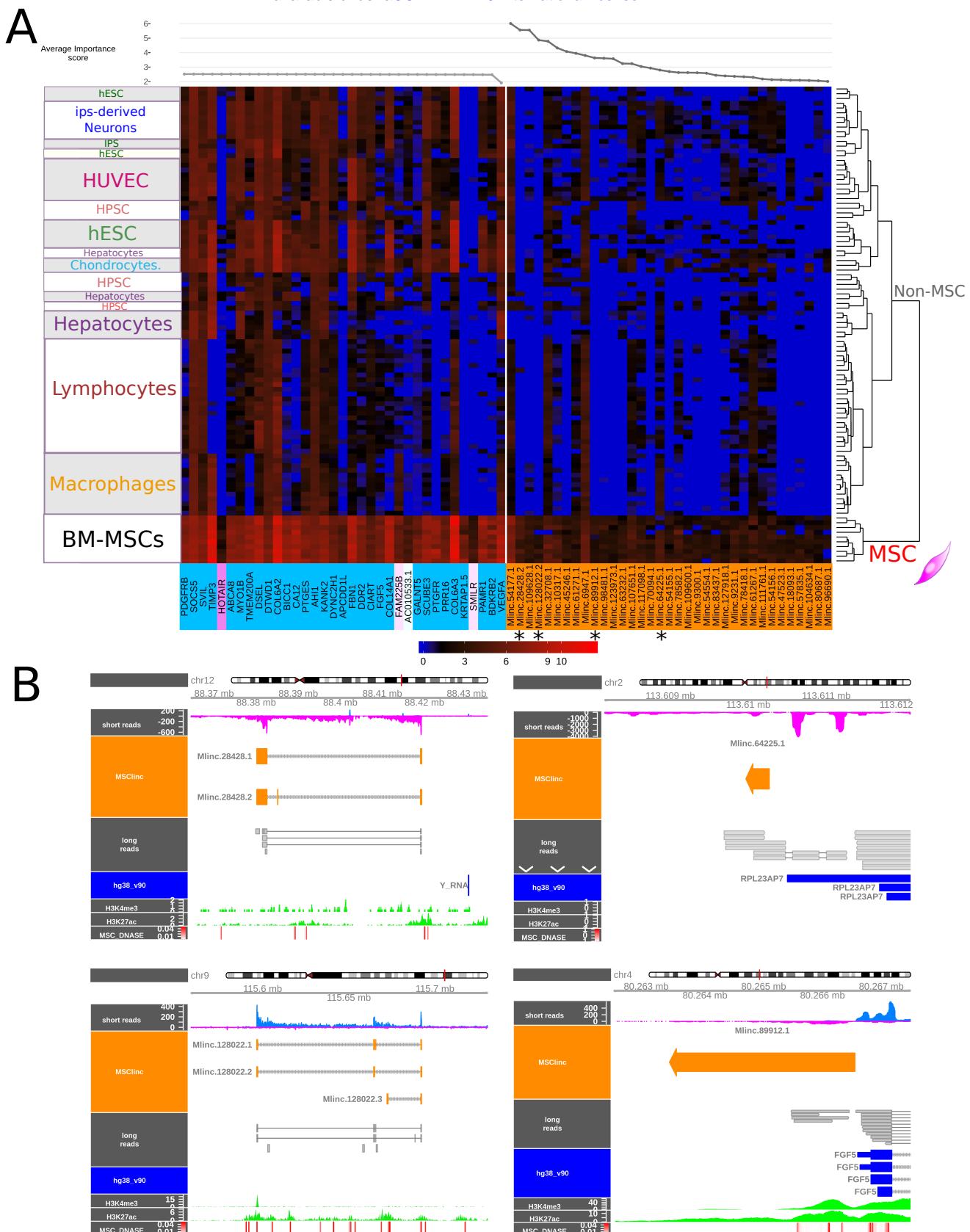


Figure 4

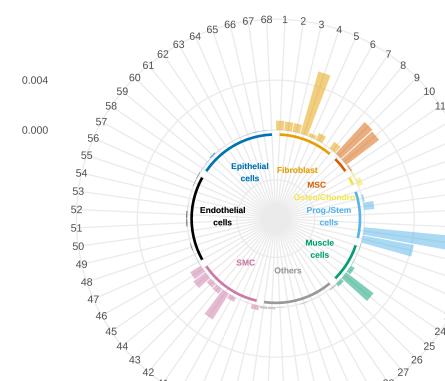
bioRxiv preprint doi: <https://doi.org/10.1101/2020.03.09.976001>; this version posted March 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

A

id	variable	n
1	bronchus fibroblast of lung	2
2	cardiac atrium fibroblast	2
3	cardiac ventricle fibroblast	2
4	fibroblast of dermis	2
5	fibroblast of lung	2
6	fibroblast of the aortic adventitia	2
7	fibroblast of villous mesenchyme	2
8	pericardium fibroblast	2
9	MSC of adipose	2
10	MSC of the bone marrow	2
11	MSC of Wharton's jelly	2
12	articular chondrocyte	2
13	osteoblast	2
14	H1-hESC	5
15	H7-hESC	2
16	hair follicle dermal papilla cell	2
17	hematopoietic multipotent progenitor cell	1
18	neural progenitor cell	2
19	skeletal muscle satellite cell	2
20	subcutaneous preadipocyte	2
21	cardiac muscle cell	2
22	myocyte	2
23	myometrial cell	2
24	myotube	2
25	regular cardiac myocyte	2
26	skeletal muscle myoblast	2
27	astrocyte	2
28	bipolar neuron	2
29	dendritic cell	20
30	foreskin keratinocyte	6
31	hair follicular keratinocyte	2
32	hepatocyte	2
33	melanocyte of skin	4
34	mesangial cell	2
35	mononuclear cell	1
36	placental pericyte	2
37	aortic SMC	2
38	bronchial smooth muscle cell	2
39	smooth muscle cell	2
40	smooth muscle cell of bladder	2
41	smooth muscle cell of the coronary artery	2
42	smooth muscle cell of the pulmonary artery	2
43	smooth muscle cell of the umbilical artery	2
44	smooth muscle cell of trachea	2
45	uterine smooth muscle cell	2
46	endothelial cell	2
47	dermis blood vessel endothelial cell	2
48	dermis lymphatic vessel endothelial cell	2
49	dermis microvascular lymphatic vessel endothelial cell	2
50	endometrial microvascular endothelial cells	2
51	endothelial cell of coronary artery	2
52	glomerular endothelial cell	2
53	lung microvascular endothelial cell	2
54	mammary microvascular endothelial cell	2
55	pulmonary artery endothelial cell	2
56	thoracic aorta endothelial cell	2
57	vein endothelial cell	2
58	airway epithelial cell	2
59	bronchial epithelial cell	2
60	epithelial cell of alveolus of lung	2
61	epithelial cell of proximal tubule	2
62	epithelial cell of umbilical artery	2
63	kidney epithelial cell	2
64	mammary epithelial cell	1
65	epithelial cell of viscerocranial mucosa	2
66	placental epithelial cell	2
67	renal cortical epithelial cell	2
68	tracheal epithelial cell	2

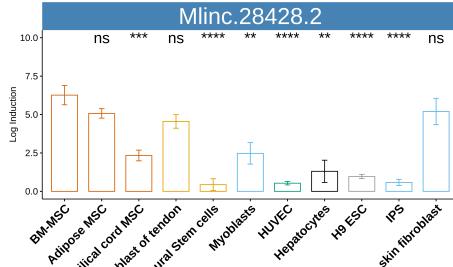
B

Mlnc.28428.2

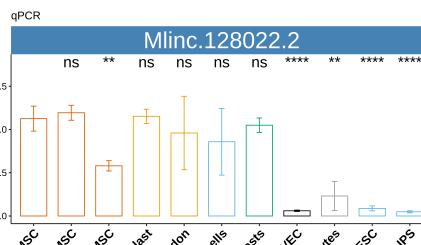
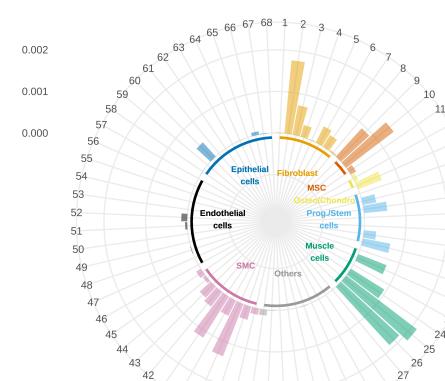


C

Mlnc.28428.2



Mlnc.128022.2



Mlnc.89912.1

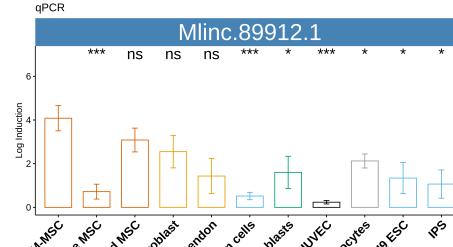
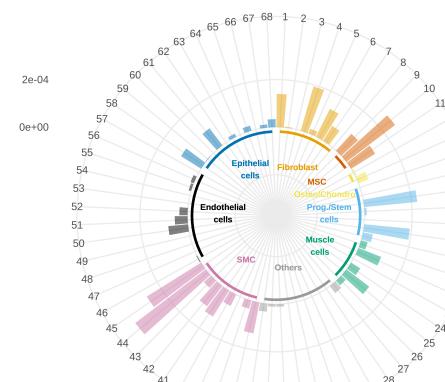
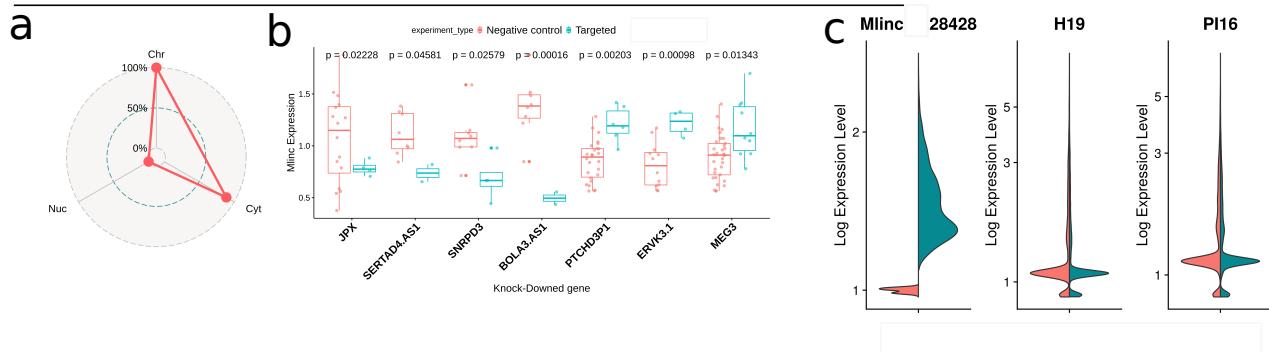


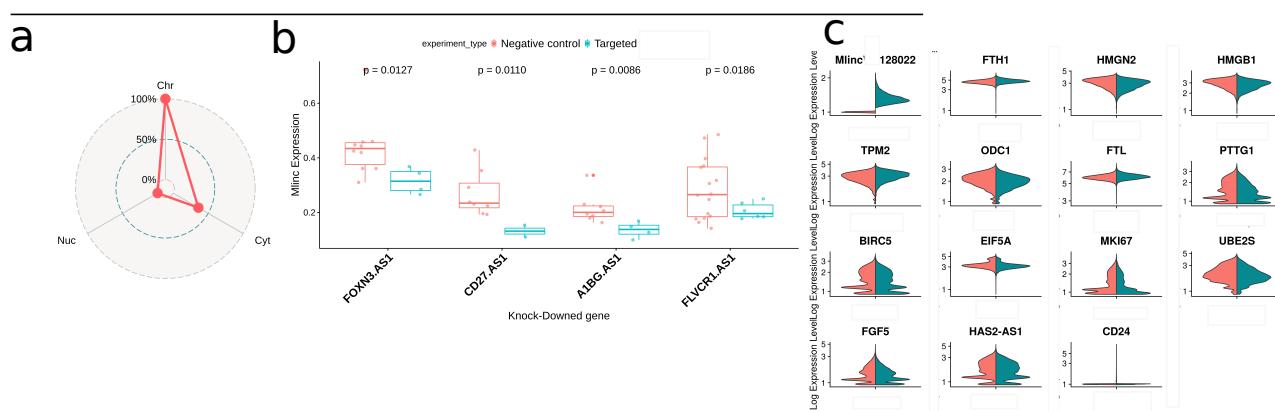
Figure 5

bioRxiv preprint doi: <https://doi.org/10.1101/2020.03.09.976001>; this version posted March 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

A: Mlinc.28428



B: Mlinc.128022



C : Mlinc.89912

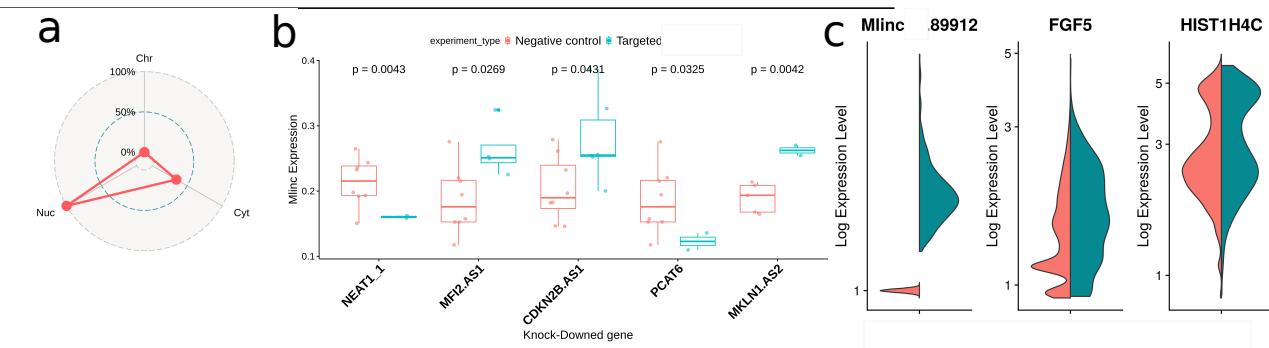
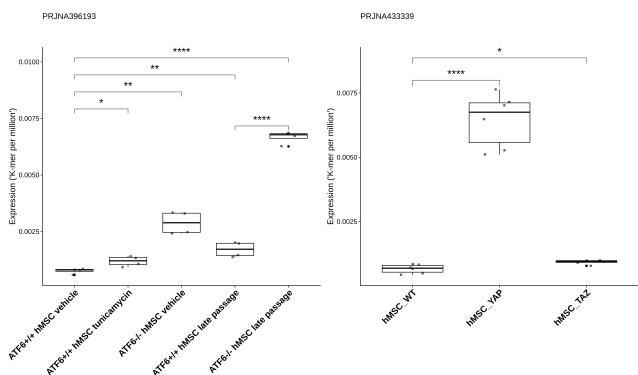


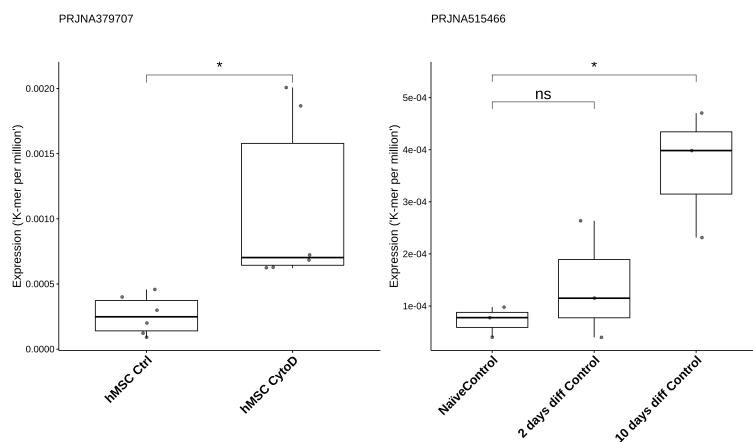
Figure 6

bioRxiv preprint doi: <https://doi.org/10.1101/2020.03.09.976001>; this version posted March 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

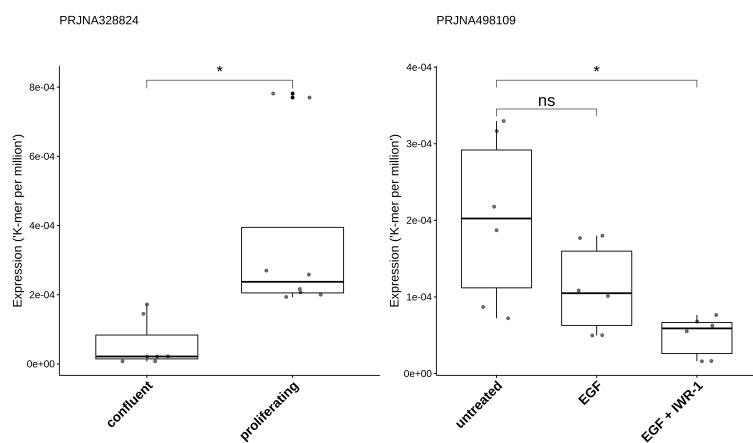
A: Mlinc.28428



B: Mlinc.128022



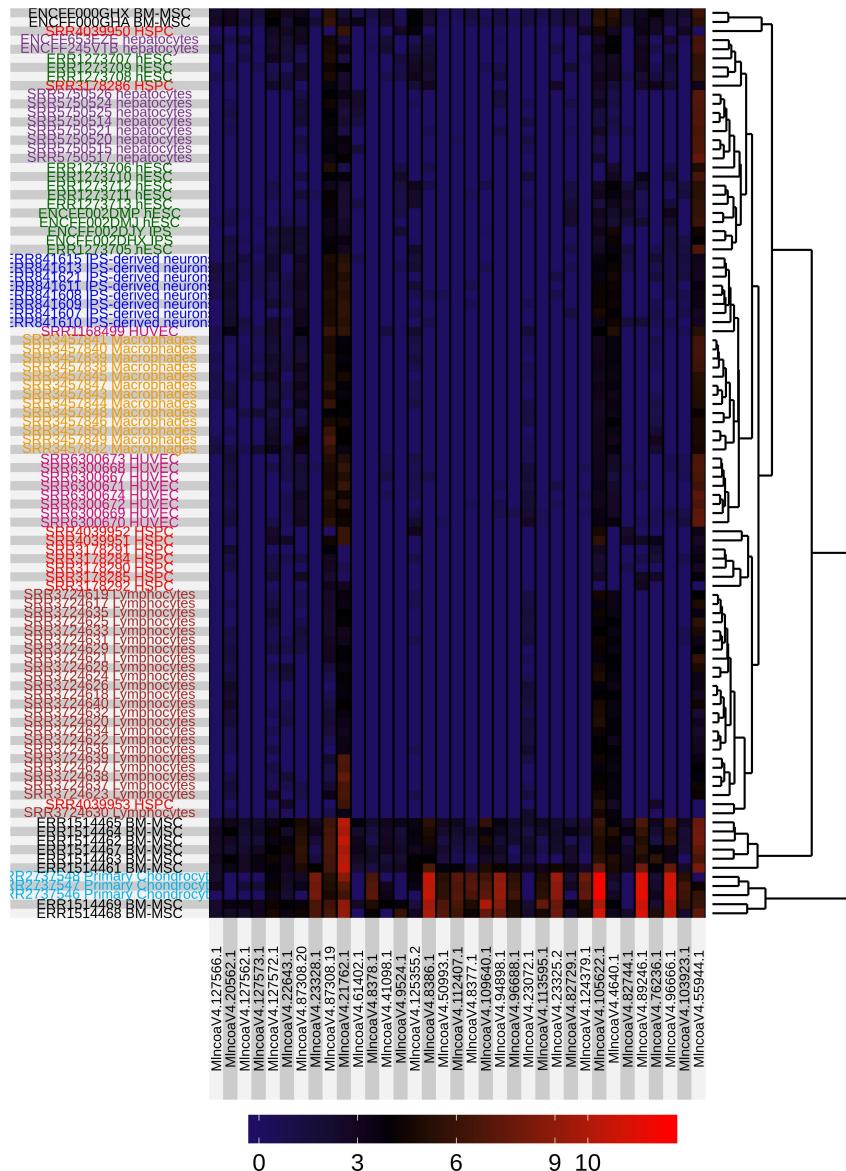
C : Mlinc.89912



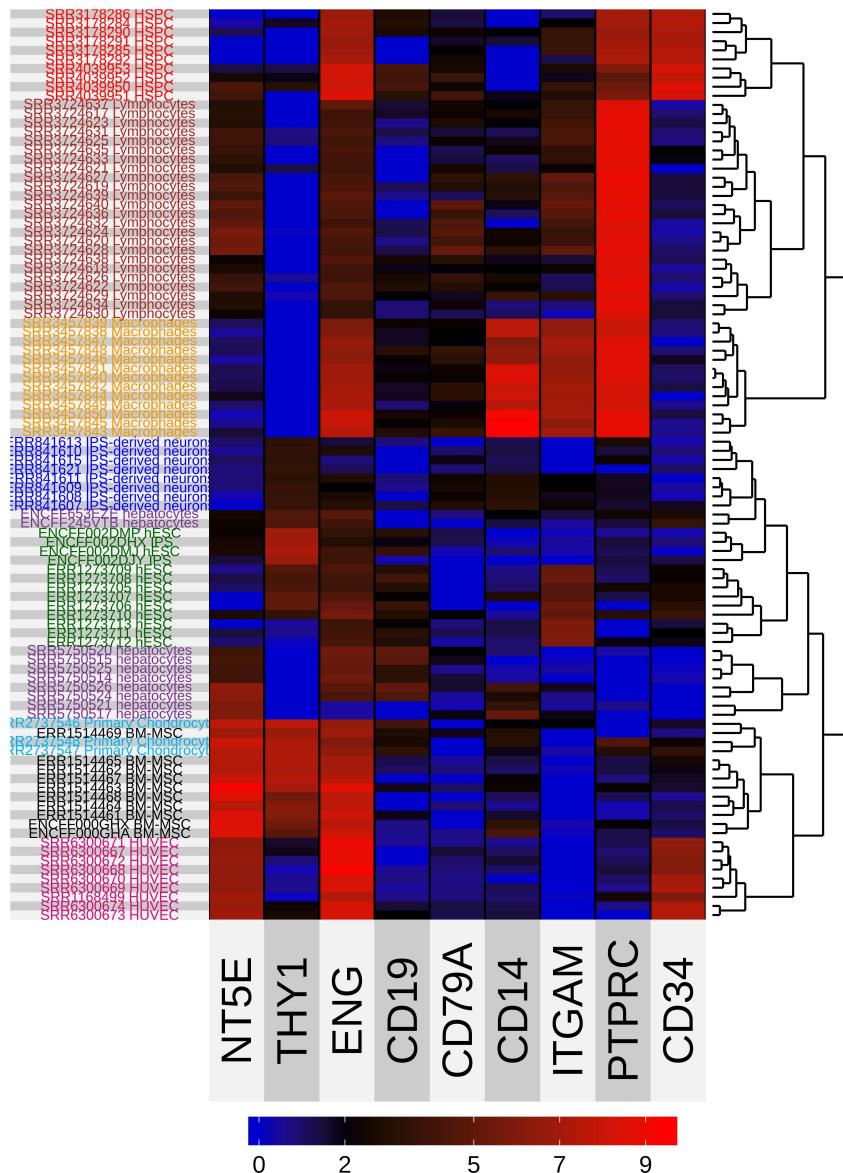
Supplementary figures

sebastien.riquier

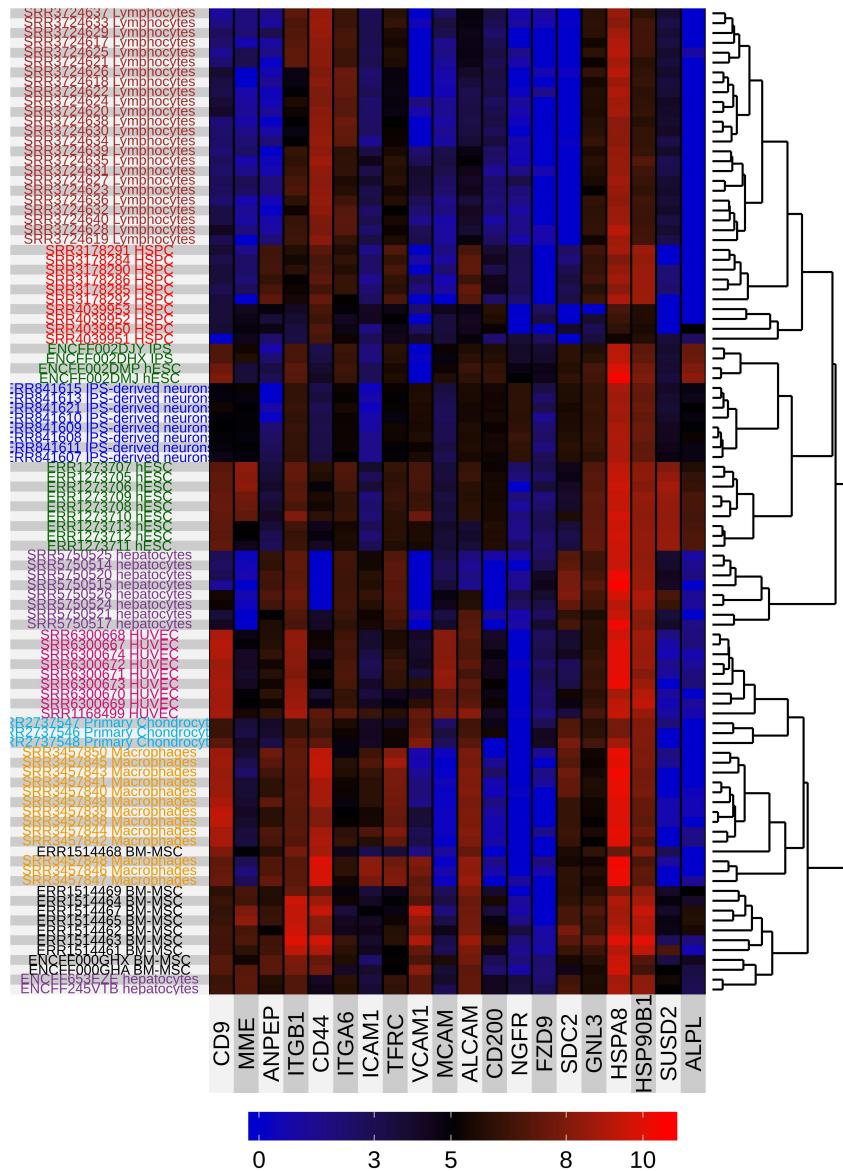
January 2020



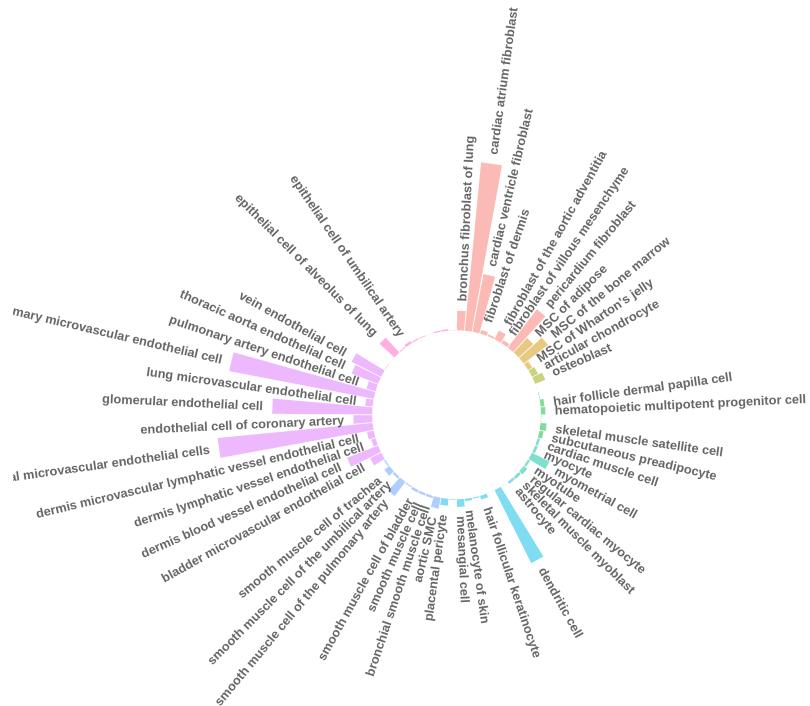
Supplementary Figure 1: Expression of Mloanc (Antisens unannotated RNA) selected after feature selection in the differential analysis cohort



Supplementary Figure 2: Expression of ISCT's MSC markers in the differential analysis cohort THY1 = CD90, NT5E = CD73, ENG = CD105, ITGAM = CD11B, PTPRC = CD45

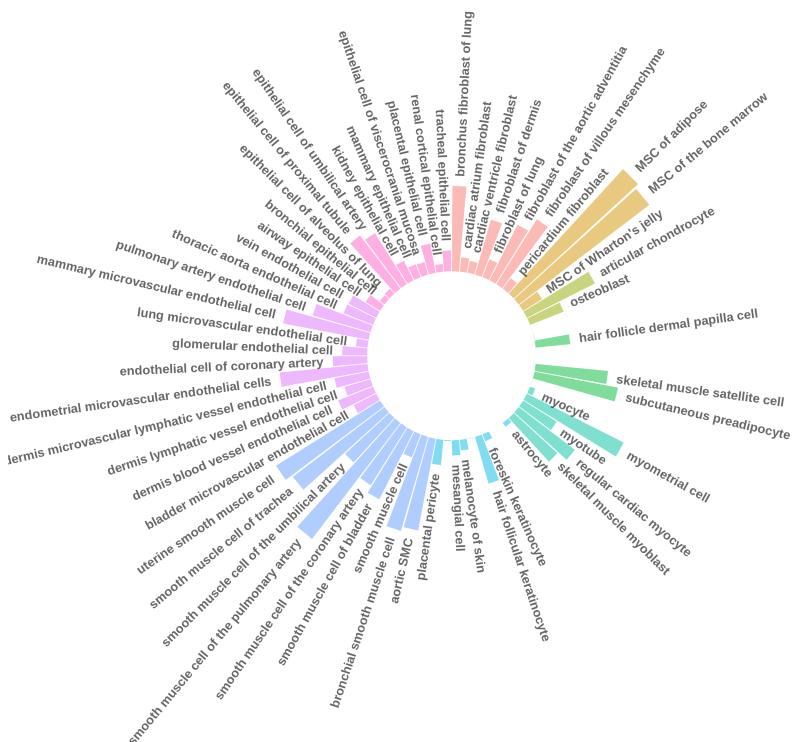


ENG-ENST00000344849



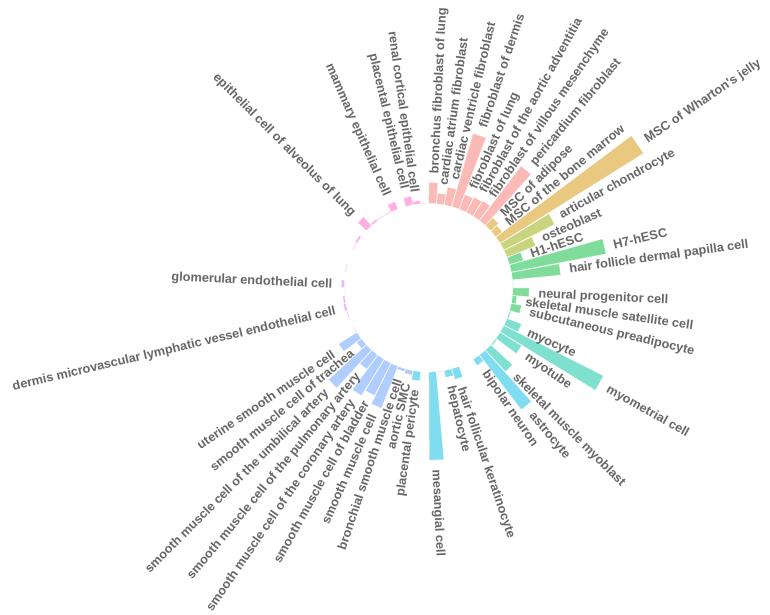
Supplementary Figure 4: Relative expression of 3 positive markers of ENG (CD105) across ENCODE's ribodepleted RNAseq datas, made by K-mer quantification, normalized in kmer by million

NT5E-ENST00000257770



Supplementary Figure 5: Relative expression of 3 positive markers of NT5E (CD73) across ENCODE's ribodepleted RNAseq datas, made by K-mer quantification, normalized in kmer by million

THY1-ENST00000284240

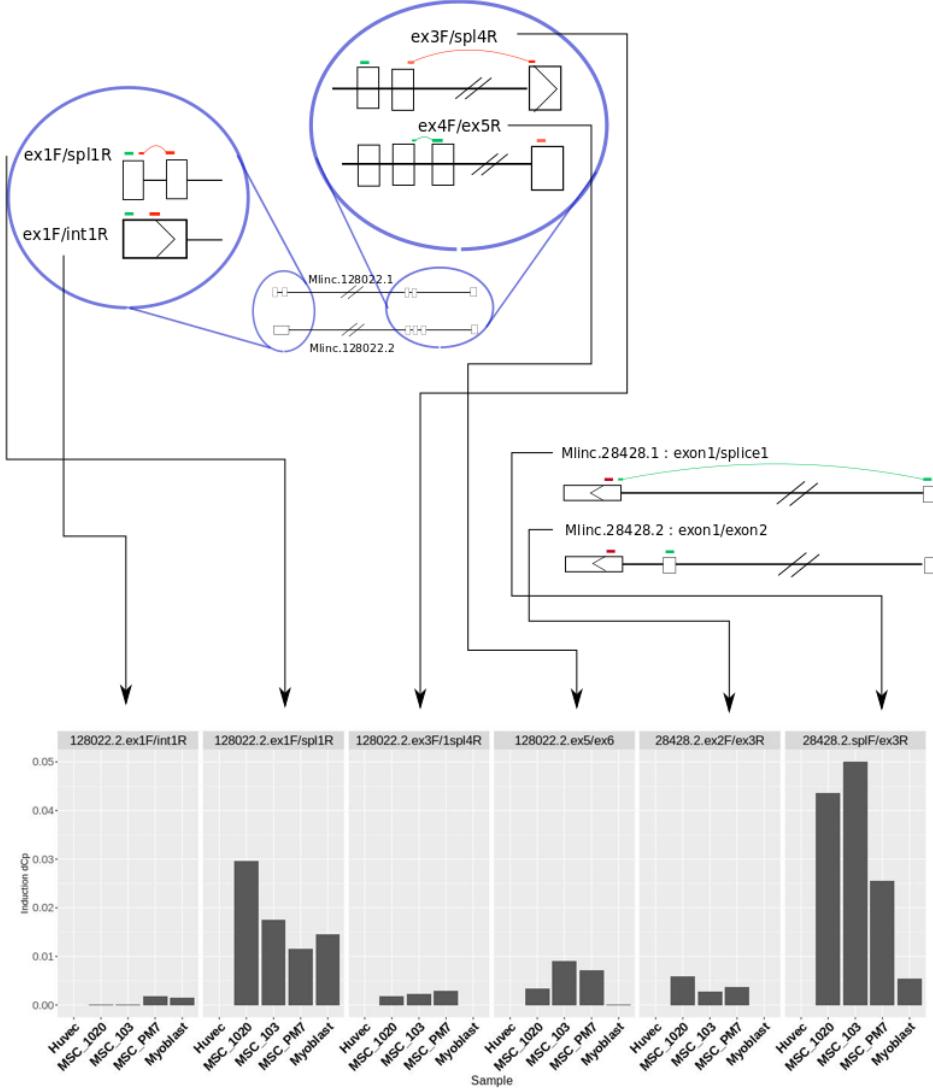


Supplementary Figure 6: Relative expression of 3 positive markers of THY1 (CD90) across ENCODE ribodepleted RNAseq datas, made by K-mer quantification, normalized in kmer by million

MlincV4.64225.1



Supplementary Figure 7: Relative expression of 3 positive markers of Mlinc.64225.1 across ENCODE's ribodepleted RNAseq datas, made by K-mer quantification, normalized in kmer by million



Supplementary Figure 8: Primer position on selected Miinc candidates and corresponding expression in MSCs, HUVECs and Myoblasts



Supplementary Figure 9: cell cycle and single cell.

Table_S1 Metadata for differential analysis

sample ID	Group	Type	SRX ID	external ID	sra sample title	specie	project title	project ID	nb total reads	length reads	detail
ENCF000GHA	MSC	BM-MSC			total RNA-seqof mesenchymal stem cell of the bone marrow	Homo sapiens	ENCODE				
ENCF000GHX	MSC	BM-MSC			total RNA-seqof mesenchymal stem cell of the bone marrow	Homo sapiens	ENCODE				
ENCF002DHX	Not	IPS			total RNA-seq GM23338	Homo sapiens	ENCODE				
ENCF002DJY	Not	IPS			total RNA-seq GM23338	Homo sapiens	ENCODE				
ENCF002DMJ	Not	hESC			total RNA-seq H7	Homo sapiens	ENCODE				
ENCF002DMP	Not	hESC			total RNA-seq H7	Homo sapiens	ENCODE				
ERR1273705	Not	hESC	ERX1345312		Transcriptional profiling of human naive embryonic stem cells	Homo sapiens	Transcriptional profiling of human naive embryonic stem cells	PRJEB12748	67286349	125	
ERR1273706	Not	hESC	ERX1345313		Transcriptional profiling of human naive embryonic stem cells	Homo sapiens	Transcriptional profiling of human naive embryonic stem cells	PRJEB12748	18924239	125	
ERR1273707	Not	hESC	ERX1345314		Transcriptional profiling of human naive embryonic stem cells	Homo sapiens	Transcriptional profiling of human naive embryonic stem cells	PRJEB12748	55745935	125	
ERR1273708	Not	hESC	ERX1345315		Transcriptional profiling of human naive embryonic stem cells	Homo sapiens	Transcriptional profiling of human naive embryonic stem cells	PRJEB12748	28988509	125	
ERR1273709	Not	hESC	ERX1345316		Transcriptional profiling of human naive embryonic stem cells	Homo sapiens	Transcriptional profiling of human naive embryonic stem cells	PRJEB12748	NA	NA	
ERR1273710	Not	hESC	ERX1345317		Transcriptional profiling of human naive embryonic stem cells	Homo sapiens	Transcriptional profiling of human naive embryonic stem cells	PRJEB12748	NA	NA	
ERR1273711	Not	hESC	ERX1345318		Transcriptional profiling of human naive embryonic stem cells	Homo sapiens	Transcriptional profiling of human naive embryonic stem cells	PRJEB12748	NA	NA	
ERR1273712	Not	hESC	ERX1345319		Transcriptional profiling of human naive embryonic stem cells	Homo sapiens	Transcriptional profiling of human naive embryonic stem cells	PRJEB12748	NA	NA	
ERR1273713	Not	hESC	ERX1345320		Transcriptional profiling of human naive embryonic stem cells	Homo sapiens	Transcriptional profiling of human naive embryonic stem cells	PRJEB12748	NA	NA	
ERR1514461	MSC	BM-MSC	ERX1585480		RNASeq of musculoskeletal ageing using human mesenchymal stem cells and their tissue constructs	Homo sapiens	RNASeq of musculoskeletal ageing using human mesenchymal stem cells and their tissue constructs	PRJEB14688	32467641	96.41	
ERR1514462	MSC	BM-MSC	ERX1585481		RNASeq of musculoskeletal ageing using human mesenchymal stem cells and their tissue constructs	Homo sapiens	RNASeq of musculoskeletal ageing using human mesenchymal stem cells and their tissue constructs	PRJEB14688	34192962	96.30	
ERR1514463	MSC	BM-MSC	ERX1585482		RNASeq of musculoskeletal ageing using human mesenchymal stem cells and their tissue constructs	Homo sapiens	RNASeq of musculoskeletal ageing using human mesenchymal stem cells and their tissue constructs	PRJEB14688	28164147	95.51	
ERR1514464	MSC	BM-MSC	ERX1585483		RNASeq of musculoskeletal ageing using human mesenchymal stem cells and their tissue constructs	Homo sapiens	RNASeq of musculoskeletal ageing using human mesenchymal stem cells and their tissue constructs	PRJEB14688	28633110	96.97	
ERR1514465	MSC	BM-MSC	ERX1585484		RNASeq of musculoskeletal ageing using human mesenchymal stem cells and their tissue constructs	Homo sapiens	RNASeq of musculoskeletal ageing using human mesenchymal stem cells and their tissue constructs	PRJEB14688	25135660	96.96	
ERR1514467	MSC	BM-MSC	ERX1585486		RNASeq of musculoskeletal ageing using human mesenchymal stem cells and their tissue constructs	Homo sapiens	RNASeq of musculoskeletal ageing using human mesenchymal stem cells and their tissue constructs	PRJEB14688	28240410	96.79	
ERR1514468	MSC	BM-MSC	ERX1585487		RNASeq of musculoskeletal ageing using human mesenchymal stem cells and their tissue constructs	Homo sapiens	RNASeq of musculoskeletal ageing using human mesenchymal stem cells and their tissue constructs	PRJEB14688	29410953	96.97	
ERR1514469	MSC	BM-MSC	ERX1585488		RNASeq of musculoskeletal ageing using human mesenchymal stem cells and their tissue constructs	Homo sapiens	RNASeq of musculoskeletal ageing using human mesenchymal stem cells and their tissue constructs	PRJEB14688	28966083	97.22	
ERR841607	Not	IPS-derived neurons	ERX921697		illumina HiSeq 2500 paired end sequencing; RNA-seq of NEAT1 knockdown in neuronal excitability	Homo sapiens	RNA-seq of NEAT1 knockdown in neuronal excitability	PRJEB9006	14282989	101	KCI
ERR841608	Not	IPS-derived neurons	ERX921709		illumina HiSeq 2500 paired end sequencing; RNA-seq of NEAT1 knockdown in neuronal excitability	Homo sapiens	RNA-seq of NEAT1 knockdown in neuronal excitability	PRJEB9006	22188866	101	KCI
ERR841609	Not	IPS-derived neurons	ERX921694		illumina HiSeq 2500 paired end sequencing; RNA-seq of NEAT1 knockdown in neuronal excitability	Homo sapiens	RNA-seq of NEAT1 knockdown in neuronal excitability	PRJEB9006	22616373	101	KCI
ERR841610	Not	IPS-derived neurons	ERX921708		illumina HiSeq 2500 paired end sequencing; RNA-seq of NEAT1 knockdown in neuronal excitability	Homo sapiens	RNA-seq of NEAT1 knockdown in neuronal excitability	PRJEB9006	13192833	101	NA
ERR841611	Not	IPS-derived neurons	ERX921705		illumina HiSeq 2500 paired end sequencing; RNA-seq of NEAT1 knockdown in neuronal excitability	Homo sapiens	RNA-seq of NEAT1 knockdown in neuronal excitability	PRJEB9006	15290840	101	KCI
ERR841613	Not	IPS-derived neurons	ERX921699		illumina HiSeq 2500 paired end sequencing; RNA-seq of NEAT1 knockdown in neuronal excitability	Homo sapiens	RNA-seq of NEAT1 knockdown in neuronal excitability	PRJEB9006	15686003	101	NA
ERR841615	Not	IPS-derived neurons	ERX921702		illumina HiSeq 2500 paired end sequencing; RNA-seq of NEAT1 knockdown in neuronal excitability	Homo sapiens	RNA-seq of NEAT1 knockdown in neuronal excitability	PRJEB9006	14182869	101	NA
ERR841621	Not	IPS-derived neurons	ERX921696		illumina HiSeq 2500 paired end sequencing; RNA-seq of NEAT1 knockdown in neuronal excitability	Homo sapiens	RNA-seq of NEAT1 knockdown in neuronal excitability	PRJEB9006	15232542	101	NA
SRR2737546	Not	Primary Chondrocytes	SRX1358410		GSM1914777: Control 1; Homo sapiens; RNA-Seq	Homo sapiens	Transcriptome response to 4h IL-1b stimulation of primary chondrocytes	PRJNA299568	34300359	97.39	No stimulation
SRR2737547	Not	Primary Chondrocytes	SRX1358411		GSM1914778: Control 2; Homo sapiens; RNA-Seq	Homo sapiens	Transcriptome response to 4h IL-1b stimulation of primary chondrocytes	PRJNA299568	37284758	96.91	No stimulation
SRR2737548	Not	Primary Chondrocytes	SRX1358412		GSM1914779: Control 3; Homo sapiens; RNA-Seq	Homo sapiens	Transcriptome response to 4h IL-1b stimulation of primary chondrocytes	PRJNA299568	33144961	96.89	No stimulation
SRR4039950	Not	HSPC	SRX2031066		GSM2284651: HSPC day 0 rep1; Homo sapiens; RNA-Seq	Homo sapiens	Hematopoietic Stem/Progenitor Cell Expansion Without Differentiation is Achieved in Zwitterionic Hydrgols	PRJNA339416	17369972	50	
SRR4039951	Not	HSPC	SRX2031067		GSM2284652: HSPC day 0 rep2; Homo sapiens; RNA-Seq	Homo sapiens	Hematopoietic Stem/Progenitor Cell Expansion Without Differentiation is Achieved in Zwitterionic Hydrgols	PRJNA339416	18053125	50	
SRR4039952	Not	HSPC	SRX2031068		GSM2284653: HSPC day 24 rep1; Homo sapiens; RNA-Seq	Homo sapiens	Hematopoietic Stem/Progenitor Cell Expansion Without Differentiation is Achieved in Zwitterionic Hydrgols	PRJNA339416	20682268	50	
SRR4039953	Not	HSPC	SRX2031069		GSM2284654: HSPC day 24 rep2; Homo sapiens; RNA-Seq	Homo sapiens	Hematopoietic Stem/Progenitor Cell Expansion Without Differentiation is Achieved in Zwitterionic Hydrgols	PRJNA339416	19282208	50	
SRR3178284	Not	HSPC	SRX1592574		GSM2066393: huCB stem ctrl_1rep1; Homo sapiens; RNA-Seq	Homo sapiens	miR-126 Orchestrates an Oncogenic Program in B-Cell Precursor Acute Lymphoblastic Leukemia	PRJNA312538	25398470	100.79	huCB stem ctrl
SRR3178285	Not	HSPC	SRX1592575		GSM2066394: huCB stem ctrl_1rep2; Homo sapiens; RNA-Seq	Homo sapiens	miR-126 Orchestrates an Oncogenic Program in B-Cell Precursor Acute Lymphoblastic Leukemia	PRJNA312538	5926147	100.77	huCB stem ctrl
SRR3178286	Not	HSPC	SRX1592576		GSM2066395: huCB stem ctrl_1rep3; Homo sapiens; RNA-Seq	Homo sapiens	miR-126 Orchestrates an Oncogenic Program in B-Cell Precursor Acute Lymphoblastic Leukemia	PRJNA312538	7615479	100.78	huCB stem ctrl
SRR3178290	Not	HSPC	SRX1592580		GSM2066399: huCB progenitors ctrl_1rep1; Homo sapiens; RNA-Seq	Homo sapiens	miR-126 Orchestrates an Oncogenic Program in B-Cell Precursor Acute Lymphoblastic Leukemia	PRJNA312538	36341406	100.77	huCB progenitors ctrl
SRR3178291	Not	HSPC	SRX1592581		GSM2066400: huCB progenitors ctrl_1rep2; Homo sapiens; RNA-Seq	Homo sapiens	miR-126 Orchestrates an Oncogenic Program in B-Cell Precursor Acute Lymphoblastic Leukemia	PRJNA312538	6431588	100.78	huCB progenitors ctrl
SRR3178292	Not	HSPC	SRX1592582		GSM2066401: huCB progenitors ctrl_1rep3; Homo sapiens; RNA-Seq	Homo sapiens	miR-126 Orchestrates an Oncogenic Program in B-Cell Precursor Acute Lymphoblastic Leukemia	PRJNA312538	3602619	100.77	huCB progenitors ctrl
ENCF245VTB	Not	hepatocytes			total RNA-seqof hepatocyte	Homo sapiens	ENCODE				
ENCF653EZE	Not	hepatocytes			total RNA-seqof hepatocyte	Homo sapiens	ENCODE				
SRR1168499	Not	HUVEC	SRX470246		GSM1327344: HUVEC; Homo sapiens; RNA-Seq	Homo sapiens	Transcriptomic analysis reveals novel long non-coding RNAs critical for vertebrate development [RNA-Seq]	PRJNA238180	97606136	90	
SRR3457838	Not	Macrophages	SRX1733620		GSM2135456: 140421_S2_CNT_A ; Homo sapiens; RNA-Seq	Homo sapiens	Single-cell qPCR, and bulk-level transcriptomics and epigenomics of treated macrophages	PRJNA315538	62681472	50	CNT
SRR3457839	Not	Macrophages	SRX1733621		GSM2135457: 140421_S2_CNT_B ; Homo sapiens; RNA-Seq	Homo sapiens	Single-cell qPCR, and bulk-level transcriptomics and epigenomics of treated macrophages	PRJNA315538	66405838	50	CNT
SRR3457840	Not	Macrophages	SRX1733622		GSM2135458: 140421_S2_IL10_A ; Homo sapiens; RNA-Seq	Homo sapiens	Single-cell qPCR, and bulk-level transcriptomics and epigenomics of treated macrophages	PRJNA315538	70016991	50	IL10
SRR3457841	Not	Macrophages	SRX1733623		GSM2135459: 140421_S2_IL10_B ; Homo sapiens; RNA-Seq	Homo sapiens	Single-cell qPCR, and bulk-level transcriptomics and epigenomics of treated macrophages	PRJNA315538	80257003	50	IL10
SRR3457842	Not	Macrophages	SRX1733624		GSM2135460: 140425_S3_CNT ; Homo sapiens; RNA-Seq	Homo sapiens	Single-cell qPCR, and bulk-level transcriptomics and epigenomics of treated macrophages	PRJNA315538	37671849	75	CNT

Table_S1 Metadata for differential analysis

SRR3457843	Not	Macrophages	SRX1733625	GSM2135461: 140825_S3_IL10_ ; Homo sapiens; RNA-Seq	Homo sapiens	Single-cell qPCR, and bulk-level transcriptomics and epigenomics of treated macrophages	PRJNA315538	40575802	75 IL10	NA
SRR3457844	Not	Macrophages	SRX1733626	GSM2135462: 140825_S1_CNT_ ; Homo sapiens; RNA-Seq	Homo sapiens	Single-cell qPCR, and bulk-level transcriptomics and epigenomics of treated macrophages	PRJNA315538	28548046	75 CNT	NA
SRR3457845	Not	Macrophages	SRX1733627	GSM2135463: 140825_S1_IL10_ ; Homo sapiens; RNA-Seq	Homo sapiens	Single-cell qPCR, and bulk-level transcriptomics and epigenomics of treated macrophages	PRJNA315538	44848576	75 IL10	NA
SRR3457846	Not	Macrophages	SRX1733628	GSM2135464: 140825_S1_IFNgTNFa_ ; Homo sapiens; RNA-Seq	Homo sapiens	Single-cell qPCR, and bulk-level transcriptomics and epigenomics of treated macrophages	PRJNA315538	33920601	75 IFNgTNFa	NA
SRR3457847	Not	Macrophages	SRX1733629	GSM2135465: 140825_S2_IFNgTNFa_ ; Homo sapiens; RNA-Seq	Homo sapiens	Single-cell qPCR, and bulk-level transcriptomics and epigenomics of treated macrophages	PRJNA315538	26580763	75 IFNgTNFa	NA
SRR3457848	Not	Macrophages	SRX1733630	GSM2135466: 140825_S3_IFNgTNFa_ ; Homo sapiens; RNA-Seq	Homo sapiens	Single-cell qPCR, and bulk-level transcriptomics and epigenomics of treated macrophages	PRJNA315538	28565135	75 IFNgTNFa	NA
SRR3457849	Not	Macrophages	SRX1733631	GSM2135467: 140825_S2_CNT_ ; Homo sapiens; RNA-Seq	Homo sapiens	Single-cell qPCR, and bulk-level transcriptomics and epigenomics of treated macrophages	PRJNA315538	41595399	75 CNT	NA
SRR3457850	Not	Macrophages	SRX1733632	GSM2135468: 140825_S2_IL10_ ; Homo sapiens; RNA-Seq	Homo sapiens	Single-cell qPCR, and bulk-level transcriptomics and epigenomics of treated macrophages	PRJNA315538	38443748	75 IL10	NA
SRR5750514	Not	hepatocytes	SRX2950738	Dual RNAseq: Human hepatocytes and Plasmodium berghei	Homo sapiens	Dual RNAseq - Human hepatocytes infected with Plasmodium berghei	PRJNA390648	62983452	99.38	Infected with Plasmodium berghei
SRR5750515	Not	hepatocytes	SRX2950737	Dual RNAseq: Human hepatocytes and Plasmodium berghei	Homo sapiens	Dual RNAseq - Human hepatocytes infected with Plasmodium berghei	PRJNA390648	46318597	99.38	NA
SRR5750517	Not	hepatocytes	SRX2950735	Dual RNAseq: Human hepatocytes and Plasmodium berghei	Homo sapiens	Dual RNAseq - Human hepatocytes infected with Plasmodium berghei	PRJNA390648	63391464	99.29	NA
SRR5750520	Not	hepatocytes	SRX2950732	Dual RNAseq: Human hepatocytes and Plasmodium berghei	Homo sapiens	Dual RNAseq - Human hepatocytes infected with Plasmodium berghei	PRJNA390648	57010988	99.33	NA
SRR5750521	Not	hepatocytes	SRX2950731	Dual RNAseq: Human hepatocytes and Plasmodium berghei	Homo sapiens	Dual RNAseq - Human hepatocytes infected with Plasmodium berghei	PRJNA390648	47651456	99.41	Infected with Plasmodium berghei
SRR5750524	Not	hepatocytes	SRX2950728	Dual RNAseq: Human hepatocytes and Plasmodium berghei	Homo sapiens	Dual RNAseq - Human hepatocytes infected with Plasmodium berghei	PRJNA390648	47486674	99.53	NA
SRR5750525	Not	hepatocytes	SRX2950727	Dual RNAseq: Human hepatocytes and Plasmodium berghei	Homo sapiens	Dual RNAseq - Human hepatocytes infected with Plasmodium berghei	PRJNA390648	52736152	99.44	Infected with Plasmodium berghei
SRR5750526	Not	hepatocytes	SRX2950726	Dual RNAseq: Human hepatocytes and Plasmodium berghei	Homo sapiens	Dual RNAseq - Human hepatocytes infected with Plasmodium berghei	PRJNA390648	49216420	99.16	NA
SRR6300667	Not	HUVEC	SRX3401581	GSM2859866: hyp_12h_n1_1; Homo sapiens; RNA-Seq	Homo sapiens	Total RNA deep sequencing (ribosomal depleted) of human umbilical vein endothelial cells exposed to hypoxia (0.2%) for 12h and 24h or kept under normoxic conditions.	PRJNA418883	39712730	94	Hypoxia 0.2% O2, 5% CO2 humidified atmosphere
SRR6300668	Not	HUVEC	SRX3401582	GSM2859867: hyp_12h_n2_1; Homo sapiens; RNA-Seq	Homo sapiens	Total RNA deep sequencing (ribosomal depleted) of human umbilical vein endothelial cells exposed to hypoxia (0.2%) for 12h and 24h or kept under normoxic conditions.	PRJNA418883	43357452	94	Hypoxia 0.2% O2, 5% CO2 humidified atmosphere
SRR6300669	Not	HUVEC	SRX3401583	GSM2859868: hyp_24h_n1_1; Homo sapiens; RNA-Seq	Homo sapiens	Total RNA deep sequencing (ribosomal depleted) of human umbilical vein endothelial cells exposed to hypoxia (0.2%) for 12h and 24h or kept under normoxic conditions.	PRJNA418883	41045511	94	Hypoxia 0.2% O2, 5% CO2 humidified atmosphere
SRR6300670	Not	HUVEC	SRX3401584	GSM2859869: hyp_24h_n2_1; Homo sapiens; RNA-Seq	Homo sapiens	Total RNA deep sequencing (ribosomal depleted) of human umbilical vein endothelial cells exposed to hypoxia (0.2%) for 12h and 24h or kept under normoxic conditions.	PRJNA418883	44557122	94	Hypoxia 0.2% O2, 5% CO2 humidified atmosphere
SRR6300671	Not	HUVEC	SRX3401585	GSM2859870: norm_12h_n1_1; Homo sapiens; RNA-Seq	Homo sapiens	Total RNA deep sequencing (ribosomal depleted) of human umbilical vein endothelial cells exposed to hypoxia (0.2%) for 12h and 24h or kept under normoxic conditions.	PRJNA418883	42304183	94	Normoxia 20% O2, 5% CO2 humidified atmosphere
SRR6300672	Not	HUVEC	SRX3401586	GSM2859871: norm_12h_n2_1; Homo sapiens; RNA-Seq	Homo sapiens	Total RNA deep sequencing (ribosomal depleted) of human umbilical vein endothelial cells exposed to hypoxia (0.2%) for 12h and 24h or kept under normoxic conditions.	PRJNA418883	42748680	94	Normoxia 20% O2, 5% CO2 humidified atmosphere
SRR6300673	Not	HUVEC	SRX3401587	GSM2859872: norm_24h_n1_1; Homo sapiens; RNA-Seq	Homo sapiens	Total RNA deep sequencing (ribosomal depleted) of human umbilical vein endothelial cells exposed to hypoxia (0.2%) for 12h and 24h or kept under normoxic conditions.	PRJNA418883	42771271	94	Normoxia 20% O2, 5% CO2 humidified atmosphere
SRR6300674	Not	HUVEC	SRX3401588	GSM2859873: norm_24h_n2_1; Homo sapiens; RNA-Seq	Homo sapiens	Total RNA deep sequencing (ribosomal depleted) of human umbilical vein endothelial cells exposed to hypoxia (0.2%) for 12h and 24h or kept under normoxic conditions.	PRJNA418883	42131852	94	Normoxia 20% O2, 5% CO2 humidified atmosphere
SRR3724617	Not	Lymphocytes	SRX1881712	GSM2218995: Healthy1-CD4Memory; Homo sapiens; RNA-Seq	Homo sapiens	T cells from paroxysmal nocturnal hemoglobinuria patients show an altered TNFR signaling pathway	PRJNA327044	29296358	100	healthy
SRR3724618	Not	Lymphocytes	SRX1881713	GSM2218996: Healthy1-CD4Naive; Homo sapiens; RNA-Seq	Homo sapiens	T cells from paroxysmal nocturnal hemoglobinuria patients show an altered TNFR signaling pathway	PRJNA327044	27598581	100	healthy
SRR3724619	Not	Lymphocytes	SRX1881714	GSM2218997: Healthy1-CD8Memory; Homo sapiens; RNA-Seq	Homo sapiens	T cells from paroxysmal nocturnal hemoglobinuria patients show an altered TNFR signaling pathway	PRJNA327044	26351498	100	healthy
SRR3724620	Not	Lymphocytes	SRX1881715	GSM2218998: Healthy1-CD8Naive; Homo sapiens; RNA-Seq	Homo sapiens	T cells from paroxysmal nocturnal hemoglobinuria patients show an altered TNFR signaling pathway	PRJNA327044	29464342	100	healthy
SRR3724621	Not	Lymphocytes	SRX1881716	GSM2218999: Healthy2-CD4Memory; Homo sapiens; RNA-Seq	Homo sapiens	T cells from paroxysmal nocturnal hemoglobinuria patients show an altered TNFR signaling pathway	PRJNA327044	48167581	100	healthy
SRR3724622	Not	Lymphocytes	SRX1881717	GSM2219000: Healthy2-CD4Naive; Homo sapiens; RNA-Seq	Homo sapiens	T cells from paroxysmal nocturnal hemoglobinuria patients show an altered TNFR signaling pathway	PRJNA327044	42220340	100	healthy
SRR3724623	Not	Lymphocytes	SRX1881718	GSM2219001: Healthy2-CD8Memory; Homo sapiens; RNA-Seq	Homo sapiens	T cells from paroxysmal nocturnal hemoglobinuria patients show an altered TNFR signaling pathway	PRJNA327044	38563382	100	healthy
SRR3724624	Not	Lymphocytes	SRX1881719	GSM2219002: Healthy2-CD8Naive; Homo sapiens; RNA-Seq	Homo sapiens	T cells from paroxysmal nocturnal hemoglobinuria patients show an altered TNFR signaling pathway	PRJNA327044	45267949	100	healthy
SRR3724625	Not	Lymphocytes	SRX1881720	GSM2219003: Healthy3-CD4Memory; Homo sapiens; RNA-Seq	Homo sapiens	T cells from paroxysmal nocturnal hemoglobinuria patients show an altered TNFR signaling pathway	PRJNA327044	45930063	100	healthy
SRR3724626	Not	Lymphocytes	SRX1881721	GSM2219004: Healthy3-CD4Naive; Homo sapiens; RNA-Seq	Homo sapiens	T cells from paroxysmal nocturnal hemoglobinuria patients show an altered TNFR signaling pathway	PRJNA327044	40731526	100	healthy
SRR3724627	Not	Lymphocytes	SRX1881722	GSM2219005: Healthy3-CD8Memory; Homo sapiens; RNA-Seq	Homo sapiens	T cells from paroxysmal nocturnal hemoglobinuria patients show an altered TNFR signaling pathway	PRJNA327044	47542258	100	healthy
SRR3724628	Not	Lymphocytes	SRX1881723	GSM2219006: Healthy3-CD8Naive; Homo sapiens; RNA-Seq	Homo sapiens	T cells from paroxysmal nocturnal hemoglobinuria patients show an altered TNFR signaling pathway	PRJNA327044	53718245	100	healthy
SRR3724629	Not	Lymphocytes	SRX1881724	GSM2219007: PNH1-CD4Memory; Homo sapiens; RNA-Seq	Homo sapiens	T cells from paroxysmal nocturnal hemoglobinuria patients show an altered TNFR signaling pathway	PRJNA327044	30025375	100	paroxysmal nocturnal hemoglobinuria (PNH)
SRR3724630	Not	Lymphocytes	SRX1881725	GSM2219008: PNH1-CD4Naive; Homo sapiens; RNA-Seq	Homo sapiens	T cells from paroxysmal nocturnal hemoglobinuria patients show an altered TNFR signaling pathway	PRJNA327044	29638535	100	paroxysmal nocturnal hemoglobinuria (PNH)
SRR3724631	Not	Lymphocytes	SRX1881726	GSM2219009: PNH1-CD8Memory; Homo sapiens; RNA-Seq	Homo sapiens	T cells from paroxysmal nocturnal hemoglobinuria patients show an altered TNFR signaling pathway	PRJNA327044	30553368	100	paroxysmal nocturnal hemoglobinuria (PNH)
SRR3724632	Not	Lymphocytes	SRX1881727	GSM2219010: PNH1-CD8Naive; Homo sapiens; RNA-Seq	Homo sapiens	T cells from paroxysmal nocturnal hemoglobinuria patients show an altered TNFR signaling pathway	PRJNA327044	34722293	100	paroxysmal nocturnal hemoglobinuria (PNH)

Table_S1 Metadata for differential analysis

SRR3724633	Not	Lymphocytes	SRX1881728		GSM2219011: PNH2-CD4Memory; Homo sapiens; RNA-Seq	Homo sapiens	T cells from paroxysmal nocturnal hemoglobinuria patients show an altered TNFR signaling pathway	PRJNA327044	32466930	100	paroxysmal nocturnal hemoglobinuria (PNH)
SRR3724634	Not	Lymphocytes	SRX1881729		GSM2219012: PNH2-CD4Naive; Homo sapiens; RNA-Seq	Homo sapiens	T cells from paroxysmal nocturnal hemoglobinuria patients show an altered TNFR signaling pathway	PRJNA327044	29139809	100	paroxysmal nocturnal hemoglobinuria (PNH)
SRR3724635	Not	Lymphocytes	SRX1881730		GSM2219013: PNH2-CD8Memory; Homo sapiens; RNA-Seq	Homo sapiens	T cells from paroxysmal nocturnal hemoglobinuria patients show an altered TNFR signaling pathway	PRJNA327044	28506728	100	paroxysmal nocturnal hemoglobinuria (PNH)
SRR3724636	Not	Lymphocytes	SRX1881731		GSM2219014: PNH2-CD8Naive; Homo sapiens; RNA-Seq	Homo sapiens	T cells from paroxysmal nocturnal hemoglobinuria patients show an altered TNFR signaling pathway	PRJNA327044	30666327	100	paroxysmal nocturnal hemoglobinuria (PNH)
SRR3724637	Not	Lymphocytes	SRX1881732		GSM2219015: PNH3-CD4Memory; Homo sapiens; RNA-Seq	Homo sapiens	T cells from paroxysmal nocturnal hemoglobinuria patients show an altered TNFR signaling pathway	PRJNA327044	49485440	100	paroxysmal nocturnal hemoglobinuria (PNH)
SRR3724638	Not	Lymphocytes	SRX1881733		GSM2219016: PNH3-CD4Naive; Homo sapiens; RNA-Seq	Homo sapiens	T cells from paroxysmal nocturnal hemoglobinuria patients show an altered TNFR signaling pathway	PRJNA327044	43428307	100	paroxysmal nocturnal hemoglobinuria (PNH)
SRR3724639	Not	Lymphocytes	SRX1881734		GSM2219017: PNH3-CD8Memory; Homo sapiens; RNA-Seq	Homo sapiens	T cells from paroxysmal nocturnal hemoglobinuria patients show an altered TNFR signaling pathway	PRJNA327044	41998683	100	paroxysmal nocturnal hemoglobinuria (PNH)
SRR3724640	Not	Lymphocytes	SRX1881735		GSM2219018: PNH3-CD8Naive; Homo sapiens; RNA-Seq	Homo sapiens	T cells from paroxysmal nocturnal hemoglobinuria patients show an altered TNFR signaling pathway	PRJNA327044	52622177	100	paroxysmal nocturnal hemoglobinuria (PNH)

Table_S2 Metadata for ENCODE kmer research

Original.files	label	group	Accession	Assay.Nickname	Biosample.summary	Lab	Project	Species	Life.stage	Age	Age.Units	Treatment
ENcff371KWE	dendritic cell	6: Others	ENCSR892ZUK	total RNA-seq	dendritic cell treated with 100 ng/mL Lipopolysaccharide for 2 hours	Manuel Garber, UMass	GGR	Homo sapiens	unknown	unknown		Lipopolysaccharide
ENcff534HFO	dendritic cell	6: Others	ENCSR519NFO	total RNA-seq	dendritic cell treated with 100 ng/mL Lipopolysaccharide for 1 hour	Manuel Garber, UMass	GGR	Homo sapiens	unknown	unknown		Lipopolysaccharide
ENcff126HPJ	endometrial microvascular endothelial cells	8: Endothelial cells	ENCSR919MZM	total RNA-seq	endometrial microvascular endothelial cells female adult (34 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	34	year	
ENcff502JCS	endometrial microvascular endothelial cells	8: Endothelial cells	ENCSR919MZM	total RNA-seq	endometrial microvascular endothelial cells female adult (34 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	34	year	
ENcff001QZS	smooth muscle cell of bladder	7: SMC	ENCSR000AAC	total RNA-seq	smooth muscle cell of bladder female adult (53 years) and male adult (62 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	53.62	year	
ENcff001RAO	smooth muscle cell of bladder	7: SMC	ENCSR000AAC	total RNA-seq	smooth muscle cell of bladder female adult (53 years) and male adult (62 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	53.62	year	
ENcff001QZC	renal cortical epithelial cell	9: Epithelial cells	ENCSR000AAQ	total RNA-seq	renal cortical epithelial cell female adult (69 years) and male adult (64 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	69.84	year	
ENcff001RAU	renal cortical epithelial cell	9: Epithelial cells	ENCSR000AAQ	total RNA-seq	renal cortical epithelial cell female adult (69 years) and male adult (64 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	69.84	year	
ENcff001QZK	bladder microvascular endothelial cell	8: Endothelial cells	ENCSR000AAB	total RNA-seq	bladder microvascular endothelial cell male adult (46 years) and male adult (60 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	46.6	year	
ENcff001RAI	bladder microvascular endothelial cell	8: Endothelial cells	ENCSR000AAB	total RNA-seq	bladder microvascular endothelial cell male adult (46 years) and male adult (60 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	46.6	year	
ENcff001QZG	dermis blood vessel endothelial cell	8: Endothelial cells	ENCSR000AAI	total RNA-seq	dermis blood vessel endothelial cell female child (16 years) and male child (13 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	child	13.16	year	
ENcff001RAQ	dermis blood vessel endothelial cell	8: Endothelial cells	ENCSR000AAI	total RNA-seq	dermis blood vessel endothelial cell female child (16 years) and male child (13 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	child	13.16	year	
ENcff000ILE	skeletal muscle satellite cell	4: Progenitor or Stem cell	ENCSR000CUI	total RNA-seq	skeletal muscle satellite cell female adult (64 years) and male adult (21 year)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	21.64	year	
ENcff000ILQ	skeletal muscle satellite cell	4: Progenitor or Stem cell	ENCSR000CUI	total RNA-seq	skeletal muscle satellite cell female adult (64 years) and male adult (21 year)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	21.64	year	
ENcff000GFS	mononuclear cell	6: Others	ENCSR000CUT	total RNA-seq	mononuclear cell female adult (52 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	52	year	
ENcff810KXA	dendritic cell	6: Others	ENCSR682CFV	total RNA-seq	dendritic cell treated with 100 ng/mL Lipopolysaccharide for 2 hours	Manuel Garber, UMass	GGR	Homo sapiens	unknown	unknown		Lipopolysaccharide
ENcff928LQX	dendritic cell	6: Others	ENCSR652RSO	total RNA-seq	dendritic cell treated with 100 ng/mL Lipopolysaccharide for 6 hours	Manuel Garber, UMass	GGR	Homo sapiens	unknown	unknown		Lipopolysaccharide
ENcff816SSU	dendritic cell	6: Others	ENCSR536GUD	total RNA-seq	dendritic cell treated with 100 ng/mL Lipopolysaccharide for 4 hours	Manuel Garber, UMass	GGR	Homo sapiens	unknown	unknown		Lipopolysaccharide
ENcff720QTK	dendritic cell	6: Others	ENCSR870DRE	total RNA-seq	dendritic cell treated with 0 ng/mL Lipopolysaccharide for 0 hours	Manuel Garber, UMass	GGR	Homo sapiens	unknown	unknown		Lipopolysaccharide
ENcff816QOY	H1-hESC	4: Progenitor or Stem cell	ENCSR537BCG	total RNA-seq	H1-hESC	Joe Ecker, Salk	Roadmap	Homo sapiens	embryonic	unknown		
ENcff476PHC	H1-hESC	4: Progenitor or Stem cell	ENCSR537BCG	total RNA-seq	H1-hESC	Joe Ecker, Salk	Roadmap	Homo sapiens	embryonic	unknown		
ENcff821YWS	H1-hESC	4: Progenitor or Stem cell	ENCSR537BCG	total RNA-seq	H1-hESC	Joe Ecker, Salk	Roadmap	Homo sapiens	embryonic	unknown		
ENcff712SHP	H1-hESC	4: Progenitor or Stem cell	ENCSR537BCG	total RNA-seq	H1-hESC	Joe Ecker, Salk	Roadmap	Homo sapiens	embryonic	unknown		
ENcff793PLC	H1-hESC	4: Progenitor or Stem cell	ENCSR537BCG	total RNA-seq	H1-hESC	Joe Ecker, Salk	Roadmap	Homo sapiens	embryonic	unknown		
ENcff000FKR	thoracic aorta endothelial cell	8: Endothelial cells	ENCSR000CUK	total RNA-seq	thoracic aorta endothelial cell female adult (22 years) and male adult (55 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	55.22	year	
ENcff000FLI	thoracic aorta endothelial cell	8: Endothelial cells	ENCSR000CUK	total RNA-seq	thoracic aorta endothelial cell female adult (22 years) and male adult (55 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	55.22	year	
ENcff000GBQ	hair follicle dermal papilla cell	4: Progenitor or Stem cell	ENCSR000CUB	total RNA-seq	hair follicle dermal papilla cell female adult (47 years) and female adult (70 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	70.47	year	
ENcff000GBT	hair follicle dermal papilla cell	4: Progenitor or Stem cell	ENCSR000CUB	total RNA-seq	hair follicle dermal papilla cell female adult (47 years) and female adult (70 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	70.47	year	
ENcff656QHH	myometrial cell	5: Muscle cells	ENCSR371VGV	total RNA-seq	myometrial cell female adult (34 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	34	year	
ENcff212TDB	myometrial cell	5: Muscle cells	ENCSR371VGV	total RNA-seq	myometrial cell female adult (34 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	34	year	
ENcff995TMT	dendritic cell	6: Others	ENCSR332RS	total RNA-seq	dendritic cell treated with 100 ng/mL Lipopolysaccharide for 4 hours	Manuel Garber, UMass	GGR	Homo sapiens	unknown	unknown		Lipopolysaccharide
ENcff605VBT	glomerular endothelial cell	8: Endothelial cells	ENCSR878EUT	total RNA-seq	glomerular endothelial cell female embryo (22 weeks) and male embryo (22 weeks)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	embryonic	22	week	
ENcff509JLH	glomerular endothelial cell	8: Endothelial cells	ENCSR878EUT	total RNA-seq	glomerular endothelial cell female embryo (22 weeks) and male embryo (22 weeks)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	embryonic	22	week	
ENcff000FLO	articular chondrocyte	character(0)	ENCSR000CUE	total RNA-seq	articular chondrocyte of knee joint female adult (56 years) and male adult (64 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	64.56	year	
ENcff000FMM	articular chondrocyte	character(0)	ENCSR000CUE	total RNA-seq	articular chondrocyte of knee joint female adult (56 years) and male adult (64 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	64.56	year	
ENcff875NIU	dendritic cell	6: Others	ENCSR178SNP	total RNA-seq	dendritic cell treated with 100 ng/mL Lipopolysaccharide for 1 hour	Manuel Garber, UMass	GGR	Homo sapiens	unknown	unknown		Lipopolysaccharide
ENcff000GXC	fibroblast of villous mesenchyme	1: Fibroblast	ENCSR000CUL	total RNA-seq	fibroblast of villous mesenchyme female newborn and male newborn	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	newborn	unknown		
ENcff000GXV	fibroblast of villous mesenchyme	1: Fibroblast	ENCSR000CUL	total RNA-seq	fibroblast of villous mesenchyme female newborn and male newborn	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	newborn	unknown		
ENcff236EYN	hair follicular keratinocyte	6: Others	ENCSR680USE	total RNA-seq	hair follicular keratinocyte male adult (55 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	55	year	
ENcff917MHA	hair follicular keratinocyte	6: Others	ENCSR680USE	total RNA-seq	hair follicular keratinocyte male adult (55 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	55	year	
ENcff142YQX	mesangial cell	6: Others	ENCSR198TKA	total RNA-seq	mesangial cell NONE and female embryo (21 week)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	unknown,embryo	unknown,21	week	
ENcff644UKX	mesangial cell	6: Others	ENCSR198TKA	total RNA-seq	mesangial cell NONE and female embryo (21 week)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	unknown,embryo	unknown,21	week	
ENcff927MXX	myotube	5: Muscle cells	ENCSR828TEI	total RNA-seq	myotube originated from skeletal muscle myoblast	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	unknown	unknown		
ENcff999QZD	myotube	5: Muscle cells	ENCSR828TEI	total RNA-seq	myotube originated from skeletal muscle myoblast	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	unknown	unknown		
ENcff911ELB	foreskin keratinocyte	6: Others	ENCSR034RPU	total RNA-seq	foreskin keratinocyte male newborn (2-4 days)	Michael Snyder, Stanford	GGR	Homo sapiens	newborn	2-4	day	
ENcff578CTE	foreskin keratinocyte	6: Others	ENCSR034RPU	total RNA-seq	foreskin keratinocyte male newborn (2-4 days)	Michael Snyder, Stanford	GGR	Homo sapiens	newborn	2-4	day	
ENcff000GNB	vein endothelial cell	8: Endothelial cells	ENCSR000CUG	total RNA-seq	vein endothelial cell male adult (48 years) and male adult (52 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	52.48	year	

Table_S2 Metadata for ENCODE kmer research

ENcff000GNC	vein endothelial cell	8: Endothelial cells	ENCSR000CUG	total RNA-seq	vein endothelial cell male adult (48 years) and male adult (52 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	52.48year	
ENcff001QZM	smooth muscle cell of trachea	7: SMC	ENCSR000AAS	total RNA-seq	smooth muscle cell of trachea male adult (28 years) and male adult (56 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	56.28year	
ENcff001RDQ	smooth muscle cell of trachea	7: SMC	ENCSR000AAS	total RNA-seq	smooth muscle cell of trachea male adult (28 years) and male adult (56 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	56.28year	
ENcff001RCA	smooth muscle cell of the pulmonary artery	7: SMC	ENCSR000AAN	total RNA-seq	smooth muscle cell of the pulmonary artery male adult (26 years) and male adult (28 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	28.26year	
ENcff001RDO	smooth muscle cell of the pulmonary artery	7: SMC	ENCSR000AAN	total RNA-seq	smooth muscle cell of the pulmonary artery male adult (26 years) and male adult (28 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	28.26year	
ENcff000GXK	placental pericyte	6: Others	ENCSR000CTX	total RNA-seq	placental pericyte female newborn and male newborn	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	newborn	unknown	
ENcff000GKZ	placental pericyte	6: Others	ENCSR000CTX	total RNA-seq	placental pericyte female newborn and male newborn	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	newborn	unknown	
ENcff000HWI	fibroblast of dermis	1: Fibroblast	ENCSR000CUH	total RNA-seq	fibroblast of dermis female adult (44 years) and female adult (55 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	55.44year	
ENcff000HXB	fibroblast of dermis	1: Fibroblast	ENCSR000CUH	total RNA-seq	fibroblast of dermis female adult (44 years) and female adult (55 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	55.44year	
ENcff644CJJ	bipolar neuron	6: Others	ENCSR968WKR	total RNA-seq	bipolar neuron originated from GM23338 treated with 0.5 µg/mL doxycycline hyclate for 4 days	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	53year	doxycycline hyclate
ENcff040TFC	bipolar neuron	6: Others	ENCSR968WKR	total RNA-seq	bipolar neuron originated from GM23338 treated with 0.5 µg/mL doxycycline hyclate for 4 days	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	53year	doxycycline hyclate
ENcff000GFU	MSC of adipose	character(0)	ENCSR000CTZ	total RNA-seq	mesenchymal stem cell of adipose	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	42.37year	
ENcff000GGR	MSC of adipose	character(0)	ENCSR000CTZ	total RNA-seq	mesenchymal stem cell of adipose	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	42.37year	
ENcff001RBG	lung microvascular endothelial cell	8: Endothelial cells	ENCSR000AAP	total RNA-seq	lung microvascular endothelial cell female adult (55 years) and male adult (63 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	63.55year	
ENcff001RBW	lung microvascular endothelial cell	8: Endothelial cells	ENCSR000AAP	total RNA-seq	lung microvascular endothelial cell female adult (55 years) and male adult (63 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	63.55year	
ENcff840QXH	bronchus fibroblast of lung	1: Fibroblast	ENCSR620NSN	total RNA-seq	bronchus fibroblast of lung	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	unknown	unknown	
ENcff480YLN	bronchus fibroblast of lung	1: Fibroblast	ENCSR620NSN	total RNA-seq	bronchus fibroblast of lung	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	unkown	unknown	
ENcff001RAK	tracheal epithelial cell	9: Epithelial cells	ENCSR000AAR	total RNA-seq	tracheal epithelial cell male adult (21 year) and male adult (68 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	21.68year	
ENcff001RBS	tracheal epithelial cell	9: Epithelial cells	ENCSR000AAR	total RNA-seq	tracheal epithelial cell male adult (21 year) and male adult (68 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	21.68year	
ENcff001QZE	epithelial cell of viscerocranial mucosa	character(0)	ENCSR000AAL	total RNA-seq	nasal cavity respiratory epithelium epithelial cell of viscerocranial mucosa female adult (70 years) and male adult (46 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	46.7year	
ENcff001QZI	epithelial cell of viscerocranial mucosa	character(0)	ENCSR000AAL	total RNA-seq	nasal cavity respiratory epithelium epithelial cell of viscerocranial mucosa female adult (70 years) and male adult (46 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	46.7year	
ENcff001RBA	uterine smooth muscle cell	7: SMC	ENCSR000AAV	total RNA-seq	uterine smooth muscle cell female adult (48 years) and female adult (50 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	48.5year	
ENcff001RCC	uterine smooth muscle cell	7: SMC	ENCSR000AAV	total RNA-seq	uterine smooth muscle cell female adult (48 years) and female adult (50 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	48.5year	
ENcff001QZQ	aortic SMC	character(0)	ENCSR000AAA	total RNA-seq	aortic smooth muscle cell male adult (21 year) and male adult (54 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	54.21year	
ENcff001RAM	aortic SMC	character(0)	ENCSR000AAA	total RNA-seq	aortic smooth muscle cell male adult (21 year) and male adult (54 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	54.21year	
ENcff000IGA	melanocyte of skin	6: Others	ENCSR000CUQ	total RNA-seq	melanocyte of skin male child (1 year) and male child (3 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	child	3.1year	
ENcff000IGU	melanocyte of skin	6: Others	ENCSR000CUQ	total RNA-seq	melanocyte of skin male child (1 year) and male child (3 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	child	3.1year	
ENcff446LHV	dendritic cell	6: Others	ENCSR022BYE	total RNA-seq	dendritic cell treated with 0 ng/mL Lipopolysaccharide for 0 hours	Manuel Garber, UMass	GGR	Homo sapiens	unknown	unknown	Lipopolysaccharide
ENcff817RIH	dendritic cell	6: Others	ENCSR227MWL	total RNA-seq	dendritic cell treated with 100 ng/mL Lipopolysaccharide for 2 hours	Manuel Garber, UMass	GGR	Homo sapiens	unknown	unknown	Lipopolysaccharide
ENcff662HMZ	dendritic cell	6: Others	ENCSR975YGW	total RNA-seq	dendritic cell treated with 100 ng/mL Lipopolysaccharide for 2 hours	Manuel Garber, UMass	GGR	Homo sapiens	unknown	unknown	Lipopolysaccharide
ENcff001RAG	bronchial smooth muscle cell	7: SMC	ENCSR000AAE	total RNA-seq	bronchial smooth muscle cell male adult (52 years) and male adult (59 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	52.59year	
ENcff001RAS	bronchial smooth muscle cell	7: SMC	ENCSR000AAE	total RNA-seq	bronchial smooth muscle cell male adult (52 years) and male adult (59 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	52.59year	
ENcff001RAA	endothelial cell of coronary artery	8: Endothelial cells	ENCSR000AAF	total RNA-seq	endothelial cell of coronary artery female adult (41 year) and male adult (77 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	41.77year	
ENcff001RBC	endothelial cell of coronary artery	8: Endothelial cells	ENCSR000AAF	total RNA-seq	endothelial cell of coronary artery female adult (41 year) and male adult (77 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	41.77year	
ENcff001QZY	smooth muscle cell of the coronary artery	7: SMC	ENCSR000AAG	total RNA-seq	smooth muscle cell of the coronary artery female adult (53 years) and male adult (55 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	53.55year	
ENcff001RCG	smooth muscle cell of the coronary artery	7: SMC	ENCSR000AAG	total RNA-seq	smooth muscle cell of the coronary artery female adult (53 years) and male adult (55 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	53.55year	
ENcff001QZU	pulmonary artery endothelial cell	8: Endothelial cells	ENCSR000AAM	total RNA-seq	pulmonary artery endothelial cell male adult (23 years) and male adult (52 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	23.52year	
ENcff001RCE	pulmonary artery endothelial cell	8: Endothelial cells	ENCSR000AAM	total RNA-seq	pulmonary artery endothelial cell male adult (23 years) and male adult (52 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	23.52year	
ENcff071TOB	dendritic cell	6: Others	ENCSR084JIA	total RNA-seq	dendrite cell treated with 100 ng/mL Lipopolysaccharide for 1 hour	Manuel Garber, UMass	GGR	Homo sapiens	unknown	unknown	Lipopolysaccharide
ENcff001RAE	regular cardiac myocyte	5: Muscle cells	ENCSR000AAH	total RNA-seq	regular cardiac myocyte female adult (51 year) and male adult (48 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	48.51year	
ENcff001RBM	regular cardiac myocyte	5: Muscle cells	ENCSR000AAH	total RNA-seq	regular cardiac myocyte female adult (51 year) and male adult (48 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	48.51year	
ENcff001RAY	dermis lymphatic vessel endothelial cell	8: Endothelial cells	ENCSR000AAJ	total RNA-seq	dermis lymphatic vessel endothelial cell female adult (45 years) and male child (6 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	child,adult	6.45year	

Table_S2 Metadata for ENCODE kmer research

ENCF001RBQ	dermis lymphatic vessel endothelial cell	8: Endothelial cells	ENCSR000AAJ	total RNA-seq	dermis lymphatic vessel endothelial cell female adult (45 years) and male child (6 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	child,adult	6.45	year	
ENCF001RAW	dermis microvascular lymphatic vessel endothelial cell	8: Endothelial cells	ENCSR000AAK	total RNA-seq	dermis microvascular lymphatic vessel endothelial cell female adult (38 years) and female adult (64 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	38.64	year	
ENCF001RBE	dermis microvascular lymphatic vessel endothelial cell	8: Endothelial cells	ENCSR000AAK	total RNA-seq	dermis microvascular lymphatic vessel endothelial cell female adult (38 years) and female adult (64 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	38.64	year	
ENCF365NAH	foreskin keratinocyte	6: Others	ENCSR527SSD	total RNA-seq	foreskin keratinocyte male newborn (2-4 days) treated with 1.2 mM calcium for 5.5 days	Michael Snyder, Stanford	GGR	Homo sapiens	newborn	2-4	day	calcium
ENCF668JVG	foreskin keratinocyte	6: Others	ENCSR527SSD	total RNA-seq	foreskin keratinocyte male newborn (2-4 days) treated with 1.2 mM calcium for 5.5 days	Michael Snyder, Stanford	GGR	Homo sapiens	newborn	2-4	day	calcium
ENCF649ASL	dendritic cell	6: Others	ENCSR571IUZ	total RNA-seq	dendritic cell treated with 100 ng/mL Lipopolysaccharide for 4 hours	Manuel Garber, UMass	GGR	Homo sapiens	unknown	unknown		Lipopolysaccharide
ENCF123JND	dendritic cell	6: Others	ENCSR760INS	total RNA-seq	dendritic cell treated with 100 ng/mL Lipopolysaccharide for 6 hours	Manuel Garber, UMass	GGR	Homo sapiens	unknown	unknown		Lipopolysaccharide
ENCF945OFS	dendritic cell	6: Others	ENCSR670MXT	total RNA-seq	dendritic cell treated with 100 ng/mL Lipopolysaccharide for 2 hours	Manuel Garber, UMass	GGR	Homo sapiens	unknown	unknown		Lipopolysaccharide
ENCF124RBL	dendritic cell	6: Others	ENCSR476TKU	total RNA-seq	dendritic cell treated with 100 ng/mL Lipopolysaccharide for 4 hours	Manuel Garber, UMass	GGR	Homo sapiens	unknown	unknown		Lipopolysaccharide
ENCF365UNT	dendritic cell	6: Others	ENCSR475TVN	total RNA-seq	dendritic cell treated with 100 ng/mL Lipopolysaccharide for 6 hours	Manuel Garber, UMass	GGR	Homo sapiens	unknown	unknown		Lipopolysaccharide
ENCF658PGS	dendritic cell	6: Others	ENCSR707JVU	total RNA-seq	dendritic cell treated with 100 ng/mL Lipopolysaccharide for 4 hours	Manuel Garber, UMass	GGR	Homo sapiens	unknown	unknown		Lipopolysaccharide
ENCF609HPU	dendritic cell	6: Others	ENCSR916YUR	total RNA-seq	dendritic cell treated with 100 ng/mL Lipopolysaccharide for 6 hours	Manuel Garber, UMass	GGR	Homo sapiens	unknown	unknown		Lipopolysaccharide
ENCF000GLZ	placental epithelial cell	9: Epithelial cells	ENCSR000CUP	total RNA-seq	placental epithelial cell female newborn and male newborn	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	newborn	unknown		
ENCF000GMA	placental epithelial cell	9: Epithelial cells	ENCSR000CUP	total RNA-seq	placental epithelial cell female newborn and male newborn	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	newborn	unknown		
ENCF000GYE	subcutaneous preadipocyte	4: Progenitor or Stem cell	ENCSR000CUM	total RNA-seq	subcutaneous preadipocyte female adult (62 years) and male adult (65 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	65.62	year	
ENCF000GZD	subcutaneous preadipocyte	4: Progenitor or Stem cell	ENCSR000CUM	total RNA-seq	subcutaneous preadipocyte female adult (62 years) and male adult (65 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	65.62	year	
ENCF947BAG	dendritic cell	6: Others	ENCSR146PLL	total RNA-seq	dendritic cell treated with 100 ng/mL Lipopolysaccharide for 4 hours	Manuel Garber, UMass	GGR	Homo sapiens	unknown	unknown		Lipopolysaccharide
ENCF659EDN	mammary microvascular endothelial cell	8: Endothelial cells	ENCSR815UVL	total RNA-seq	mammary microvascular endothelial cell female adult (26 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	26	year	
ENCF143UMB	mammary microvascular endothelial cell	8: Endothelial cells	ENCSR815UVL	total RNA-seq	mammary microvascular endothelial cell female adult (26 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	26	year	
ENCF001QZW	epithelial cell of umbilical artery	9: Epithelial cells	ENCSR000AAT	total RNA-seq	epithelial cell of umbilical artery female newborn and male newborn	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	newborn	unknown		
ENCF001RBC	epithelial cell of umbilical artery	9: Epithelial cells	ENCSR000AAT	total RNA-seq	epithelial cell of umbilical artery female newborn and male newborn	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	newborn	unknown		
ENCF001RBU	smooth muscle cell of the umbilical artery	7: SMC	ENCSR000AUU	total RNA-seq	smooth muscle cell of the umbilical artery female newborn and male newborn	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	newborn	unknown		
ENCF001RBY	smooth muscle cell of the umbilical artery	7: SMC	ENCSR000AUU	total RNA-seq	smooth muscle cell of the umbilical artery female newborn and male newborn	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	newborn	unknown		
ENCF002DMJ	H7-hESC	4: Progenitor or Stem cell	ENCSR490SQH	total RNA-seq	H7-hESC	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	embryonic	unknown		
ENCF002DMP	H7-hESC	4: Progenitor or Stem cell	ENCSR490SQH	total RNA-seq	H7-hESC	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	embryonic	unknown		
ENCF000IHC	melanocyte of skin	6: Others	ENCSR000CUR	total RNA-seq	melanocyte of skin female adult (52 years) and male adult (55 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	55.52	year	
ENCF000IHY	melanocyte of skin	6: Others	ENCSR000CUR	total RNA-seq	melanocyte of skin female adult (52 years) and male adult (55 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	55.52	year	
ENCF575SLU	cardiac ventricle fibroblast	1: Fibroblast	ENCSR369RVN	total RNA-seq	cardiac ventricle fibroblast NONE and male adult (18 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult,unknown	18,unknown	year	
ENCF157UXG	cardiac ventricle fibroblast	1: Fibroblast	ENCSR369RVN	total RNA-seq	cardiac ventricle fibroblast NONE and male adult (18 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult,unknown	18,unknown	year	
ENCF689YVG	pericardium fibroblast	1: Fibroblast	ENCSR362HMX	total RNA-seq	pericardium fibroblast NONE and female embryo (20 weeks)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	embryonic,unkn	20,unknown	week	
ENCF005EEW	pericardium fibroblast	1: Fibroblast	ENCSR362HMX	total RNA-seq	pericardium fibroblast NONE and female embryo (20 weeks)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	embryonic,unkn	20,unknown	week	
ENCF746KSI	epithelial cell of alveolus of lung	9: Epithelial cells	ENCSR897KTO	total RNA-seq	epithelial cell of alveolus of lung NONE and female embryo (21 week)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	unknown,embry	unknown,21	week	
ENCF408FWI	epithelial cell of alveolus of lung	9: Epithelial cells	ENCSR897KTO	total RNA-seq	epithelial cell of alveolus of lung NONE and female embryo (21 week)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	unknown,embry	unknown,21	week	
ENCF109IUU	kidney epithelial cell	9: Epithelial cells	ENCSR373BDG	total RNA-seq	kidney epithelial cell male embryo (22 weeks) and male newborn	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	embryonic,newb	22,unknown	week	
ENCF999ZER	kidney epithelial cell	9: Epithelial cells	ENCSR373BDG	total RNA-seq	kidney epithelial cell male embryo (22 weeks) and male newborn	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	embryonic,newb	22,unknown	week	
ENCF353STY	cardiac atrium fibroblast	1: Fibroblast	ENCSR110BDY	total RNA-seq	cardiac atrium fibroblast male child (2 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	child	2	year	
ENCF-F73GPU	cardiac atrium fibroblast	1: Fibroblast	ENCSR110BDY	total RNA-seq	cardiac atrium fibroblast male child (2 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	child	2	year	
ENCF000GJC	osteoblast	3: Osteo/Chondro lineage	ENCSR000CUF	total RNA-seq	osteoblast female adult (56 years) and male adult (62 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	56.62	year	
ENCF000GKC	osteoblast	3: Osteo/Chondro lineage	ENCSR000CUF	total RNA-seq	osteoblast female adult (56 years) and male adult (62 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	56.62	year	
ENCF000GET	mammary epithelial cell	9: Epithelial cells	ENCSR000CUN	total RNA-seq	mammary epithelial cell female adult (23 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	23	year	
ENCF353SPR	cardiac muscle cell	5: Muscle cells	ENCSR379YAE	total RNA-seq	cardiac muscle cell originated from RUES2	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	embryonic	unknown		
ENCF030WXU	cardiac muscle cell	5: Muscle cells	ENCSR379YAE	total RNA-seq	cardiac muscle cell originated from RUES2	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	embryonic	unknown		
ENCF472QXK	foreskin keratinocyte	6: Others	ENCSR959LTT	total RNA-seq	foreskin keratinocyte male newborn (2-4 days) treated with 1.2 mM calcium for 2.5 days	Michael Snyder, Stanford	GGR	Homo sapiens	newborn	2-4	day	calcium
ENCF218VQQ	foreskin keratinocyte	6: Others	ENCSR959LTT	total RNA-seq	foreskin keratinocyte male newborn (2-4 days) treated with 1.2 mM calcium for 2.5 days	Michael Snyder, Stanford	GGR	Homo sapiens	newborn	2-4	day	calcium

Table_S2 Metadata for ENCODE kmer research

ENCF000GIC	MSC of Wharton's jelly	character(0)	ENCSR000CUO	total RNA-seq	mesenchymal stem cell of Wharton's jelly	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	newborn	unknown		
ENCF000GIV	MSC of Wharton's jelly	character(0)	ENCSR000CUO	total RNA-seq	mesenchymal stem cell of Wharton's jelly	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	newborn	unknown		
ENCF548JWS	smooth muscle cell	7: SMC	ENCSR052FJA	total RNA-seq	smooth muscle cell originated from H9	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	embryonic	5day		
ENCF199QMY	smooth muscle cell	7: SMC	ENCSR052FJA	total RNA-seq	smooth muscle cell originated from H9	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	embryonic	5day		
ENCF001RAC	bronchial epithelial cell	9: Epithelial cells	ENCSR000AAD	total RNA-seq	bronchial epithelial cell female adult (40 years) and male adult (68 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	68.4/year		
ENCF001RB0	bronchial epithelial cell	9: Epithelial cells	ENCSR000AAD	total RNA-seq	bronchial epithelial cell female adult (40 years) and male adult (68 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	68.4/year		
ENCF653EZE	hepatocyte	6: Others	ENCSR908ZAS	total RNA-seq	hepatocyte originated from H9	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	embryonic	5day		
ENCF245VTB	hepatocyte	6: Others	ENCSR908ZAS	total RNA-seq	hepatocyte originated from H9	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	embryonic	5day		
ENCF996KMK	neural progenitor cell	4: Progenitor or Stem cell	ENCSR244ISQ	total RNA-seq	neural progenitor cell originated from H9	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	embryonic	5day		
ENCF939FVE	neural progenitor cell	4: Progenitor or Stem cell	ENCSR244ISQ	total RNA-seq	neural progenitor cell originated from H9	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	embryonic	5day		
ENCF002DMH	skeletal muscle myoblast	5: Muscle cells	ENCSR444WHQ	total RNA-seq	skeletal muscle myoblast	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	unknown	unknown		
ENCF002DML	skeletal muscle myoblast	5: Muscle cells	ENCSR444WHQ	total RNA-seq	skeletal muscle myoblast	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	unknown	unknown		
ENCF119TIR	myocyte	5: Muscle cells	ENCSR894WMQ	total RNA-seq	myocyte originated from LHCN-M2	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	41year		
ENCF700EUG	myocyte	5: Muscle cells	ENCSR894WMQ	total RNA-seq	myocyte originated from LHCN-M2	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	41year		
ENCF000GHA	MSC of the bone marrow	character(0)	ENCSR000CUJ	total RNA-seq	mesenchymal stem cell of the bone marrow	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	57.6/year		
ENCF000GHX	MSC of the bone marrow	character(0)	ENCSR000CUJ	total RNA-seq	mesenchymal stem cell of the bone marrow	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	57.6/year		
ENCF644PFE	airway epithelial cell	9: Epithelial cells	ENCSR822SUG	total RNA-seq	airway epithelial cell	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	unknown	unknown		
ENCF846YCY	airway epithelial cell	9: Epithelial cells	ENCSR822SUG	total RNA-seq	airway epithelial cell	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	unknown	unknown		
ENCF541NUM	epithelial cell of proximal tubule	9: Epithelial cells	ENCSR118TVR	total RNA-seq	epithelial cell of proximal tubule	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	unknown	unknown		
ENCF479JTM	epithelial cell of proximal tubule	9: Epithelial cells	ENCSR118TVR	total RNA-seq	epithelial cell of proximal tubule	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	unknown	unknown		
ENCF000FJX	fibroblast of the aortic adventitia	1: Fibroblast	ENCSR000CUJ	total RNA-seq	fibroblast of the aortic adventitia female adult (24 years) and male adult (47 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	47.24/year		
ENCF000FKI	fibroblast of the aortic adventitia	1: Fibroblast	ENCSR000CUJ	total RNA-seq	fibroblast of the aortic adventitia female adult (24 years) and male adult (47 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	47.24/year		
ENCF000EUS	hematopoietic multipotent progenitor cell	4: Progenitor or Stem cell	ENCSR000CUA	total RNA-seq	hematopoietic multipotent progenitor cell	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	unknown	unknown		
ENCF393GQZ	astrocyte	6: Others	ENCSR233JT	total RNA-seq	astrocyte	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	unknown	unknown		
ENCF424DEO	astrocyte	6: Others	ENCSR233JT	total RNA-seq	astrocyte	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	unknown	unknown		
ENCF001QZO	fibroblast of lung	1: Fibroblast	ENCSR000AAO	total RNA-seq	fibroblast of lung female adult (83 years) and male adult (23 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	83.23/year		
ENCF001RBI	fibroblast of lung	1: Fibroblast	ENCSR000AAO	total RNA-seq	fibroblast of lung female adult (83 years) and male adult (23 years)	Thomas Gingeras, CSHL	ENCODE	Homo sapiens	adult	83.23/year		

Table_S3 Meataadata for single cell analysis and FANTOM6 kmer research

Run ID	titles	experiment_type	fraction	run_no	target_gene_symbol	specie	study_title	Study ID	nreads	read_length	library_selection	perturb_id	cell_type_alias
SRR7294023	single-cell RNA-seq of human adipose-derived mesenchymal stem cells	10X genomics sequencing	NA	NA	NA	Homo sapiens	single-cell transcriptomic sequencing of 24,370 adipose-derived mesenchymal stem cells	PRJNA472816	176545104	149.75	PolyA	NA	Ad-MSCs
SRR7253635	single-cell RNA-seq of human adipose-derived mesenchymal stem cells	10X genomics sequencing	NA	NA	NA	Homo sapiens	single-cell transcriptomic sequencing of 24,370 adipose-derived mesenchymal stem cells	PRJNA472816	265719530	149.99	PolyA	NA	Ad-MSCs
SRR7363187	single-cell RNA-seq of human adipose-derived mesenchymal stem cells	10X genomics sequencing	NA	NA	NA	Homo sapiens	single-cell transcriptomic sequencing of 24,370 adipose-derived mesenchymal stem cells	PRJNA472816	352873537	149.98	PolyA	NA	Ad-MSCs
Run ID	titles	experiment_type	fraction	run_no	target_gene_symbol	specie	study_title	Study ID	nreads	read_length	library_selection	perturb_id	cell_type_alias
DRR177156	RDh10602_ATCACG.experiment	Reference	cytoplasmic	0 NA		Homo sapiens	FANTOM6 project	PRJDB7993	33503637	51 inverse rRNA	NA		HDF (neonatal)
DRR177155	RDh10602_ACTTGA.experiment	Reference	cytoplasmic	0 NA		Homo sapiens	FANTOM6 project	PRJDB7993	30096528	51 inverse rRNA	NA		HDF (neonatal)
DRR177154	RDh10601_TTAGGC.experiment	Reference	nuclear	0 NA		Homo sapiens	FANTOM6 project	PRJDB7993	35912547	51 inverse rRNA	NA		HDF (neonatal)
DRR177153	RDh10601_GATCAG.experiment	Reference	nuclear	0 NA		Homo sapiens	FANTOM6 project	PRJDB7993	35267146	51 inverse rRNA	NA		HDF (neonatal)
DRR177152	RDh10600_ATCACG.experiment	Reference	chromatin	0 NA		Homo sapiens	FANTOM6 project	PRJDB7993	33268855	51 inverse rRNA	NA		HDF (neonatal)
DRR177151	RDh10600_ACTTGA.experiment	Reference	chromatin	0 NA		Homo sapiens	FANTOM6 project	PRJDB7993	34620706	51 inverse rRNA	NA		HDF (neonatal)
DRR176801	CNh10765 ACC.experiment	Targeted	NA	14 RP11-54A9.1		Homo sapiens	FANTOM6 project	PRJDB7993	17143511	47 CAGE	ASO_G0257219_AD_05		HDF (neonatal)
DRR176800	CNh10764 TAC.experiment	Targeted	NA	13 TERC		Homo sapiens	FANTOM6 project	PRJDB7993	22288472	47 CAGE	ASO_G0270141_AD_10		HDF (neonatal)
DRR176799	CNh10764 GCT.experiment	Negative control	NA	14 NA		Homo sapiens	FANTOM6 project	PRJDB7993	16779827	47 CAGE	ASO_Lipo		HDF (neonatal)
DRR176798	CNh10764 GCG.experiment	Targeted	NA	14 RP5-886K2.3		Homo sapiens	FANTOM6 project	PRJDB7993	13412787	47 CAGE	ASO_G0236810_AD_02		HDF (neonatal)
DRR176797	CNh10764 CAC.experiment	Targeted	NA	14 RP5-886K2.3		Homo sapiens	FANTOM6 project	PRJDB7993	13768161	47 CAGE	ASO_G0236810_AD_02		HDF (neonatal)
DRR176796	CNh10764 ATG.experiment	Experiment control	NA	13 MALAT1		Homo sapiens	FANTOM6 project	PRJDB7993	18335819	47 CAGE	ASO_MALAT1		HDF (neonatal)
DRR176731	CNh10640 ACG.experiment	Targeted	NA	12 DNM3OS		Homo sapiens	FANTOM6 project	PRJDB7993	14547994	47 CAGE	ASO_G0230630_AD_04		HDF (neonatal)
DRR176730	CNh10640 ACC.experiment	Targeted	NA	12 EMX2OS		Homo sapiens	FANTOM6 project	PRJDB7993	16086917	47 CAGE	ASO_G02239847_AD_04		HDF (neonatal)
DRR176729	CNh10639 TAC.experiment	Experiment control	NA	12 MALAT1		Homo sapiens	FANTOM6 project	PRJDB7993	14896166	47 CAGE	ASO_MALAT1		HDF (neonatal)
DRR176728	CNh10639 GCT.experiment	Targeted	NA	12 TUG1		Homo sapiens	FANTOM6 project	PRJDB7993	14574794	47 CAGE	ASO_G253352_AD_08		HDF (neonatal)
DRR176727	CNh10639 GCG.experiment	Targeted	NA	12 C11orH95		Homo sapiens	FANTOM6 project	PRJDB7993	12726968	47 CAGE	ASO_G0188070_AD_05		HDF (neonatal)
DRR176726	CNh10639 CAC.experiment	Targeted	NA	12 LINC00339		Homo sapiens	FANTOM6 project	PRJDB7993	13664802	47 CAGE	ASO_G0218510_AD_03		HDF (neonatal)
DRR176725	CNh10639 ATG.experiment	Targeted	NA	12 TERC		Homo sapiens	FANTOM6 project	PRJDB7993	12597402	47 CAGE	ASO_G0270141_AD_10		HDF (neonatal)
DRR176724	CNh10639 AGT.experiment	Negative control	NA	12 NA		Homo sapiens	FANTOM6 project	PRJDB7993	11157330	47 CAGE	ASO_NC_A		HDF (neonatal)
DRR176659	CNh10639 TAC.experiment	Negative control	NA	12 NA		Homo sapiens	FANTOM6 project	PRJDB7993	14568347	47 CAGE	ASO_NC_A		HDF (neonatal)
DRR176658	CNh10630 GCT.experiment	Targeted	NA	12 LINC00339		Homo sapiens	FANTOM6 project	PRJDB7993	14061748	47 CAGE	ASO_G0218510_AD_06		HDF (neonatal)
DRR176657	CNh10630 GCG.experiment	Targeted	NA	12 RP11-417E7.1		Homo sapiens	FANTOM6 project	PRJDB7993	12365942	47 CAGE	ASO_G0223485_AD_06		HDF (neonatal)
DRR176656	CNh10630 CAC.experiment	Targeted	NA	12 C11orH95		Homo sapiens	FANTOM6 project	PRJDB7993	13926089	47 CAGE	ASO_G0188070_AD_05		HDF (neonatal)
DRR176655	CNh10630 ATG.experiment	Targeted	NA	12 ERVK1-13		Homo sapiens	FANTOM6 project	PRJDB7993	15897264	47 CAGE	ASO_G0260565_AD_06		HDF (neonatal)
DRR176654	CNh10630 AGT.experiment	Targeted	NA	12 RP11-417E7.1		Homo sapiens	FANTOM6 project	PRJDB7993	11413198	47 CAGE	ASO_G0223485_AD_04		HDF (neonatal)
DRR176653	CNh10630 ACG.experiment	Targeted	NA	12 RP6-99M1.2		Homo sapiens	FANTOM6 project	PRJDB7993	14016329	47 CAGE	ASO_G0270069_AD_07		HDF (neonatal)
DRR176652	CNh10630 ACC.experiment	Targeted	NA	12 RP11-834C11.4		Homo sapiens	FANTOM6 project	PRJDB7993	20414932	47 CAGE	ASO_G0250742_AD_04		HDF (neonatal)
DRR176587	CNh10609 GCT.experiment	Targeted	NA	10 NR2F1-AS1		Homo sapiens	FANTOM6 project	PRJDB7993	535389	47 CAGE	ASO_G0237187_01		HDF (neonatal)
DRR176586	CNh10609 GCG.experiment	Targeted	NA	10 RP6-109B7.3		Homo sapiens	FANTOM6 project	PRJDB7993	10382541	47 CAGE	ASO_G0241990_07		HDF (neonatal)
DRR176585	CNh10605 CAC.experiment	Targeted	NA	10 LINC00674		Homo sapiens	FANTOM6 project	PRJDB7993	13123013	47 CAGE	ASO_G0237854_05		HDF (neonatal)
DRR176584	CNh10605 ATG.experiment	Experiment control	NA	10 MALAT1		Homo sapiens	FANTOM6 project	PRJDB7993	14008394	47 CAGE	ASO_MALAT1		HDF (neonatal)
DRR176583	CNh10605 AGT.experiment	Targeted	NA	10 CKMT2-AS1		Homo sapiens	FANTOM6 project	PRJDB7993	16349504	47 CAGE	ASO_G0247572_09		HDF (neonatal)
DRR176582	CNh10605 ACG.experiment	Targeted	NA	11 LINC00883		Homo sapiens	FANTOM6 project	PRJDB7993	14714634	47 CAGE	ASO_G0243701_01		HDF (neonatal)
DRR176581	CNh10605 ACC.experiment	Targeted	NA	11 TP53TG1		Homo sapiens	FANTOM6 project	PRJDB7993	16141896	47 CAGE	ASO_G0182165_AD_10		HDF (neonatal)
DRR176580	CNh10604 TAC.experiment	Targeted	NA	10 LINC00707		Homo sapiens	FANTOM6 project	PRJDB7993	12703583	47 CAGE	ASO_G0238266_03		HDF (neonatal)
DRR176515	CNh10584 ATG.experiment	Targeted	NA	9 RP11-95P2.1		Homo sapiens	FANTOM6 project	PRJDB7993	11169499	47 CAGE	ASO_G0262468_07		HDF (neonatal)
DRR176514	CNh10584 AGT.experiment	Targeted	NA	9 RP11-95P2.1		Homo sapiens	FANTOM6 project	PRJDB7993	13068899	47 CAGE	ASO_G0262468_04		HDF (neonatal)
DRR176513	CNh10584 ACG.experiment	Targeted	NA	9 LINC00667		Homo sapiens	FANTOM6 project	PRJDB7993	11868738	47 CAGE	ASO_G0263753_01		HDF (neonatal)
DRR176512	CNh10584 ACC.experiment	Targeted	NA	9 RP11-95P2.1		Homo sapiens	FANTOM6 project	PRJDB7993	19111059	47 CAGE	ASO_G0262468_02		HDF (neonatal)
DRR176511	CNh10584 TAC.experiment	Targeted	NA	9 RP11-333E1.1		Homo sapiens	FANTOM6 project	PRJDB7993	11345714	47 CAGE	ASO_G0261879_05		HDF (neonatal)
DRR176510	CNh10584 GCT.experiment	Targeted	NA	9 RP11-95P2.1		Homo sapiens	FANTOM6 project	PRJDB7993	10783975	47 CAGE	ASO_G0262468_02		HDF (neonatal)
DRR176509	CNh10583 GCG.experiment	Targeted	NA	9 RP11-553K8.5		Homo sapiens	FANTOM6 project	PRJDB7993	12579190	47 CAGE	ASO_G0261573_03		HDF (neonatal)
DRR176508	CNh10583 CAC.experiment	Targeted	NA	9 RP11-553K8.5		Homo sapiens	FANTOM6 project	PRJDB7993	13226029	47 CAGE	ASO_G0261573_02		HDF (neonatal)
DRR176443	CNh10542 AGT.experiment	Targeted	NA	5 CATG00000016989.1		Homo sapiens	FANTOM6 project	PRJDB7993	13134737	47 CAGE	ASO_C0020443_06		HDF (neonatal)
DRR176442	CNh10543 ACC.experiment	Targeted	NA	5 CATG00000079799.1		Homo sapiens	FANTOM6 project	PRJDB7993	11721046	47 CAGE	ASO_C0098586_03		HDF (neonatal)
DRR176441	CNh10543 ACG.experiment	Negative control	NA	5 NA		Homo sapiens	FANTOM6 project	PRJDB7993	12028179	47 CAGE	ASO_NC_A		HDF (neonatal)
DRR176440	CNh10542 TAC.experiment	Reference	NA	0 NA		Homo sapiens	FANTOM6 project	PRJDB7993	1175326	47 CAGE	NA		HDF (neonatal)
DRR176439	CNh10542 GCT.experiment	Negative control	NA	5 NA		Homo sapiens	FANTOM6 project	PRJDB7993	15072508	47 CAGE	ASO_NC_A		HDF (neonatal)
DRR176438	CNh10542 CAC.experiment	Negative control	NA	6 NA		Homo sapiens	FANTOM6 project	PRJDB7993	12572386	47 CAGE	ASO_NC_A		HDF (neonatal)
DRR176437	CNh10542 ATG.experiment	Targeted	NA	6 CTC-338M12.4		Homo sapiens	FANTOM6 project	PRJDB7993	10514355	47 CAGE	ASO_G0233937_05		HDF (neonatal)
DRR176436	CNh10542 AGT.experiment	Targeted	NA	6 RP11-220L1.1		Homo sapiens	FANTOM6 project	PRJDB7993	11847706	47 CAGE	ASO_G0233137_04		HDF (neonatal)
DRR176371	CNh10532 ACC.experiment	Targeted	NA	6 AC092295.7		Homo sapiens	FANTOM6 project	PRJDB7993	11750669	47 CAGE	ASO_G0233527_07		HDF (neonatal)
DRR176370	CNh10532 TAC.experiment	Targeted	NA	5 TK1N1-AS1		Homo sapiens	FANTOM6 project	PRJDB7993	10753751	47 CAGE	ASO_G0186615_03		HDF (neonatal)
DRR176369	CNh10532 GCT.experiment	Targeted	NA	6 AC092295.7		Homo sapiens	FANTOM6 project	PRJDB7993	10343841	47 CAGE	ASO_G0233527_02		HDF (neonatal)
DRR176368	CNh10533 GCG.experiment	Negative control	NA	5 NA		Homo sapiens	FANTOM6 project	PRJDB7993	9051548	47 CAGE	ASO_NC_A		HDF (neonatal)
DRR176367	CNh10533 CAC.experiment	Targeted	NA	5 RP11-296O14.3		Homo sapiens	FANTOM6 project	PRJDB7993	10389408	47 CAGE	ASO_G0203739_10		HDF (neonatal)
DRR176366	CNh10533 ATG.experiment	Targeted	NA	6 AC092295.7		Homo sapiens	FANTOM6 project	PRJDB7993	10291057	47 CAGE	ASO_G0233527_07		HDF (neonatal)
DRR176365	CNh10533 AGT.experiment	Negative control	NA	5 NA		Homo sapiens	FANTOM6 project	PRJDB7993	9027563	47 CAGE	ASO_NC_A		HDF (neonatal)
DRR176364	CNh10533 ACG.experiment	Targeted	NA	6 U52111.14		Homo sapiens	FANTOM6 project	PRJDB7993	11119652	47 CAGE	ASO_G0232725_05		HDF (neonatal)
DRR176299	CNh10524 TAC.experiment	Targeted	NA	5 CATG00000020700.1		Homo sapiens	FANTOM6 project	PRJDB7993	19711422	47 CAGE	ASO_C0020984_09		HDF (neonatal)
DRR176298	CNh10524 GCT.experiment	Targeted	NA	4 CD99P1		Homo sapiens	FANTOM6 project	PRJDB7993	16834564	47 CAGE	ASO_G0223737_05		HDF (neonatal)
DRR176297	CNh10524 GCG.experiment	Targeted	NA	4 RP11-395B7.4		Homo sapiens	FANTOM6 project	PRJDB7993	12730668	47 CAGE	ASO_G0227053_03		HDF (neonatal)
DRR176296	CNh10524 CAC.experiment	Negative control	NA	4 NA		Homo sapiens	FANTOM6 project	PRJDB7993	16433061	47 CAGE	ASO_NC_A		HDF (neonatal)
DRR176295	CNh10524 AGT.experiment	Targeted	NA	4 CD99P1		Homo sapiens	FANTOM6 project	PRJDB7993	17651165	47 CAGE	ASO_G0223773_06		HDF (neonatal)

Table_S4 metadata for kmer research in MSC in different conditions

Run ID	Run Title	Experiment ID	EXTERNAL_ID	specie	study_title	project ID	nreads	read_length	cell type	passages	treatment
SRR5878104	GSM2720343: ATF6+/+ hMSC vehicle-2 RNA-seq; Homo sapiens; RNA-Seq	SRX3045407	GSM2720343	Homo sapiens	ATF6 safeguards organelle homeostasis and cellular aging in human mesenchymal stem cells	PRJNA396193	2991402	101	MSC	p5	vehicle
SRR5878103	GSM2720342: ATF6+/+ hMSC vehicle-1 RNA-seq; Homo sapiens; RNA-Seq	SRX3045406	GSM2720342	Homo sapiens	ATF6 safeguards organelle homeostasis and cellular aging in human mesenchymal stem cells	PRJNA396193	2957936	101	MSC	p5	vehicle
SRR5878102	GSM2720341: ATF6+/+ hMSC tunicamycin-2 RNA-seq; Homo sapiens; RNA-Seq	SRX3045405	GSM2720341	Homo sapiens	ATF6 safeguards organelle homeostasis and cellular aging in human mesenchymal stem cells	PRJNA396193	11944303	101	MSC	p5	tunicamycin
SRR5878101	GSM2720340: ATF6+/+ hMSC tunicamycin-1 RNA-seq; Homo sapiens; RNA-Seq	SRX3045404	GSM2720340	Homo sapiens	ATF6 safeguards organelle homeostasis and cellular aging in human mesenchymal stem cells	PRJNA396193	9700764	101	MSC	p5	tunicamycin
SRR5878100	GSM2720339: ATF6-/ hMSC tunicamycin-2 RNA-seq; Homo sapiens; RNA-Seq	SRX3045403	GSM2720339	Homo sapiens	ATF6 safeguards organelle homeostasis and cellular aging in human mesenchymal stem cells	PRJNA396193	11592138	101	MSC	p5	tunicamycin
SRR5878099	GSM2720338: ATF6-/ hMSC tunicamycin-1 RNA-seq; Homo sapiens; RNA-Seq	SRX3045402	GSM2720338	Homo sapiens	ATF6 safeguards organelle homeostasis and cellular aging in human mesenchymal stem cells	PRJNA396193	10540607	101	MSC	p5	tunicamycin
SRR5878098	GSM2720337: ATF6-/ hMSC vehicle-2 RNA-seq; Homo sapiens; RNA-Seq	SRX3045401	GSM2720337	Homo sapiens	ATF6 safeguards organelle homeostasis and cellular aging in human mesenchymal stem cells	PRJNA396193	9985873	101	MSC	p5	vehicle
SRR5878097	GSM2720336: ATF6-/ hMSC vehicle-1 RNA-seq; Homo sapiens; RNA-Seq	SRX3045400	GSM2720336	Homo sapiens	ATF6 safeguards organelle homeostasis and cellular aging in human mesenchymal stem cells	PRJNA396193	9924834	101	MSC	p5	vehicle
SRR5878096	GSM2720335: ATF6+/+ hWAPC vehicle-2 RNA-seq; Homo sapiens; RNA-Seq	SRX3045399	GSM2720335	Homo sapiens	ATF6 safeguards organelle homeostasis and cellular aging in human mesenchymal stem cells	PRJNA396193	32613469	101	WAPC	p5	vehicle
SRR5878095	GSM2720334: ATF6+/+ hWAPC vehicle-1 RNA-seq; Homo sapiens; RNA-Seq	SRX3045398	GSM2720334	Homo sapiens	ATF6 safeguards organelle homeostasis and cellular aging in human mesenchymal stem cells	PRJNA396193	11973184	101	WAPC	p5	vehicle
SRR5878094	GSM2720333: ATF6+/+ hWAPC tunicamycin-2 RNA-seq; Homo sapiens; RNA-Seq	SRX3045397	GSM2720333	Homo sapiens	ATF6 safeguards organelle homeostasis and cellular aging in human mesenchymal stem cells	PRJNA396193	10129610	101	WAPC	p5	tunicamycin
SRR5878093	GSM2720332: ATF6+/+ hWAPC tunicamycin-1 RNA-seq; Homo sapiens; RNA-Seq	SRX3045396	GSM2720332	Homo sapiens	ATF6 safeguards organelle homeostasis and cellular aging in human mesenchymal stem cells	PRJNA396193	12628610	101	WAPC	p5	tunicamycin
SRR5878092	GSM2720331: ATF6-/ hWAPC tunicamycin-2 RNA-seq; Homo sapiens; RNA-Seq	SRX3045395	GSM2720331	Homo sapiens	ATF6 safeguards organelle homeostasis and cellular aging in human mesenchymal stem cells	PRJNA396193	10730944	101	WAPC	p5	tunicamycin
SRR5878091	GSM2720330: ATF6-/ hWAPC tunicamycin-1 RNA-seq; Homo sapiens; RNA-Seq	SRX3045394	GSM2720330	Homo sapiens	ATF6 safeguards organelle homeostasis and cellular aging in human mesenchymal stem cells	PRJNA396193	9530313	101	WAPC	p5	tunicamycin
SRR5878090	GSM2720329: ATF6-/ hWAPC vehicle-2 RNA-seq; Homo sapiens; RNA-Seq	SRX3045393	GSM2720329	Homo sapiens	ATF6 safeguards organelle homeostasis and cellular aging in human mesenchymal stem cells	PRJNA396193	10217050	101	WAPC	p5	vehicle
SRR5878089	GSM2720328: ATF6-/ hWAPC vehicle-1 RNA-seq; Homo sapiens; RNA-Seq	SRX3045392	GSM2720328	Homo sapiens	ATF6 safeguards organelle homeostasis and cellular aging in human mesenchymal stem cells	PRJNA396193	10306775	101	WAPC	p5	vehicle
SRR5878080	GSM2720319: ATF6+/+ hMSC late passage-2 RNA-seq; Homo sapiens; RNA-Seq	SRX3045383	GSM2720319	Homo sapiens	ATF6 safeguards organelle homeostasis and cellular aging in human mesenchymal stem cells	PRJNA396193	8182392	143.61	MSC	p10	NA
SRR5878079	GSM2720318: ATF6+/+ hMSC late passage-1 RNA-seq; Homo sapiens; RNA-Seq	SRX3045382	GSM2720318	Homo sapiens	ATF6 safeguards organelle homeostasis and cellular aging in human mesenchymal stem cells	PRJNA396193	6753082	147.91	MSC	p10	NA
SRR5878078	GSM2720317: ATF6-/ hMSC late passage-2 RNA-seq; Homo sapiens; RNA-Seq	SRX3045381	GSM2720317	Homo sapiens	ATF6 safeguards organelle homeostasis and cellular aging in human mesenchymal stem cells	PRJNA396193	7782173	147.47	MSC	p10	NA
SRR5878077	GSM2720316: ATF6-/ hMSC late passage-1 RNA-seq; Homo sapiens; RNA-Seq	SRX3045380	GSM2720316	Homo sapiens	ATF6 safeguards organelle homeostasis and cellular aging in human mesenchymal stem cells	PRJNA396193	8331869	145.24	MSC	p10	NA
Run ID	Run Title	Experiment ID	EXTERNAL_ID	specie	study_title	project ID	nreads	read_length	cell type	passages	treatment
SRR6680317	GSM2983811: hMSC_TAZ_3; Homo sapiens; RNA-Seq	SRX3656633	GSM2983811	Homo sapiens	Upregulation of FOXD1 by YAP alleviates senescence and osteoarthritis	PRJNA433339	10429618	150	MSC		TAZ KO
SRR6680316	GSM2983810: hMSC_TAZ_2; Homo sapiens; RNA-Seq	SRX3656632	GSM2983810	Homo sapiens	Upregulation of FOXD1 by YAP alleviates senescence and osteoarthritis	PRJNA433339	11433785	150	MSC		TAZ KO
SRR6680315	GSM2983809: hMSC_TAZ_1; Homo sapiens; RNA-Seq	SRX3656631	GSM2983809	Homo sapiens	Upregulation of FOXD1 by YAP alleviates senescence and osteoarthritis	PRJNA433339	11009653	150	MSC		TAZ KO
SRR6680314	GSM2983808: hMSC_YAP_3; Homo sapiens; RNA-Seq	SRX3656630	GSM2983808	Homo sapiens	Upregulation of FOXD1 by YAP alleviates senescence and osteoarthritis	PRJNA433339	11196860	150	MSC		YAP KO
SRR6680313	GSM2983807: hMSC_YAP_2; Homo sapiens; RNA-Seq	SRX3656629	GSM2983807	Homo sapiens	Upregulation of FOXD1 by YAP alleviates senescence and osteoarthritis	PRJNA433339	11261647	150	MSC		YAP KO
SRR6680312	GSM2983806: hMSC_YAP_1; Homo sapiens; RNA-Seq	SRX3656628	GSM2983806	Homo sapiens	Upregulation of FOXD1 by YAP alleviates senescence and osteoarthritis	PRJNA433339	10926717	150	MSC		YAP KO
SRR6680311	GSM2983805: hMSC_WT_3; Homo sapiens; RNA-Seq	SRX3656627	GSM2983805	Homo sapiens	Upregulation of FOXD1 by YAP alleviates senescence and osteoarthritis	PRJNA433339	11048601	150	MSC		WT
SRR6680310	GSM2983804: hMSC_WT_2; Homo sapiens; RNA-Seq	SRX3656626	GSM2983804	Homo sapiens	Upregulation of FOXD1 by YAP alleviates senescence and osteoarthritis	PRJNA433339	10156337	150	MSC		WT
SRR6680309	GSM2983803: hMSC_WT_1; Homo sapiens; RNA-Seq	SRX3656625	GSM2983803	Homo sapiens	Upregulation of FOXD1 by YAP alleviates senescence and osteoarthritis	PRJNA433339	10341635	150	MSC		WT
Run ID	Run Title	Experiment ID	EXTERNAL_ID	specie	study_title	project ID	nreads	read_length	cell type	passages	treatment
SRR5357818	GSM2544142: MSC 211 CytoD; Homo sapiens; RNA-Seq	SRX2653659	GSM2544142	Homo sapiens	Osteogenic programming of adipose-derived mesenchymal stem cells using a fungal metabolite that suppresses the Polycomb protein EZH2	PRJNA379707	25484965	51	adipose derived mesenchymal stem cells	5-7	CytoD
SRR5357817	GSM2544141: MSC 211 Ctrl; Homo sapiens; RNA-Seq	SRX2653658	GSM2544141	Homo sapiens	Osteogenic programming of adipose-derived mesenchymal stem cells using a fungal metabolite that suppresses the Polycomb protein EZH2	PRJNA379707	26464456	51	adipose derived mesenchymal stem cells	5-7	Ctrl

Table_S4 metadata for kmer research in MSC in different conditions

SRR	GSM	SRX	GSM	Homo	Osteogenic programming of adipose-derived mesenchymal stem cells using a fungal metabolite that suppresses the Polycomb protein EZH2.	PRJNA379707	24280737	51	adipose derived mesenchymal stem cells	5-7	CytoD
SRR5357816	GSM2544140: MSC 283 CytoD; Homo sapiens; RNA-Seq	SRX2653657	GSM2544140	Homo sapiens	Osteogenic programming of adipose-derived mesenchymal stem cells using a fungal metabolite that suppresses the Polycomb protein EZH2.	PRJNA379707	21040712	51	adipose derived mesenchymal stem cells	5-7	Ctrl
SRR5357815	GSM2544139: MSC 283 Ctrl; Homo sapiens; RNA-Seq	SRX2653656	GSM2544139	Homo sapiens	Osteogenic programming of adipose-derived mesenchymal stem cells using a fungal metabolite that suppresses the Polycomb protein EZH2.	PRJNA379707	35139902	51	adipose derived mesenchymal stem cells	5-7	CytoD
SRR5357814	GSM2544138: MSC 258 CytoD; Homo sapiens; RNA-Seq	SRX2653655	GSM2544138	Homo sapiens	Osteogenic programming of adipose-derived mesenchymal stem cells using a fungal metabolite that suppresses the Polycomb protein EZH2.	PRJNA379707	34623656	51	adipose derived mesenchymal stem cells	5-7	Ctrl
SRR5357813	GSM2544137: MSC 258 Ctrl; Homo sapiens; RNA-Seq	SRX2653654	GSM2544137	Homo sapiens	Osteogenic programming of adipose-derived mesenchymal stem cells using a fungal metabolite that suppresses the Polycomb protein EZH2.	PRJNA379707	17969174	44	human Mesenchymal Stem Cells	p4-p10	10d osteo Diff
Run ID	Run Title	Experiment ID	EXTERNAL_ID	specie	study_title	project ID	nreads	read_length	cell type	passages	treatment
SRR8447838	GSM3564546: 10 days diff Control siRNA-C RNA-seq; Homo sapiens; RNA-Seq	SRX5254873	SAMN10755603	Homo sapiens	PLZF targets developmental enhancers for activation during osteogenic differentiation of human mesenchymal stem cells (RNA-seq)	PRJNA515466	18943085	44	human Mesenchymal Stem Cells	p4-p10	10d osteo Diff
SRR8447837	GSM3564545: 10 days diff Control siRNA-B RNA-seq; Homo sapiens; RNA-Seq	SRX5254872	SAMN10755604	Homo sapiens	PLZF targets developmental enhancers for activation during osteogenic differentiation of human mesenchymal stem cells (RNA-seq)	PRJNA515466	26461685	44	human Mesenchymal Stem Cells	p4-p10	10d osteo Diff
SRR8447836	GSM3564544: 10 days diff Control siRNA-A RNA-seq; Homo sapiens; RNA-Seq	SRX5254871	SAMN10755605	Homo sapiens	PLZF targets developmental enhancers for activation during osteogenic differentiation of human mesenchymal stem cells (RNA-seq)	PRJNA515466	27117845	44	human Mesenchymal Stem Cells	p4-p10	10d osteo Diff
SRR8447835	GSM3564543: 2 days diff Control siRNA-C RNA-seq; Homo sapiens; RNA-Seq	SRX5254870	SAMN10755606	Homo sapiens	PLZF targets developmental enhancers for activation during osteogenic differentiation of human mesenchymal stem cells (RNA-seq)	PRJNA515466	26162294	44	human Mesenchymal Stem Cells	p4-p10	2d osteo Diff
SRR8447834	GSM3564542: 2 days diff Control siRNA-B RNA-seq; Homo sapiens; RNA-Seq	SRX5254869	SAMN10755607	Homo sapiens	PLZF targets developmental enhancers for activation during osteogenic differentiation of human mesenchymal stem cells (RNA-seq)	PRJNA515466	19724867	44	human Mesenchymal Stem Cells	p4-p10	2d osteo Diff
SRR8447833	GSM3564541: 2 days diff Control siRNA-A RNA-seq; Homo sapiens; RNA-Seq	SRX5254868	SAMN10755608	Homo sapiens	PLZF targets developmental enhancers for activation during osteogenic differentiation of human mesenchymal stem cells (RNA-seq)	PRJNA515466	24266576	44	human Mesenchymal Stem Cells	p4-p10	Naive
SRR8447832	GSM3564540: NaiveControl siRNA-C RNA-seq; Homo sapiens; RNA-Seq	SRX5254867	SAMN10755609	Homo sapiens	PLZF targets developmental enhancers for activation during osteogenic differentiation of human mesenchymal stem cells (RNA-seq)	PRJNA515466	25910627	44	human Mesenchymal Stem Cells	p4-p10	Naive
SRR8447831	GSM3564539: NaiveControl siRNA-B RNA-seq; Homo sapiens; RNA-Seq	SRX5254866	SAMN10755610	Homo sapiens	PLZF targets developmental enhancers for activation during osteogenic differentiation of human mesenchymal stem cells (RNA-seq)	PRJNA515466	24849706	44	human Mesenchymal Stem Cells	p4-p10	Naive
Run ID	Run Title	Experiment ID	EXTERNAL_ID	specie	study_title	project ID	nreads	read_length	cell type	passages	treatment
SRR3882937	GSM2231443; AMSC_19_proliferating; Homo sapiens; RNA-Seq	SRX1939121	GSM2231443	Homo sapiens	Transcriptome Sequencing of Adipose-Derived Mesenchymal Stromal Cells	PRJNA328824	29447967	50			
SRR3882936	GSM2231442; AMSC_19_confluent; Homo sapiens; RNA-Seq	SRX1939120	GSM2231442	Homo sapiens	Transcriptome Sequencing of Adipose-Derived Mesenchymal Stromal Cells	PRJNA328824	43754165	51	Adipose-derived Mesenchymal Stromal Cell		
SRR3882935	GSM2231441; AMSC_4_proliferating; Homo sapiens; RNA-Seq	SRX1939119	GSM2231441	Homo sapiens	Transcriptome Sequencing of Adipose-Derived Mesenchymal Stromal Cells	PRJNA328824	34428851	50	Adipose-derived Mesenchymal Stromal Cell		
SRR3882934	GSM2231440; AMSC_4_confluent; Homo sapiens; RNA-Seq	SRX1939118	GSM2231440	Homo sapiens	Transcriptome Sequencing of Adipose-Derived Mesenchymal Stromal Cells	PRJNA328824	42327600	51	Adipose-derived Mesenchymal Stromal Cell		
SRR3882933	GSM2231439; AMSC_3_proliferating; Homo sapiens; RNA-Seq	SRX1939117	GSM2231439	Homo sapiens	Transcriptome Sequencing of Adipose-Derived Mesenchymal Stromal Cells	PRJNA328824	33767562	51	Adipose-derived Mesenchymal Stromal Cell		
SRR3882932	GSM2231438; AMSC_3_confluent; Homo sapiens; RNA-Seq	SRX1939116	GSM2231438	Homo sapiens	Transcriptome Sequencing of Adipose-Derived Mesenchymal Stromal Cells	PRJNA328824	44801550	51	Adipose-derived Mesenchymal Stromal Cell		
SRR3882931	GSM2231437; AMSC_1_proliferating; Homo sapiens; RNA-Seq	SRX1939115	GSM2231437	Homo sapiens	Transcriptome Sequencing of Adipose-Derived Mesenchymal Stromal Cells	PRJNA328824	37501459	50	Adipose-derived Mesenchymal Stromal Cell		
SRR3882930	GSM2231436; AMSC_1_confluent; Homo sapiens; RNA-Seq	SRX1939114	GSM2231436	Homo sapiens	Transcriptome Sequencing of Adipose-Derived Mesenchymal Stromal Cells	PRJNA328824	34863048	51	Adipose-derived Mesenchymal Stromal Cell		
Run ID	Run Title	Experiment ID	EXTERNAL_ID	specie	study_title	project ID	nreads	read_length	cell type	source_name	treatment
SRR8099005	GSM3444217; EGF_IWR1_S03; Homo sapiens; RNA-Seq	SRX4925800	GSM3444217	Homo sapiens	Epidemal growth factor activates β -catenin via integrin-linked kinase to control proliferation of mesenchymal stromal cells.	PRJNA498109	14005865	151	Mesenchymal stromal cells	femoral heads/knee bone explants	EGF + IWR-1
SRR8099004	GSM3444216; EGF_IWR1_S02; Homo sapiens; RNA-Seq	SRX4925799	GSM3444216	Homo sapiens	Epidemal growth factor activates β -catenin via integrin-linked kinase to control proliferation of mesenchymal stromal cells.	PRJNA498109	11575790	151	Mesenchymal stromal cells	femoral heads/knee bone explants	EGF + IWR-1
SRR8099003	GSM3444215; EGF_IWR1_S01; Homo sapiens; RNA-Seq	SRX4925798	GSM3444215	Homo sapiens	Epidemal growth factor activates β -catenin via integrin-linked kinase to control proliferation of mesenchymal stromal cells.	PRJNA498109	15666602	151	Mesenchymal stromal cells	femoral heads/knee bone explants	EGF + IWR-1
SRR8099002	GSM3444214; EGF_S03; Homo sapiens; RNA-Seq	SRX4925797	GSM3444214	Homo sapiens	Epidemal growth factor activates β -catenin via integrin-linked kinase to control proliferation of mesenchymal stromal cells.	PRJNA498109	15786160	151	Mesenchymal stromal cells	femoral heads/knee bone explants	EGF
SRR8099001	GSM3444213; EGF_S02; Homo sapiens; RNA-Seq	SRX4925796	GSM3444213	Homo sapiens	Epidemal growth factor activates β -catenin via integrin-linked kinase to control proliferation of mesenchymal stromal cells.	PRJNA498109	19843529	151	Mesenchymal stromal cells	femoral heads/knee bone explants	EGF
SRR8099000	GSM3444212; EGF_S01; Homo sapiens; RNA-Seq	SRX4925795	GSM3444212	Homo sapiens	Epidemal growth factor activates β -catenin via integrin-linked kinase to control proliferation of mesenchymal stromal cells.	PRJNA498109	19132337	151	Mesenchymal stromal cells	femoral heads/knee bone explants	EGF
SRR8098999	GSM3444211; untreated_S03; Homo sapiens; RNA-Seq	SRX4925794	GSM3444211	Homo sapiens	Epidemal growth factor activates β -catenin via integrin-linked kinase to control proliferation of mesenchymal stromal cells.	PRJNA498109	13571062	151	Mesenchymal stromal cells	femoral heads/knee bone explants	untreated
SRR8098998	GSM3444210; untreated_S02; Homo sapiens; RNA-Seq	SRX4925793	GSM3444210	Homo sapiens	Epidemal growth factor activates β -catenin via integrin-linked kinase to control proliferation of mesenchymal stromal cells.	PRJNA498109	16677487	151	Mesenchymal stromal cells	femoral heads/knee bone explants	untreated
SRR8098997	GSM3444209; untreated_S01; Homo sapiens; RNA-Seq	SRX4925792	GSM3444209	Homo sapiens	Epidemal growth factor activates β -catenin via integrin-linked kinase to control proliferation of mesenchymal stromal cells.	PRJNA498109	17232973	151	Mesenchymal stromal cells	femoral heads/knee bone explants	untreated

Table_S5 genes names positively enriched in MSC

gene_name
 KRTAP1_5
 FAM180A
 GALNT5
 CCKAR
 LINC01423
 KCNK2
 ITGBL1
 APCDD1L
 BDKRB2
 FGf7
 LRRK15
 PTGFR
 AL355607.2
 KCNK15
 FGf5
 SLC14A1
 MYLK
 SMILR
 GDF5
 SBSN
 RFX8
 CA12
 IL13RA2
 EYA2
 C2CD6
 DLX5
 PENK
 AL359834.1
 C1S
 RAB27B
 Z97200.1
 AC022467.1
 AL160153.1
 FAP
 NTM
 FAM133CP
 FOXD1
 C11orf87
 ISLR
 SFRP4
 HOXC10
 PRR16
 HOXC11
 AC091182.2
 FAM25B
 LINC02454
 LAMA3
 THBS2
 PDGFRA
 LIMA1
 AC007336.3
 BICC1
 OLFM1
 LINC01605
 KIAA1755
 AC105046.1
 IGFBP6
 SATB2
 PDGFRB
 CLCA2
 GPR1
 BLID
 EGR1
 FGF10
 PAMR1
 LINC00968
 PCOLCE
 POPDC3
 C1R
 CCN5
 COL8A1
 SSPN
 APCDD1L.DT
 AGAP11
 SOCS5
 AL355102.1
 AP003071.4
 MOXD1
 CLMP
 GLYATL2
 NTNG1
 AC114284.1

Table_S5 genes names positively enriched in MSC

LMO7
 RF01891
 PQLC2L
 CCDC68
 TNFRSF11B
 COL6A3
 PTHLH
 PRRX1
 PAM
 PALM2.AKAP2
 SIX2
 AC099066.2
 MIR31HG
 PDE1C
 SLC16A4
 VGLL3
 EGR3
 ABCA8
 SERPINE2
 ITGA11
 IL1RL2
 PTGES
 AC010533.1
 ANXA10
 FOXF2
 HSPB2
 WNT5A
 PPP1BP1
 CCN6
 AL118522.1
 GREM1
 HOATAIR
 FMN2
 BEND6
 ROS1
 IL6
 AC041040.1
 AC110597.1
 RANBP9L
 LM07.AS1
 LINC01614
 PABPC1P4
 ANTXR2
 SVIL
 DDR2
 KANK2
 HSPB2.C11orf52
 MIR3609
 LTB2
 COL14A1
 NCAM2
 TSLP
 NEGR1
 HAS1
 ANCPTL5
 DCBLD2
 Z99289.2
 CCN4
 FLG
 IFNE
 ADAMTS6
 XG
 ABCA6
 SEMA5A
 COL6A2
 LINC01060
 SNAI2
 SHISAL1
 SCARA5
 FNDC1
 TNC
 PAPPA
 CYGB
 ADAM12
 FAM87B
 TMEM119
 GPR29
 PTGIS
 LEPR
 BX005019.1
 TMEM233
 PCDHGA12
 VIT

Table_S5 genes names positively enriched in MSC

DSEL
DIRAS3
GLTB2
CFAP300
AC012531.2
PKD2
IGFBP3
MEDAG
PNLIPRP3
ANKRD29
ADAMTS5
PLAT
CXCL12
TFPI2
KCND2
CCDC80
TMEM200A
SVEP1
ADGRl4
MFAP5
CYBRD1
ALDH1A3
RPL35P6
STC1
AC093627.6
IRX3
FIBIN
CHL1
NEAT1
STEAP2
SYNPO2
GFR1
ETV1
TBX15
PLA2R1
LMOD1
NTSE
PDE7B
HOXC9
ECM2
NEXN
ADAMTS1
LAYN
VSTM4
SIX1
AOX1
ANKRD35
C1GALT1
AC079949.2
SERTAD4
SCUBE3
TWIST1
HSPB7
VEGFC
CD248
STMN2
THY1
LINC02029
AC087821.1
MAP1A
TMEM158
ZNF385D
STEAP1
DYNC2H1
SPOCK1
GEM
MRGPRF
EBF1
LBP
WNT5B
CDCP1
FAM3C
LOXL1
HSPB6
GXYLT2
RN06_1045P
AC012531.1
GBE1
COL12A1
HOXC6
BAALC
C11orf96
KCTD16

Table_S5 genes names positively enriched in MSC

DAAM2
CFAP299
LRRN4CL
CALHM5
CDR1
LINC00601
LPAR1
HEG1
AH1
HAS2
UF
CTHRC1
AC006058.1
PHLDA1
FOSL1
RNF182
RN6_530P
RBPJ
OLFML2A
EYA1
AC112721.2
LYNX1
ADAMTS2
DCN
EDIL3
AC234772.2
MAP4K4
TIMP1
COL1A2
CRYAB
PSD3
AL359095.1
TNFAIP6
C1QTNF1
CTSK
GLIS1
PPL
GREM2
NLRP10
SMIM3
ITGB1
PRRX2
PAPPAA_S2
ANKRD28
FMO2
HSPD1P11
GLRB
MEOX2
ZCCHC24
MKX
NFX
AC022034.1
AC113383.1
TRAM2
IRX5
CCDC110
PEAK1
DIO2
SCG2
VEGFA
BMERB1
C8orf48
MIR100HG
CDH11
MCFD2
GAS1RR
AC107308.1
EPS8
FOXC2
CARMN
SLT2_IT1
GALNT1
USP53
SPATA18
ENDDO1
SNED1
B3GNT9
SGIP1
AC011586.2
SLC5A3
COL10A1
ABC49
LINC02202

Table_S5 genes names positively enriched in MSC

PACERR
PLA2G5
FBN1
SOD3
COL6A1
RORB
GABRE
ABI3BP
ACTA2
MIR137HG
CORO2B
ADIRF
TLC2D
LAMA4
CALB2
OLFML3
INHBA
ZFHX4
GDNF
ARHGAP21
FRMPD4
GSC
BDNF
NFIC
HIP1A
PCDHGA7
TAGLN
TGFB1
AL121748.1
BDKRB1
KY
KCN52
AC015922.3
HTRA1
CTF1
Z99289.3
CFH
HHPL2
NPR3
LINC01705
ELOVL3
OSMR
KCTD4
ITGA10
SGCD
SLC25A27
ECM1
SH2D4A
EMILIN1
LRRTM2
LINC01119
ID4
DST
LURAP1L
STEAP4
AC117489.1
GDF6
PEX5L
TIMP3
ANTXR1
MIR222HG
GAS6.DT
CNTNAP4
MXRA5
SERTA4.AS1
HOXC5
SYTL5
IIFTM10
DOK5
ARID5B
SLC1A1
HEPH
NID2
SNORD114.1
FHL2
TBX2
HOXC8
PDZRN3
AP000525.1
CCBE1
REP15
SYNC
MT1L

HGF
FRMD6
RHOBTB3
PYGO1
TBX18
DGKI
SEC24AP1
NAALADL2
TMEM132B
NA.
AC092807.3
BAG2
GSTM2
VASN
POM121L9P
ZFHX4.AS1
CIART
FAM20A
FGF1
SLC2A12
AC027335.1
EPHAS5
RAB23
CEP170
CPEB1
EDNRA
DNALI1
NNMT
AC016397.1
AC007673.1
LUM
CDKN2B
MYL9
IL7
RCAN1
SGCB
PTGS2
KCNJ15
EPDR1
ZMAT3
BNIP3L
VCAN
RN7SKP11
BRINP1
EBF3
LTBP3
QSOX1
PPP1R3C
HMCN1
SCN2A
BMPER
COL16A1
TENM4
AC058822.1
PXK
PLPPR4
POSTN
FOXC1
SNX9
RPSAP52
SATB2.AS1
TPM2
TTC8
LINC02511
AP001434.1
RASSF8
PRKD1
GRAMD2B
LINC01050
SNAP25
AC092691.1
ASPH
PPFIA2
SHOX2
GRIA3
RFTN2
LHFPL6
RNU6_1226P
CASC4
SRPX
FAM114A1
NBL1
TM9SF3

Table_S5 genes names positively enriched in MSC

Table_S5 genes names positively enriched in MSC

PCDHGB7
 COMP
 LRP1
 CMY5
 TWIST2
 MICU1
 NEK7
 LEP
 PALM2
 SNAPC1
 RSU1P3
 INMT.MINDY4
 SNHG18
 SPAG16
 SH3PX2A
 ALPK2
 EPB41L2
 LINC00973
 MEIS3P1
 CLGN
 PCOLCE2
 AC134043.2
 PCDHGA3
 FST
 GAS6
 IGFBP4
 ADAMTS1
 SERP2
 ARHGAP20
 PROS1
 SYNDIG1
 EMX2OS
 SFRP1
 CFAP69
 UACA
 NPHP3.ACAD11
 LTBP1
 SETD7
 HRH1
 FHOD3
 GAS1
 GGT5
 PCDHGB3
 BHLE41
 ABCC9
 SUGCT
 MTMR11
 CEMIP
 TFP1
 EEF1B2P3
 KCNJ8
 ANGPT1
 EFEMP2
 RGS4
 CCDC89
 SYAP1
 AC112229.3
 EGFR
 LINC01139
 EGFLAM
 FMOD
 AL359091.2
 LYPD6B
 PCDHGA10
 STEAP1B
 AHR
 SULF1
 TNFSF11
 HOXC4
 SRPX2
 EHD2
 CST1
 MXRA7
 PNMA2
 TREM1
 IGFBP5
 TMEM26
 AP000526.1
 FGF2
 KCNMA1
 SMURF2
 AC093908.1
 KTN1

Table_S5 genes names positively enriched in MSC

CUL4B
CALD1
IGF1
MSRB3
NTRK2
GRIN2A
EY44
GALNT15
FMO3
MAGI2
AC096537.1
DYNCL2
CPXM2
DTWD1
TMEM100
AC112198.2
MXRA8
NFL3
ATP8B1
NR4A1
NME5
NET1
AC092645.1
AGTR1
AC107021.2
SSC5D
CHI3L1
PCDHB7
PABPC5
RPL13A/P5
ARHGAP12
CCN3
MMP2
TNKS1BP1
STXBP5
CRNDE
PLEKHH2
JCAD
DDO
RGS17
AC013268.1
PTX3
CAV1
ANGPTL2
NFASC
SLC6A15
WDR63
LINC00632
SLC4A4
LNC8RLR
DLX2
MX1
SMIM10
PCDH18
P4HA1
BTC
OLFML2B
NHS
LINC00839
CREB3L1
NAV3
PCDHGB5
STXBP1
HMGB1P21
DOCK5
RIMS1
PRPH2
ARHGAP6
PDGFc
CERCAM
TWSG1
MIR8485
FER1L6
NKX3.2
PRUNE2
HYDIN2
PCDHGC3
TMEM47
ERRFI1
MLPH
AC008163.1
FZD1
PCDHGA2

Table_S5 genes names positively enriched in MSC

HOXA13
 LOXL2
 GJB2
 RASSF9
 NUCB2
 TIPARP
 AL391822.1
 AC026124.1
 DUXAP8
 SNORD114.7
 CFTP12
 C5orf58
 CCDC74B
 CPA4
 BHMT2
 AP000892.3
 CEP126
 LRATD1
 CCDC170
 SPTLC3
 ANKH
 TK2
 OSR1
 EGLN1
 GLIS3
 NOTCH3
 C14orf39
 PCDHGA4
 IFI44L
 RHOQ
 RF00181
 EXT1
 TSC22D1
 SMC02
 GNA14
 EFS
 CCPG1
 PCDHGB2
 BVES
 ZC2HC1A
 ARMC9
 ALDH1L2
 AC073591.1
 TCEAL7
 FAT1
 GJA1
 SEMA3C
 FGFR1
 CPEB2
 AC006213.4
 PLCE1AS1
 IFIT1
 DNMM3OS
 ENPP2
 AC007750.1
 AC002480.1
 AZ12
 FAM155A.1T1
 OSR2
 TRIOBP
 CFAP36
 FNDC3B
 AKR1C2
 Z93022.1
 VDR
 ADORA1
 GPC6
 PTprm
 AC012321.1
 THAP12P7
 DSTN
 ESM1
 FUT11
 RCAN2
 COL1A1
 REEP3
 KIAA1549L
 AC004264.1
 FSTL1
 EMP1
 C4orf47
 ITPR1PL2
 SNORD114.14

MYPN
PLEKHA5
TENT5A
BANK1
PODN
C8orf88
IL1R1
IGFBP7
ZNF280D
C12orf56
ADAMTS9.AS2
AC080038.1
AC016397.2
KCN51
NR2F1.AS1
AEBP1
AC022509.3
IRX6
SRGAP1
EDA2R
LINC01106
PCDHGC5
FOXL1
GLS
LINC02381
TMEM92.AS1
CAST
LINC01117
NRN1
HECW2
RPS12P5
FN1
AC073115.2
PDLIM5
SCARA3
TRPC4
RAET1G
TMEM171
NR0B1
HOXA2
GPR176
MRPS6
ANO3
FAM171B
FAM155A
SPAG9
MYO1B
VCAM1
CYP1B1
FBXO32
COL21A1
AL035078.1
AL354811.1
KCNE4
CHRD1
AC009093.1
SLC9A7
FAM13C
CYP2U1
RNF180
MME
FANK1
EHBP1
PRICKLE2
SLC39A6
CACNA2D1
THSD4
SOX9
TNFRSF19
PDE5A
GOLGA2
DYRK3
DLC1
NDUFA4L2
C14orf132
EVC
AC138969.1
HMGA2.AS1
SLC2A10
AL137026.1
CEP290
SNX33
SMOC1

Table_S5 genes names positively enriched in MSC

AC093535.2
LXN
ACOX2
BNC2
NALCN
LOXL4
ACSS3
NBPF10
CAV2
MAP3K20
DIXDC1
RECK
MET
AC013565.1
MAGI2.AS3
FER
VCAN.AS1
TCEAL4
SNORD11A.4
CACNA1A
GOS2
PLAC9
RNU6.786P
KLF10
PLSCR4
BCL2L2.PABPN1
COL11A1
TTL17
HDLBP
TP63
ANGPTL1
ZNF214
TNS2
TDRD6
NEK10
NRP1
CSGALNACT2
TCEAL3
HHIP1
SPOCD1
ADD3
CHST3
DSE
EGR2
TM4SF1
CACNA1C
ARHGAP22
MIR22HG
AXL
ARSJ
BNC1
B3GALNT1
LARP6
ELL2
CAPN2
PTPN14
DIRAS1
AKR1C1
SEMA3A
HK2
ENAH
COP22
WDR19
CLDN11
RND3
LRP11
NGF
CD109
BX322650.1
VEGFB
TLN2
SEC24D
ARNT2
FKBP7
ATOH8
NFATC4
MEG3
TMEM263
PRCD
PCYQX1
ASAP3
SMAD6
PCDHGA11

Table_S5 genes names positively enriched in MSC

Table_S5 genes names positively enriched in MSC

GPX8
BMS1P7
HSD11B1
AL157702.2
HCG11
ARHGAP23
RPS3AP6
WDFY3.AS2
AC244669.2
TPBG
SLC22A3
COL3A1
LRP12
CDH13
ROR1
SLC20A1
KRT16
NA..1
EEA1
AC096531.2
KIRREL1
CDKN2A
ANK2
ACAN
ENOX1
REXO2
CEP63
AC008269.1
AC233263.6
RNF24
LNCTAM34A
HTR7
AC090527.2
VEPH1
LINC00702
NRXN2
PDE1A
DZIP1
MAB21L2
TMTC1
NKAPL
CCND1
LAMC1
FAM126A
ITGA7
SNORD114.2
ISM1
CCDC158
TCEAL9
LINC01303
TRH
CYS1
AC008522.1
PARP3
LDHA
MAP9
UHRF2
ABC85
NRG2
RILPL1
EREG
DYNLT3
PLOD2
RNASE4
ABCA13
AKR1C3
ANKRD30B
NEK11
RN7.119P
ARMH4
PDE3A
FOXC2.AS1
IKBIP
ARHGEF12
ANKRD53
SUCLG2.AS1
RN7SL396P
MDM2
IFIT3
PPP1R14C
FAS
AKAP2
AC069079.1

Table_S5 genes names positively enriched in MSC

CLEC3B
PTCHD4
CSF1
TCAF1
FAM225A
CALU
AL162171.1
B4GALT1
PCDHGA6
DLX6
MTCL1
PPARGC1A
BHLE40
WVTR1
AC098650.1
SCARB2
C19orf12
TIMP2
CDK15
CCDC8
SPIRE1
TTC3
FKBP9
HFE
STAR D13.1T1
P4HA3
AC245297.1
AMOTL1
RAB13
KITLG
CCDC74A
LINC00595
CCN1
PRSS23
ZNF471
HOXA10
AP001528.2
ENTPD3
NTN4
BIVM
CAMK2N1
AC093899.2
PDE4DIP
CCL28
CSPG4
FOSB
FYCO1
AP001970.1
MMP16
NPHP1
DPT
DUXAP10
CAVIN1
HERC4
MIR635
UBE2E3
CRTAP
ADAMTS14
TMCO3
STK17A
ARHGAP24
AL356432.3
TMF1
NPM1P40
IFT46
STC2
TMEM246
TPM1
SNORD113.2
AL157935.2
AC027277.1
TBC1D19
ST13P18
KAZALD1
NPIP9
PRKG1
ITGB8
FLOT1
ZBTB4
AC140479.4
GSTMS
SERPINF1
PTPRG

Table_S5 genes names positively enriched in MSC

NAP1L1
AHNAK2
SPRY4
LCA5
CHMP3
PRKAR1A
ROCK2
GOLGB1
HAS2_AS1
AC090229.1
TPM4
PAQR5
FBXL7
AC107982.1
DRP2
CTSA
AL049838.1
ITPRIP
RAB31
MINDY2
TICAM2
AC037198.1
SHC1
NAB2
RARG
BARX1
LOX
RF00181.1
BMPR2
PLXNA4
KIAA1614
LDB2
POGLUT3
AL391427.1
SNORD114.6
NKX3.1
SGMS2
NR2F2_AS1
CA9
SEC22B
SMIM14
PCDHGB1
BCAP29
AC048341.1
FZD7
GNG12
GAP43
CH25H
THBS3
PPFIA4
RPS6KA6
THBS1
CCL13
GDNF_AS1
AC008914.1
IRS1
EEF1A1P16
ARMCX2
DNAJB4
EMP2
STS
PCOLCE_AS1
SORBS2
TRAM111
MSANTD3
STRIP2
NDP
CAMK2D
MEG8
SPESP1
FAM66B
SLC16A2
BBOX1_AS1
KLF4
AC010655.3
WIP1
PARVA
IFT81
AK5
CC2D2A
LINC00472
GLIS2
COX20

Table_S5 genes names positively enriched in MSC

PCDHGB4
COLEC12
RAI14
CEP162
NRG1
ITGAV
RPSAP48
SEMA3E
PLCE1
MYOF
RSPH3
LRRK32
NAV1
ARMCX3
TRPA1
AC121247.1
MAP4K5
DUXAP9
SNTB2
RTN4
LDOC1
AL590560.2
LINC02516
SERPINE1
PPIL6
AC015712.2
TMED7.TICAM2
CLIP3
DLX1
CPZ
SETP6
SERINC1
IGIP
RNU6.767P
SEC16B
SLT3
CPED1
GINM1
FAM20C
FZD8
CTSO
YAP1
PRG4
ESR1
PTN
CCN2
CYP27C1
RASD2
TGFb2
TRIO
ZNF229
STAT2
TMTc3
CFL2
DCUN1D3
TTC30A
PDCD1LG2
PPP3CA
ARHGAP35
EFEMP1
GOLM4
ACSM5
AC079594.2
AC025423.1
CHSY3
MOK
AL021068.1
DZIP3
PLPP7
KIAA1217
TGFb11
MIR1245A
FAM210B
HMGN2P5
TCEA3
AC025259.3
DEPDC7
LUCAT1
RBOFOX2
ERO1A
EMX2
ZHXB
LINC00565

Table_S5 genes names positively enriched in MSC

SLC4A7
ATP2B4
TLR3
LINC01293
PTPN20
MBNL1.AS1
COL25A1
SLC13A4
ABLM3
ST5
HCFC1R1
AL022068.1
MYOZ2
AC008771.1
PTPRQ
MBOAT2
AC099494.2
SHISA4
LSM1P1
AL137024.1
HRCT1
AC100854.1
BACE1
EN1
AC021205.1
RBMS2
CDKL2
SLP1
AC092111.2
EVC2
RCN3
UNCSC
PRELP
C12orf75
AL117190.1
IFI6
TTC3P1
CTTN
FAM129B
AC110597.3
HTR2A
SAMD11
ARMCX1
TEAD1
STAC
PHLDB1
FGF14
SPEG
CBR3
ITGA1
SIX4
HTATSF1P2
SPRY4.AS1
OSBPL8
SEC31A
PTPRK
DLX3
ZNF662
NR4A2
ITGB5
C1orf198
RP1
KIF1BP
MBTPS2
TP1P1
RABGEF1
PRDM16
AL137161.1
PTPRN
DTNA
AFF4
MRV11
MEF2A
AC027031.2
FGFR11
NCKAP5
GCShp3
STX1A
ACP7
LRRK66
ZDHHC7
FOS
RBMS3

Table_S5 genes names positively enriched in MSC

AC108734.4
 PER3
 CLCF1
 NEK1
 LINC01123
 MT1M
 DKK1
 LAMB2
 AC105114.1
 AL049795.2
 YWHAE
 ZBTB38
 RF00100
 FIP1L1
 AL133346.1
 AD000090.1
 LIN7C
 PPP1R12A
 TSHZ3
 THSD1
 EIF2AK2
 TBX2.AS1
 PRDM8
 SLC41A1
 PTGDS
 ARHGEF28
 MIR497HG
 CDH2
 AC084757.3
 PDLM4
 DKK3
 RMDN2
 PAPPA.AS1
 C1RL
 LAMB1
 MFSD1
 EML1
 MTMR9LP
 CNTLN
 MMP14
 ARHGAP22.IT1
 GNP NAT1
 FZD6
 RPLP0P2
 AC011611.3
 LACC1
 UGDH
 ZFPFM2.AS1
 SNTG2.AS1
 CDK14
 MCC
 CYB5R3
 APP
 AC026785.2
 DCLK1
 RRM2B
 GLIPR1
 WNT16
 PCSK1
 THNSL2
 MIER1
 PLPP3
 KCTD1
 KIF5B
 SPIN1
 AF127577.4
 SGCE
 EOGT
 RASA1
 SNORD113.1
 NCOA7
 DCDC1
 ADH1B
 CCDC149
 HNRNPLP1
 UNCSB
 C5orf24
 AC110048.2
 DAB2
 MAP1B
 MAMLD1
 BOC
 PTPN13

Table_S5 genes names positively enriched in MSC

VCL
 NFAT5
 XYLT1
 LINC01844
 SYNM
 RDD1
 MAT2A
 SLC30A4
 NR4A3
 MFAP4
 NDRG1
 AC241952.1
 RGN
 IFIT5
 PRICKLE1
 ANKRD42
 FERM2
 ZFP91
 CSPG4P12
 RGMB
 HOXA10.AS
 EXT2
 TULP3
 TSPYLS
 LINC01268
 SNORD114.3
 COL4A2
 SLT2
 HES1
 IFI16
 BNC2.AS1
 FBLN1
 TWF1
 NFTB
 RAB9B
 AKAP11
 ENPP1
 AL163636.2
 RCN1P2
 DCUN1D4
 EVA1A
 LINC01600
 PPP2R3A
 TUSC3
 FOXQ1
 A4GALT
 ATP9A
 FEZ1
 FAM167B
 C7
 SMAD9
 LRIG3
 RN7SL602P
 RCN1
 NBPF14
 CTTNBP2NL
 PRSS3
 PPA2
 ERMAP
 LRRK17
 ALCAM
 CHD9
 N4BP2L2
 LRRKIP2
 AL137782.1
 ZBTB47
 AL162386.2
 LINC00682
 AL645922.1
 KCNT2
 CNH3
 AC092104.1
 NPY2R
 HLX
 AMOTL2
 PLCD4
 SAMD8
 NKX6.1
 AR
 AC104051.2
 AC013271.1
 FAM187A
 AL592158.1

BGN
VPS41
CFAP97
MALL
APPL2
CSNK1G3
SERPING1
TCEAL1
ROBO1
STEAP3
LRRK6
BBS10
PRSS12
DUSP4
FEMIC
RAET1E
NDUFS4
PKP1
SNX7
ZBTB41
RECQL
GHR
AL583722.1
AC090197.1
CPS1
MPP4
ENG
NPTX1
KIAA0408
ZC3HAV1L
PLOD1
PLAU
SDAD1P1
B3GALT2
ULBP2
LAPTM4A
AC020909.4
SAP30
ZNF25
LAMA1
SIGLEC15
AC004080.6
ATXN3
JAK2
IER5L
SYNJ2
EBF2
SLC2A3P1
RPL23AP32
PHEX
PPI4
RN7SKP255
SYT12
CRISPLD2
AL078599.2
MLF1
HOXD8
HOXD_AS2
MARCH4
ATP10D
SERINC3
NDNF
EXTL1
SMAD7
FAT4
OPCML
ADM
ZNF302
LIFR
PIFO
ARL4C
RF00017
LGALS3BP
WDR60
ENPP5
AC142381.3
ARRDC3
FLNC
MIR4458HG
PODNL1
LRRK49
SPRED2
NORAD

Table_S5 genes names positively enriched in MSC

Table_S5 genes names positively enriched in MSC

AL445649.1
 AC108463.1
 ZNF454
 PHLDB2
 PCDHG9
 MFAP2
 EPGN
 SMAGP
 SUSD5
 INSYN2A
 STON1
 AC105105.1
 AC023908.1
 KIRREL1.T1
 AC092614.1
 PXDN
 CCDC14
 PRRT2
 RBM43
 BBOP1
 LEPROT
 CCNG2
 CRYBG3
 TAF7L
 TCTN1
 CAVIN3
 PTPRU
 ADCY8
 HTR1F
 METTL15P1
 WASHC2C
 ADGRL2
 FPGT.TNNI3K
 COL5A2
 DIP2C
 AC092168.2
 ZNF135
 TRIM22
 ITGA3
 RMDN2.AS1
 C11orf58
 PLCB4
 SERPINB2
 TMED2
 AC073115.1
 OSMR.AS1
 SEPT11
 RUNX2
 AC103796.1
 AC080038.2
 LINC01443
 LRP1B
 CCDC36
 GIGYF2
 AP002396.5
 DZIP1L
 KIF7
 AL132780.2
 MPDZ
 PRTFDC1
 ADAM9
 SOX5
 SLC8A1
 LINC02595
 EPAS1
 SNORD113.3
 RARRES1
 PHAX
 ZNF503
 CLIC4
 PHYH
 AL160408.4
 HOXA11
 LINC01426
 PCDHA3
 LMNA
 VAX2
 KIAA0825
 SPAG17
 TPD52L1
 HOXA7
 H6PD
 KIAA1109

Table_S5 genes names positively enriched in MSC

AC112229.1
 CDC14B
 RPS6KA2
 COL4A4
 PRRG1
 CYP39A1
 SOCS6
 DDAH1
 MYO1E
 TOPORS
 FAM110B
 LYPD1
 MDH1B
 AC092053.3
 PSG5
 ACKR4
 FGFRP6
 FBLN7
 PCDHG45
 MAP7D3
 ERLIN2
 LRRK2
 KCNH1
 WDR78
 C1GALT1C1L
 C15orf65
 NCKAP1
 MIA3
 FUT8
 SOBP
 TTC37
 TCTN2
 CHMP2B
 MARVELD1
 LOXL1.AS1
 SH3RF3
 ATL3
 AC008147.2
 WASF3
 DPYSL3
 ATP2B1
 COL24A1
 TSPAN11
 AC079328.2
 DLEU2L
 GOLM1
 AC009237.3
 ELOVL4
 LRCH2
 RPL23AP7
 BX679664.3
 ADCY6
 CFJ
 STARD13
 PKM
 RUNX1T1
 FKBP10
 SOCS3
 ZNF221
 GPNMB
 ENPEP
 TRIB1
 LAMA2
 HAND2
 SYNE1
 ADAM33
 TM4SF20
 AC009237.14
 SPX
 AC096636.1
 RHBDL2
 RTL5
 RALGPS2
 WFDC2IP
 LINC00327
 SLC44A1
 SESTD1
 SNRNP27
 OXR1
 NR1D2
 AC027237.4
 TBX3
 ACTRT3

Table_S5 genes names positively enriched in MSC

BACE2
AC002480.2
COL5A1
CRIM1
SASH1
HMGB3P9
ST13
GLI3
AC112721.1
ITGA2
UGGT2
AC068580.4
TMEM136
FAM172A
FKBP14
CEBPD
KRCC1
LINC01133
TMEM200B
PGM1
C3orf67
TBC1D2
BMPR1A
RASSF8.AS1
MNS1
HMGA2
LY96
KDELR3
KLF9
APLN
TSPAN5
ZSCAN23
COLEA3
PKIG
SIAE
STARD13.AS
SNX25
AL596325.2
KIF26B
CLSTN2
ELN
CLU
FBLN5
FER1L5
FJX1
VKORC1
SREK1
BASP1
ZNF667
IL20RA
ZRANB2
AL590428.1
RN7SL68P
PCDH9
AC040160.1
ODF2L
HEXB
PRDM16.DT
LIN7A
GPR161
MRAS
CRYZL2P.SEC16B
RUFY1
RF00181.2
NR2F2
ADCY4
LRRK27
PARD3B
ESF1
FLG.AS1
SMTN
CCDC148
AL354740.1
IFI44
LRRK2
OSBPL1A
GLI2
SH3RF1
PLAUR
AKAP12
MYH15
PCDHGB9P
PJA2

Table_S5 genes names positively enriched in MSC

AC005840.1
 SEMA7A
 JUN
 SYDE1
 FOXP2
 FAXDC2
 ACTN1
 LINC02257
 NXPH4
 CALCOCO1
 CARD6
 FAM3C2
 CNTNAP1
 ABCA1
 UBE2H
 CNTN3
 DYNC2L1
 AC005618.1
 B3GALT5.AS1
 PPIC
 PMP22
 PALLD
 SPRED1
 BACH1
 LGALS3
 AC144831.1
 PGRCMC2
 MT2A
 CCDC122
 JDP2
 DUSP10
 AIM1
 ANXA2P2
 TMEM255B
 AC023043.1
 PCDHGB6
 TJP1
 DHX32
 NTSR1
 ZDHHC21
 MFGE8
 TMX3
 CDKN2B.AS1
 NRP2
 AC097534.2
 CLDN12
 CSDE1
 ZFP36L1
 TPST1
 MSC.AS1
 CDKN2C
 JCHAIN
 ZDHHC1
 RPS27L
 SEC62
 AC079600.3
 AC060766.7
 TMEF2
 ARHGAP29
 SLC25A43
 IER3
 ZNF436
 AC087893.1
 ASAP2
 RN7SL15P
 MMP19
 ARNTL2
 AL049775.1
 DNAJC13
 DCAF6
 LINC02593
 CTSF
 CADPS2
 CDH6
 LINC02016
 RPL12P38
 EHD3
 SLFN5
 KIF13A
 TMCO1
 CPT1C
 AOPEP
 AC013564.1

Table_S5 genes names positively enriched in MSC

ABL2
NA..2
EFNA5
KIFAP3
HS3ST3A1
FRMD5
FLRT2
NPC1
SERTAD2
SLK
CTSB
NA..3
ZNF300P1
AC092139.1
PSORS1C1
ADGRA2
IL13RA1
EFCAB1
PLA2G12A
NECTIN3
ENO1P3
CCDC92
IFT80
P3H2
DUSP5
SBDS
MICU3
RN7SKP163
LINC00184
LGR4
ZNF106
HSP90B1
CD27AS1
PCDHGA1
COX7A1
PNMABA
SLC2A13
LIMCH1
HSPA8P4
EDEM3
OS9
ABRAXAS1
S100A16
MYBPB
S100A13
YAP1P1
USO1
OXCT1
PLK2
NLGN1
ADAMTS12
OTUD7B
AC046143.1
AC091965.5
AC005592.1
AC073610.2
TVP23CP1
WDFY2
SPART
CCL26
FNIP1
LINC02333
AP0002369.1
SAR1A
CNRP1
POLR3D
GTF2H2C
TXNDC15
SPRY2
TEAD3
CLCN3
SLC38A2
SLC35G2
AP3S1
CD68
SLC26A4
CDKN1A
AHRR
REV3L
CTBS
ARSK
LINC02518
PHC2

Table_S5 genes names positively enriched in MSC

KCNK1
CNN1
AP4E1
FOXP1
BAG3
SLC41A2
GPC1
ANXA1
TSHZ2
BEX3
ID2.AS1
RTL8B
EV15
PRELID2
AC010198.2
BTBD8
FTH1P23
ENO1P4
RAB3B
SLC25A46
MAGI1
KATNAL2
AC092040.1
MAPK3
TBC1D8B
COPB2
UNC5B.AS1
POLI
ZNF883
AC010201.2
S100A6
RNNUATAC
AC239809.3
RBMS1
AP000915.2
RTN4RL1
INSYN2B
DPYSL2
PCDHG6
STAM2
C2CD2
GPRC5A
RAB36
AP000695.1
RAB10
TEX9
RNF11
TMEM237
AC008758.5
PLEKHA4
PEA15
ZNF281
SMARCA1
NEO1
PXDC1
ATE1.AS1
ARL4AP5
TMEM127
RN7SKP97
ACSL4
ADGRB2
CD59
NEDD8.MDP1
ZNF239
PRNP
IQGAP1
YBX3
C1orf54
CAPS2
PHLDA3
IL11
ZNF626
PGR
NUPR1
DACT1
FTH1P8
EPHA3
APLF
DENND2C
CCDC113
COL13A1
CFAP53
Z97653.2

Table_S5 genes names positively enriched in MSC

AC090204.1
 ITGB3
 PIGF
 FXR1
 PLAG1
 RPGRIP1L
 SDCCAG8
 ETV5
 APBB2
 AMZ2
 CAMLG
 P4HA2
 KCNIP4.1T1
 LINC01275
 TRPC1
 EEF1A1P7
 ETA1
 YIPF5
 LANCL3
 SSX2IP
 RPL5P34
 RPL7P1
 GNG11
 CSDC2
 AP001453.2
 ZNF354C
 AC080038.4
 PTEN
 NBEAL1
 RN7SL718P
 AC073508.2
 FGD1
 MAPK10
 SYT11
 AC016526.1
 TMEM35A
 COL8A2
 IRX2
 IFT57
 FRA10AC1
 CDO1
 IER3 AS1
 SBF2
 KCTD12
 CNN3
 RASD1
 CASC2
 INKA2 AS1
 TNS1
 ATP13A3
 RNF152
 FBXO17
 DPP4
 HTR2B
 MEIS1
 COL4A1
 AC074212.1
 C7orf25
 WASHC2A
 PTPRS
 APLP2
 GULP1
 AL133415.1
 WLS
 ADH5
 LNX1
 SPEF2
 PLA2G4A
 RTL8C
 AC124276.2
 MIR34AHG
 AVP11
 AL133453.1
 PCDHB14
 TMEM65
 F2RL2
 TRPS1
 FOSL2
 ANPEP
 RF00181.3
 CRISPLD1
 ALS90438.1
 MID1

Table_S5 genes names positively enriched in MSC

SLTRK4
CBLN3
WASHC3
HSPA4L
CCDC82
ZNF415P1
SEPT10
RIN2
LINC02377
MYO10
LINC00571
TUB
SPRED3
PLPP4
GNAI1
TRO
SH3BGR
PCDHGC4
RHOD
MIR604
PRAF2
OAT
RRBP1
AC027117.1
P3H3
PAPSS2
GPER1
METTL7A
AC090617.5
SIL1
ANKIB1
C22orf23
OPTN
CTSL
MSL3P1
AC080013.4
TRIP11
ARHGEF25
ZMYND11
MEGF8
ZSCAN30
AC109466.1
AL391117.1
EFCAB7
AC008574.1
ZNF404
PARD3
TRIQK
VMP1
MMP6
NMNAT2
NIPAL2
SOGA3
DLG5
MAB21L1
RN7SL67P
POGLUT2
PLS3
F8
AL356968.2
GOPC
TMEM14A
SRSF11
TMEM98
GNPD42
CCDC103
AC002094.1
RASL11A
SSH1
SPTY2D1
RUBCNL
HTRA3
TMEM183B
BCAR1
ZC3H6
CDK2AP1
BX470102.1
AIG1
EPHA2
SLC35F5
LINC01503
TLN1
RIPOR3

Table_S5 genes names positively enriched in MSC

AGKP1
CHMP1B
SLC7A8
ELK3
SLC4A3
STAG1
NAV2
ZNF334
ECT2L
AFAP1
ARS1
AC019117.3
NTF3
WAC.AS1
NAV2.IT1
GCC2
SCG5
RAB34
CNTNAP3P2
TENM3
ZMPSTE24
DMRTA1
SNORA24
EPN2
ANKRD2
PDK4
SKAP2
CSMD2
BNIP3
ZSWIM6
OSTC
AL109918.1
CPQ
ERLEC1
CDC42BPA
GPR88
RBMS3.AS3
STX7
MALAT1
TEK
GPRASP2
ANGPTL4
CPNE7
WDR35
CABP1
BCL10
WASL
AC135048.1
RO60
SDC4
PRDM5
IFT22
CHIC1
CHPF
KIDINS220
LMCD1
AGAP1
FHL1
ZBTB8A
AC016747.4
IGSF10
GRIA1
NCS1
RUFY3
AC006460.1
AC064799.2
ODF3L1
AC119674.1
AP005230.1
BPGM
MYO1C
PIP4P2
RAI2
CTIF
TP53BP1
COL7A1
AL391422.4
NOTCH2NL
MOCSS2
P2RX6
PIH1D2
CCNG1
LPIN3

Table_S5 genes names positively enriched in MSC

SPATA6
MFN1
AC083805.3
AC090425.1
SRPK3
AL157394.1
AC006213.2
LRRCS8
PCDHB10
AP001528.1
NPHP3
HSDL2
PCDHGA8
GATA6
GOLG44
AC079601.1
KIF3A
KRR1
HOXA3
GOLGAB0
CAMK1G
PNPLA3
SINHCAP3
TTC26
PEAR1
DRAM1
GADD45A
MACF1
IFIT2
RAP1B
HOXA6
SEPT7
Z99129.4
ERFE
UAP1
ZNF365
AP002414.2
TNFAIP8L3
FNDC4
TRIM32
RUSC2
LAMC2
MMP3
PLXNB3
HSD17B6
CAMSAP2
NR3C1
KCTD9
C3orf80
ARHGEF33
GMNC
AL590483.1
CNGA3
LINC01116
VSTM5
CCDC50
TMEM67
AC084117.1
HSPA12A
FER1L4
NBR1
RASSF10
PSAP
ZEB1
IFT88
CFAP58
AC006213.5
HBEGF
MAOB
MCHR1
TMED7
ARHGAP28
ZNF300
AC005736.1
CMBL
BCL6
GAPLINC
SLC38A5
SPTBN1
SCN8A
SLC22A23
KDELR2
MSTN

Table_S5 genes names positively enriched in MSC

APOBEC3C
AC011558.1
RHBDL1
AC135507.1
KLHL9
ZNF415
ARHGAP32
AL662795.2
ANKRD12
ZFYE9
SIM2
CTNNAL1
NUDT16
LINC00857
RBPM5
PAFAH1B2
SPTBN4
GOLGA8Q
GALNT2
AIDA
CD276
NLGN4Y
RHOXF1,AS1
AC013643.3
AC036108.2
TAF13
FAM92A
SYNPO
FBXL2
CCDC112
SPATS2L
ABHD2
DNAH5
MAP6
LINC00842
RWDD2B
DUSP1
CXXC5
TGM2
UFSP2
FAM204A
KRT222
TOX2
GJC1
RA87B
ELN,AS1
BTBD19
RASAL2
STBD1
MUC1
GJD3
CCDC47
YES1
MAN1A1
SNTB1
PUS7L
KIF3B
ZFP36
AC012370.1
OMD
TOR1AIPI1
CACNB3
TMOD2
FRMD4A
UBA5
MORN4
BX842568.2
PCAT6
SACS
PCMTD1
CTDSP2
AC110792.2
RBBP4P1
GLRX
LINC02605
GABARAPL1
NOMO1
LINCO1152
NFE2L1
RPL23AP49
AC119674.2
ALDH3B1
ADAMTS15

EFCAB6
PURA
DMKN
UTP11
ARFGAP3
RAPH1
AC007541.1
NKIRAS1
HDGFL3
ZSWIM8
TCEAL8
ZNF423
BCL2L13
AC002398.2
SLC35E4
MIRLET7BH6
AADAT
PALMD
NR1D1
CCDC90B
SH3D19
LINC00346
AC000403.1
SNORA44
CNTNAP3B
TSPAN6
OXTR
SHOX
AC102945.2
LRRCC1
SNORA40B
PLXNA1
ZSCAN26
NUAK1
RAD50
JHY
RPL34P20
TRPC3
AL356124.1
GBAP1
FTH1P10
HCFC2
WBP4
KAZN
PAPPA2
MBNL2
CXCL2
SLC2A5
RRAS
ACTR10
CADM1
TRAK2
CRYZL2P
SSB
MEIS3P2
AP003119.2
SV2A
TSPAN1
ZNF521
AC090578.1
AK6
AL136164.3
SVIL2P
STXBPS.AS1
EXOC1
FAT2
CLOCK
SNORD113.9
LNPK
RPS12P16
IGBP1
SAV1
AP000317.1
MEIS2
KRT19
MORF4L1P1
GNG12.AS1
APBA1
LIMS4
AL117327.1
TNFRSF10D
CIZ1
APOLD1

Table_S5 genes names positively enriched in MSC

Table_S5 genes names positively enriched in MSC

COPS2
RAB3GAP1
CCDC188
DUBR
SH3BP4
HOXA5
FARP1
IL17RD
MSANTD3.TMEFF1
NUMB
MIR1
STEAP3.AS1
WFDC1
SLAIN2
STAT1
MDFIC
FBLN2
PHTF1
TANK
CSRP3
SMPDL3A
RA86A
ATP7A
TRAF3IP2
FAM160B1
AL606834.1
IFI27L2
KDSR
RGS20
LZTFL1
DDIT4.AS1
SLC9C1
RBBP8
TNFRSF12A
CLUAP1
ITGA8
RN2.63P
PREPL
EXOC5
NDIFP2
HBP1
DSP
VWA5A
GCNT1
AL136116.3
AC068620.3
TRIB2
AC011455.2
MAFF
PRKAA2
DUSP6
EV12A
SEMA6D
HAPLN1
BDNF.AS
AP000941.1
RNF217
ZNF703
TLR4
JRK1
LINC01852
FGF16
AC026150.1
DDX43
TMEM43
HMGB1P6
ACTBP7
AC003101.1
CCDC91
ULBP3
SH3PXD2B
CDK20
GJA1P1
IQCB1.SCHIP1
AL731568.1
SAMD9
KCTD11
AC008740.1
C11orf74
PHF11
AF165147.1
C2orf69
TSPY1L2

Table_S5 genes names positively enriched in MSC

VANGL1
 AL162231.2
 VIM
 AC092376.2
 PERP
 LRRK43
 CCDC191
 TSC2D2
 DLX6_AS1
 CDR2L
 SNAI1
 ZCWPW2
 BMP1
 HOXA9
 ACAD11
 TGFBR2
 RRAGC
 AF230666.1
 KCNQ5
 GRIK1
 LAMP2
 NFKBIZ
 ME1
 CNH1
 DEPP1
 PDGF
 AC090099.1
 EIF4G3
 PFN2
 L3HYPDH
 PPP1R3G
 SLFN12
 AC022336.3
 IL6ST
 GHCG
 ZBTB20
 ACBD3
 HECTD2
 KCTD3
 CD44
 SDC2
 LYSDM3
 RNLS
 AL359091.5
 AL035681.1
 AHNAK
 AL096711.2
 NPPIP1
 DAPK2
 FLRT3
 EPG5
 NUDT12
 EFHC2
 SAMD9L
 TCFB3
 DOCK1
 STAG2
 PNMA8B
 CCL2
 AC010605.1
 AC006249.1
 RF00019
 EMC4
 ANKRD50
 PAPSS1
 ETFDH
 GRB14
 PAQR7
 GOLGA6L5P
 NEK6
 NRIP1
 PFKM
 AC026403.1
 IFT140
 SMARCE1
 NSA2
 AL031587.3
 SYT7
 NOTCH2
 PDLIM7
 STXBP3
 VPS13A
 LINC00922

Table_S5 genes names positively enriched in MSC

TBC1D12
LINGO2
STK17B
ADPRH
MAMDC2
AL451064.2
LINC00475
LINC00840
CTBP2P8
AC009404.1
AL393840.2
CABYR
SOX6
NFE2L2
ARHGEF10
CHST1
KANSL1L
FGF18
CSNK1A1
AC005722.2
PKD1P5
SLC12A8
TMEM87B
BMT2
PFDN1
TRIP10
RBPF4
SPRY1
R3HDMD2
MTUS2
ID1
AL356515.1
PGF
ACVR1
DDX5
PTP4A2
LIG4
AL355987.2
BMP3
CSTA
MPV17
KCN4A
NOL8
ECHDC1
PCDHBS5
LYPD6
PHYHIP
RRAGB
MDGA1
MEG9
LINC00865
SLC9A7P1
ASAP1
LAMB3
FRS2
SDCBP
KYAT3
NMD3
HR
AC012065.2
SLC30A9
AC074143.1
FIG4

Table_S6 summary of qPCR cohort

Type	Number
Adipose MSC	4
BM-MSC	12
Fibroblast of	3
H9 ESC	8
Hepatocytes	3
HUVEC	4
IPS	2
Myoblasts	3
Neural Stem c	2
skin fibroblast	10
Umbilical cor	3

Table_S7 Description of Mlnc Candidates

target_id	Sleuth qval (adj. Pval)	b (FC estimator)	length (nt)	exons	Chr	Interval	Strand	present in long reads	Genome browser remarks	REFLNC/GTEX	discriminative kmers	expressed in MSCs (kmers)	strong expression in SMCs (kmers)	Notable other expressions	FEELnc_coding_potential
Mlinc.54177.1	5.38147421885392e-10	3.64181797640709	17139	1	18	67485128-67502266	-	FALSE	No annotation (close to predicted lncRNA, overexpression in fibro)	FIBRO	FALSE	Ad/Bm	TRUE	myocytes/osteoblasts	0.336
Mlinc.28428.2	3.91322165116094e-24	4.75090994700694	2816	3	12	88380133-88419354	-	TRUE	TUCP/TCONC_00020862	FIBRO ONLY	TRUE	Ad/Bm	FALSE	dermal papilla/muscle satellites	0.18
Mlinc.109628.1	4.18652552062097e-14	3.55558466039083	3615	2	6	169163268-169168514	+	FALSE	TUCP/TCONC_00012024 (in zone)	FIBRO/ARTERY/BioMarker survie	FALSE	Ad/Bm/UC	TRUE	myocytes/chondrocytes/astrocytes/readipocytes	0.446
Mlinc.128022.2	2.24440052433889e-15	3.19030456476542	992	5	9	115556550-115692288	+	TRUE	TUCP/TCONC_00016469 (lot of variants in zone) + numerous chromatin marks	FIBRO ONLY	TRUE	Ad/Bm	TRUE	Myocytes/myoblasts	0.268
Mlinc.32708.1	0.00166963056755997	2.11405992530657	4346	1	13	75859960-75864305	+	FALSE	In close 3' region of LOM7	BRAIN (Isoform ?)	FALSE	Ad/Bm/UC	TRUE	Chondro (very strong)	0.118
Mlinc.10317.1	8.7851692806226e-11	3.69104237801407	8760	1	12	15239087-215247846	+	FALSE	In close 3' region of KCNK2, very high expression in fibro	TRUE	Ad/Bm	FALSE	oste/o/chondro	0.266	
Mlinc.45246.1	5.52855779835755e-11	3.12467428011448	3825	1	16	65129303-65133127	+	FALSE	TUCP/TCONC_00024436 (partial overlappig)	FALSE	Ad/Bm/UC	TRUE	SMC (very strong)/osteoblast (very strong)	0.156	
Mlinc.61271.1	0.000934727743803686	2.11029387027848	4902	1	24	46810683-46815584	+	FALSE	TUCP/TCONC_00003666, partial overlapping	FIBRO/MUSCLE/ADIPO/Blood	FALSE	Ad/Bm/UC	TRUE	Osteoblast (very strong), chondrocytes (very strong)	0.126
Mlinc.6947.1	0.0087759651506394	1.79363960044882	1126	1	11	48790080-148791205	-	FALSE	On repeated region	FALSE	no expression				0.348
Mlinc.89912.1	2.19410912061805e-13	3.160857751305	3159	1	4	80263395-80266553	-	TRUE	In close 5' region of FGF5 (very high in fibro)/Zone of regulation ++	TRUE	Ad/Bm/UC	TRUE	muscle satellite/papilla/myocytes	0.198	
Mlinc.98481.1	4.004588311421693e-14	3.43441525297039	1532	1	5	120687803-120689334	+	FALSE	In close 3' region of PRR16 (very high in fibro)	FALSE	Ad/Bm	TRUE	myoblast/chondro	0.084	
Mlinc.123973.1	1.61744832508345e-09	2.80103674598808	494	1	9	12823414-12823907	+	FALSE	In close 3' region of LURAP1L (high in fibro)	FALSE	Ad/Bm/UC	TRUE	hepatocytes/osteo (very strong)	0.244	
Mlinc.63232.1	1.85061125320116e-11	3.24619016103911	232	1	29	2068044-92068275	-	FALSE	repetition	FALSE	Ad/Bm	TRUE	Myocytes/myoblasts/SMC	0.368	
Mlinc.107651.1	1.43285225448328e-06	2.90795380207381	8953	1	6	130445023-130453975	+	FALSE	In close 3' region of TMEM200A (high in fibro)	OVERLAP / PARTIAL	FALSE	Ad/Bm/UC	TRUE	hepatocytes	0.328
Mlinc.117098.1	9.01415915005679e-06	2.6835474301863	1359	1	8	299352-300710	-	FALSE	lncRNA AC136777.1 in zone (10Kb)	FALSE	Ad/Bm	FALSE	neuron/muscle/endothelial	0.106	
Mlinc.70094.1	5.13172839816861e-10	2.8220665603735	946	1	2	237319841-237320786	-	FALSE	In close 3' region of COL613 (very high in fibro)	TRUE	Ad/Bm	FALSE	muscle satellite/papilla/oste/o/chondro	0.12	
Mlinc.64225.1	0.002065233114836	2.0136339988553	302	1	2	113609969-113610270	-	TRUE	LTR close to RPL23A-P7 (pseudogene + strong repetition (high in fibro))	FALSE	Bm	TRUE		0.368	
Mlinc.54151.1	1.00341251767868e-08	2.89678331657768	5908	1	18	67389037-67394944	-	FALSE	Intron of AC091042.1	FALSE	Ad/Bm	TRUE	osteo/hESC/chondro	0.402	
Mlinc.78582.1	1.60651558091348e-09	3.32612797663159	3402	1	3	30418752-30422153	-	FALSE	HSMM specific Chr marks (+ FOXA1 site)	NA	NA	NA		0.178	
Mlinc.109600.1	1.23389178442618e-08	2.57894271495465	318	1	6	168932197-168932514	-	FALSE	LINE/LTR repetitions? lncRNA in zone	FALSE	Ad/Bm	TRUE	H1-ESC/osteoblast	0.53	
Mlinc.9300.1	1.63418031048994e-08	2.66341118521261	1382	1	11	92488616-19248997	-	FALSE	variant of AL390957.1 (Chr marks ++ and FOXP1 site), lot of predicted lncRNA	FALSE	Ad/Bm/UC	TRUE	SMC (very strong)	0.144	
Mlinc.54554.1	2.57860369698677e-07	2.45953847799982	2156	1	18	75463916-75466071	+	FALSE	HSMM specific Chr marks (+FOXA1 Site)	FALSE	Bm/UC	TRUE	hepatocytes/SMC (very strong)/muscle cell	0.19	
Mlinc.83437.1	3.75186606665131e-11	2.92026510167055	2904	1	3	126213050-126215953	-	FALSE	TUCP/TCONC_00006199, lncRNA, Chr Marks LncRNA /TUCP in zone, Chr Mark, TF binding site	FALSE	Ad/Bm/UC	FALSE	mesenchial/neural prog./myocytes/epithelial	0.292	
Mlinc.127918.1	8.59549299747758e-05	2.2148444399395	2870	2	9	113624677-113628441	-	FALSE	close to an AS of PTGS2 antisense NFKB1 complex +-+; Numerous Chr mark + TF binding sites	TRUE	Ad/Bm	FALSE	Chondro/papilla	0.302	
Mlinc.9231.1	2.0958735035184e-06	2.77869731979218	2415	1	11	186681846-186684260	+	FALSE	Overlap TUCP/TCONC_I_2_00018477 (High in Chr Marks/TF binding sites)	FALSE	Ad/Bm/UC	TRUE	Mononuclear cells (very strong)	0.152	
Mlinc.78418.1	1.82921400272116e-05	2.52537568052055	2945	1	3	27646119-27649063	+	FALSE	Overlap TUCP/TCONC_I_2_00018477 (High in Chr Marks/TF binding sites)	FALSE	Ad/Bm	FALSE	epithelial/myocytes/chondro	0.232	
Mlinc.61267.1	0.00800968729326976	1.85151131179545	8834	1	2	46797931-46806764	-	FALSE	TCONC_00003666, rich zone in lncRNA, Chr mark et TF binding site, close to SOCS5 (high in fibro +-+)	TRUE	Ad/Bm	TRUE	osteo/myocytes/papilla	0.162	
Mlinc.11761.1	0.000429609840134646	2.16878061892736	1232	1	7	38054085-38053136	+	FALSE	EGFR1 Binding site/ HSMM Xhr sites	FALSE	Bm	TRUE	Myocytes/myotubes	0.084	
Mlinc.54156.1	2.76421814657461e-05	2.55034529045675	6240	1	18	67395327-67401566	-	FALSE	HSMM Chr Binding Site + FOXA1 Sites	TRUE	Ad/Bm	TRUE	smc of trachea (very strong)	0.334	
Mlinc.47523.1	6.05265557244177e-08	2.53923057783646	1558	1	17	13490975-13492532	-	FALSE	In 3' of gene HS3ST3A1 (High in fibro)	FIBRO/ DIFFUSE	FALSE	Ad	TRUE	epithelial/mesangial	0.342
Mlinc.18093.1	3.3005434599755e-09	2.70413881781352	977	1	11	12085459-12086435	-	FALSE	Close to AC124276.2 variant Chr Mark ++ and TF binding site +-+	FALSE	Ad/UC	TRUE	epithelial/keratinocytes	0.236	
Mlinc.57835.1	1.15589458808076e-08	2.42525523068851	927	1	19	44737723-44738649	+	FALSE	TUCP/TCONC_00027580	FALSE	FALSE	FALSE	Mononuclear cells (very strong)	0.436	
Mlinc.104634.1	7.05793182343196e-12	3.12285461836103	2731	1	6	57260477-57263207	-	FALSE	TUCP/TCONC_00011222 + Chr Mark (HSMM) and TF binding site	FALSE	Ad/Bm/UC	TRUE	SMC (strong)	0.314	
Mlinc.80687.1	3.47529931298406e-06	2.15786698401882	537	1	3	65018718-65019254	+	FALSE	In Intron of gene ADAM-AS	FALSE	BM	TRUE	epithelia/chondro/neural prog.	0.464	
Mlinc.96690.1	2.09413387964091e-06	2.08136277131835	637	2	5	84491916-84493310	+	FALSE		FALSE	BM/UC	FALSE	myocytes/oste/o/chondro	0.332	

Table_S7 Description of Miinc Candidates

FEELnc label (1=miRNA)	Closest_gene	Distance	TarpmiR max sites by miRNA	Primer1 (F)	Primer2 (R)
0AC110597.1		162	6	GAAGGGCC TGAAAGCCAT	CATGATTT GCCCTCGGC
0Y_RNA		-11088	4	TGT	TT
0LINC01615		-344	4		
0U2		42104	4	GTTTCTGG GCTGCTTC	AGGGGACA CTTCGAC
0LMO7		-90	5		CTCAA
0KCNK2		-1994	5		
0CDH11		-3191	4		
0LINC01118		1084	5		
0AC245389.1		-4592	4		
0FGF5		-46	5	TCGGCGTTG GAAACCA	TACAGACG GAGAGCTTC
0PRR16		-471	4	AAC	CCA
0LURAP1L		-1283	4		
0IGKV1OR2-2		33044	3		
0TMEM200A		-1960	5		
0AC136777.1		-10423	4		
0COL6A3		-3217	4		
0RPL23AP7		-232	4		
0AC091042.1		13732	5		
0LINC01985		-102592	4		
1AL109924.1		30096	4		
0AL136987.1		-27642	4		
0AC116003.2		-26959	5		
0ALDH1L1-AS2		-2881	5		
0AL162727.1		-2043	4		
0PACERR		-400	4		
0AC098614.4		-7343	5		
0LINC01118		9904	5		
0SFRP4		-28390	4		
0AC091042.1		20022	5		
0HS3ST3A1		-3157	4		
0AC124276.2		-456	4		
0BCL3		9056	5		
0RAB23		38163	4		
1ADAMTS9-AS2		-7250	4		
0AC113383.1		-1151	4		

Table_S8 statistically relevant interaction between Mlinc and proteins

Mlinc.28428.2	Mlinc.128022.2	Mlinc.89912.1			
Uniprot ID	gene ID	Uniprot ID	gene ID	Uniprot ID	gene ID
Q496A3	SPATS1	P31994	FCGR2B	P02760	AMBP
Q9UBM1	PEMT	Q9Y2T5	GPR52	Q92830	KAT2A
Q92911	SLC5A5	P14618	PKM	Q6NUU1	CHADL
P09001	MRLP3	Q9P1W8	SIRPG	Q9Y2K9	STXBPSL
Q96C25	MED8	Q9P0J1	PDP1	Q9NV29	TMEM100
P20700	LMBN1	Q8N9I0	SYT2	Q8N6M3	FITM2
Q5FYB0	ARSJ	Q86X59	C17orf82	O43915	VEGFD
O15066	KIF3B	Q8NC8	TMEM116	Q9NRD1	FBXO6
O77932	DXO	Q9P2T1	GMPR2	Q9UJH8	METRN
Q9P232	CNTN3	Q75581	LRP6	Q9BW8	APOL6
Q9NRY2	INIP	Q7L591	DOK3	Q15746	MYLK
Q12931	TRAP1	Q6F199	OR10K2	Q9BQE3	TUBA1C
Q0D2K2	KLHL30	Q9UM11	FZR1	Q96HE7	ERO1A
Q16891	IMMT	Q8WUH6	TMEM263	Q8TCG2	PI4K2B
Q9Y6M4	CSNK1G3	Q9UKN8	GTF3C4	Q99969	RARRES2
Q16706	MAN2A1	Q96A47	ISL2	O43148	RNMT
Q99946	PPRT1	Q9ULJ6	ZMI21	Q7Z6V5	ADAT2
Q96J6	JPH4	Q99829	CPNE1	Q9UIA9	XPO7
Q81ZH2	XRN1	Q29983	MICA	P08684	CYP3A4
Q8NBF1	GLIS1	Q92783	STAM	Q8N135	LG41
Q9BXF6	RAB11FIP5	Q8IYX7	SAXO1	Q96A04	TSACC
Q8TDB4	MGARP	P0DN77	OPN1MW2	Q99217	AMELX
Q9H251	CDH23	P0DN78	OPN1MW3	Q8NAE3	LINC01555
Q96RQ3	MCCC1	Q62MH5	SLC39A5	O60403	OR10H2
Q9HBR0	SLC38A10	Q5TH69	ARFGEF3	P49902	SELENOP
Q9BTU6	P14K2A	P04062	GBA	Q9Y4C5	CHST2
Q9Y2P8	RCL1	Q8IVM8	SLC22A9	Q14142	TRIM14
Q15056	EIF4H	Q96T88	UHRF1	P30085	CMPK1
Q9UMZ2	SYNRG	Q55WL8	PRAMEF19	G3V0H7	SLCO1B7
Q9GZP1	NRSN2	Q8N427	NME8	P60880	SNAP25
P0C645	OR4E1	Q8NBP7	PCSK9	P01767	IGHV3-53
Q6ZV6	KIAA1549L	Q8TCX1	DYNC2L1	Q8WUJ1	CYBD52
O75909	CCNK	Q13072	BAGE	Q6ZQN7	SLCO4C1
Q99574	SERPINI1	Q86Y27	BAGE5	P83731	RPL24
Q9Y5G7	PCDHGA6	Q96KE9	BTBD6	Q5T1H1	EYS
Q1A5X6	IQCJ	Q6UXH8	CCBE1	P05496	ATP5MC1
Q8IW5	MFSD6L	Q14409	GK3P	P78312	FAM193A
Q99607	ELF4	P56177	DLX1	P27918	CFP
Q8N6T3	ARFGAP1	Q9H190	SDCBP2	Q15834	CCDC85B
O14950	MYL12B	P23759	PAX7	Q11201	ST3GAL1
Q6ZN28	MACC1	O15409	FOXP2	Q9Y6H1	CHCHD2
O95402	MED26	P60520	GABARAPL2	P68363	TUBA1B
A8NIV6	LRRIQ4	P55809	OXCT1	Q494X3	ZNF404
A8MUM7	LGALS16	Q9NVH2	INTS7	Q96LJ7	DHRS1
P59536	TAS2R41	Q8N0Z8	PUSL1	Q86X58	RNF130
O95900	TRUB2	Q7L457	ARMCX6	P48742	LHX1
Q99593	TBX5	Q9Y5J6	TIMM10B	P15882	RPS2
P21439	ABCBC4	Q8WUJ4	HDAC7	P0DMV0	CT45A7
P20073	ANXA7	P04259	KRT6B	Q8N972	ZNF709
Q8TCC3	MRLP30	Q8IY88	BOD1L2	Q8IUY3	GRAMD2A
Q6Q788	APOA5	Q95455	TGDS	Q9NR96	TLR9
Q13443	ADAM9	Q969P6	TOP1MT	Q8NB0	POC1A
Q7L311	ARMCX2	Q3MIP1	ITPR1PL2	O14827	RASCRF2
Q969Q6	PPP2R3C	Q86YV6	MYLK4	Q17R55	FAM187B
Q8N1B4	VPS52	P04001	OPN1MW	Q9H0M5	ZNF700
P08243	ASNS	Q8NHF49	OR4X1	O43715	TRIAPI
Q05193	DNM3	P55771	PAX9	Q9P0V3	SH3BP4
Q13243	SRSF5	Q16656	NRF1	Q04771	ACVR1
Q8NDV7	TNRC6A	Q5JTZ5	C9orf152	Q6ZRS4	ITPRID1
P15941	MUC1	Q95382	MAP3K6	O60337	MARCH6
Q96A5N	TMEM143	Q2QD12	RPEL1	Q13103	SPP2
Q98XT2	CACNG6	Q92796	DLG3	Q75663	TIPLR
Q96EY8	MMAB	P05089	ARG1	Q8NBM4	UBAC2
P11498	PC	Q9Y512	SAMM50	Q8NEG0	FAM71C
Q96GC9	VMP1	P20265	POU3F2	O60656	UGT1A9
Q9HC52	CBX8	P04216	THY1	Q7L5L3	GDPD3
Q6UVK1	CSPG4	Q96R7Y	IFT140	Q9H9V4	RNF122
P08397	HMB5	O43187	IRAK2	Q7IU36	TUBA1A
P54922	ADPRH	P21397	MAOA	Q9UJL9	ZFP69B
Q96SA4	SERINC2	Q9H2S9	IKZF4	Q6NUJ1	PSAPL1
Q7L945	ZNF627	Q8NI99	ANGPTL6	Q6UXH9	PAMR1
Q8NGB6	OR4M2			P08833	IGFBP1
Q9GZX3	CHST6			Q75431	MTX2
P80748	IGLV3-21			Q6UDR6	SPINT4
Q6ZTW0	TPGS1			Q5HYN5	CT45A1
Q8TB05	UBALD1			Q5T0J3	C1orf220
Q75947	ATP5PD			Q96T68	SETDB2
Q75366	AVIL			Q9UHC3	ASIC3
P32754	HPD			Q96S25	ADO
Q92600	CNOT9			Q8NHU0	CT45A3
Q8TAS1	UHMK1			P29320	EPHA3

Table_S8 statistically relevant interaction between Mlinc and proteins

P22794	EVI2A	Q9H246	C1orf21
A8MTQ0	NOTO	Q9Y5K1	SPO11
Q9Y6U3	SCIN	Q5JVG8	ZNF506
Q96IZ0	PAWR	Q8NHX4	SPATA3
P06733	ENO1	Q9UM44	HHLA2
P10114	RAP2A	Q9BXM7	PINK1
Q5T4B2	CERCAM	Q99952	PTPN18
Q9HA82	CERS4	A0AV14	TMEM129
Q8IVT2	MISP	Q9P2R7	SCLC2A2
Q95427	PIGN	Q9BQB4	SOST
Q9NW87	IFT57	Q86X40	LRRK28
P0C024	NUDT7	Q8N7S2	DNAJC5G
Q9Y241	HIGD1A	Q96PQ6	ZNF317
P55291	CDH15	Q9P2U8	SLC17A6
Q315F7	ACOT6	Q4GOU5	CFAP221
Q9H467	CUEDC2	Q9HB1	EHTM1
Q86XN7	PROSER1	Q0P670	SPEM2
Q07075	ENPEP	Q94B06	PRKD3
Q5JPH6	EARS2	Q75695	RP2
Q6ZL45	CLEC20A	Q3MJ13	WDR72
Q9Y263	PLAA	A8MXK1	VSTM5
Q75309	CDH16	Q96NJ6	ZFP3
Q8N608	DPP10	Q639G0	STYK1
P36871	PGM1	Q9Y5L0	TNP03
Q8WXF5	CRYGN	Q5TAT6	COL13A1
Q8IYX0	ZNF679	P5964	FXYD4
Q8ND0	MAPK1IP11	Q6UXZ4	UNC5D
Q68DC2	ANKS6	P09565	GIG44
Q9NRA2	SLC17A5	P0CG47	UBB
Q5VZE5	NAAS35	Q14CZ8	HEPACAM
P13521	SCG2	Q86US8	SMG6
Q9BU23	LMF2	Q9Y6I4	USP3
P48382	RFX6	Q96FC7	LINC00526
P35222	CTNNB1	Q07654	TFF3
Q7ZTM9	GALNT5	Q96RD2	OR5R2B2
Q8WXD9	CASKIN1	Q8WWQ8	STAB2
P04843	RPN1	P78545	ELF3
Q9BYD1	MRPL13	Q96M02	C10orf90
Q9V573	IPP	Q9NRY7	PLSCR2
Q00748	CES2	Q12767	TMEM94
Q9UK45	LSM7	Q969Z0	TBRG4
Q8NFJ9	BBS1	P35410	MAS1L
O43525	KCNQ3	P50052	AGTR2
Q96783	SLC9A7	Q6VV07	PACS1
Q95104	SCAF4	Q8NH04	OR2T27
Q9HB1	PARVB	Q70EL2	USP45
Q9P2V4	LRIT1	P62244	RPS15A
Q9HCN3	TMEM8A	Q9H2A7	CXCL16
P11464	PSG1	Q9BT92	TCHP
Q9UNN4	GTF2A1L	P16455	MGMT
Q75935	DCTN3	P0DMU7	CT45A6
Q6V0L0	CYP26C1	P0DMU8	CT45A5
Q9H488	POFUT1	P81277	PRlh
Q9NRM1	ENAM	Q14192	FHL2
Q9NYP3	DONSON	Q15486	GUSBP1
O60636	TSPAN2	Q8IYB5	SMAP1
Q86Y28	BAGE4	Q12791	KCNMA1
Q6P4R8	NFRKB	Q8NGT1	OR2K2
Q95319	CELF2	O00755	WNT7A
Q9Y585	OR1A2	P23280	C46
P20339	RAB5A	Q7RT56	OTOP2
Q9UB14	STOML1	P10720	PF4V1
A4D1S5	RAB19	Q9H553	ALG2
P51523	ZNF84	Q86YD1	PTOV1
Q96199	SULCG2	Q27JB1	INF2
Q9BR39	JPH2	Q86Y38	XYLT1
Q14032	BAAT	Q95363	FARS2
Q9COK1	SLC39A8	A8MV57	MPTX1
Q9UQ16	DNM3	Q9H6P5	TASP1
P78344	EIF4G2	Q15084	PDI46
Q1ZZU3	SWI5	Q8N690	DEFB119
P55157	MTTP	Q96JN8	NEURL4
Q5TBB1	RNASEH2B	Q9H3R0	KDM4C
Q75175	CNOT3	Q96E17	RAB3C
Q14558	PRPSAP1		
P08183	ABC1		
P13498	CYBA		
Q9BRB3	PIGO		
Q96C34	RUND1		
Q9P0K8	FOXJ2		
Q15034	HERC3		
Q9UI30	TRMT112		
Q9NUQ7	UFSP2		

Table_S8 statistically relevant interaction between Mlinc and proteins

Q9HHA1	ZNF556
Q8NHS0	DNAJB8
Q9BWM5	ZNF416
Q32P51	HNRNPA1L2
Q8NFT6	DBF4B
Q8N1N5	CRIPAK
Q9Y5F6	PCDHGC5
P50454	SERPINH1
P08247	SYP
O14979	HNRNPDL
Q8NB16	XXYL1
Q13885	TUBB2A
Q9NQS3	NECTIN3
Q8NGU4	OR211P
Q9NX63	CHCHD3
O60907	TBL1X
P08697	SERPINF2
P52738	ZNF140
Q14406	CSHL1
P62333	PSMC6
Q9BTV5	FSD1
Q6U7Q0	ZNF322
Q56D32	CTDSP12
Q9H857	NT5DC2
P19105	MYL12A
A6NML5	TMEM212
Q9NSI2	FAM207A
Q6IC88	GRAMD4
Q96W7	SEC22A
Q96HL8	SH3YL1
Q8IUR5	TMT1
Q8NSR6	CCDC31
P06732	CKM
P55089	UCN
Q8N2G8	GHDC
Q96H40	ZNF486
Q6P3S6	FBXO42
P56282	POLE2
Q8TE12	LMX1A
Q9NPR9	GPR108
P54709	ATP1B3
Q95498	VNN2
Q969V3	NCLN
Q15398	DLGAP5
P43686	PSMC4
Q14118	DAG1
O43414	ERI3
P40225	THPO
A0AV96	RBM47
Q9HCM3	KIAA1549
Q5SR56	MFSD14B
Q96MG7	NSMCE3
Q96M29	TEKT5
C9JDP6	CLDN25
Q99674	CREF1
Q5VZ18	SHF
Q8IYU8	MICU2
Q9UNZ2	NSFL1C
Q9BX68	HINT2
P63124	HERV-K104
Q8NGH7	OR52L1
Q9UMD9	COL17A1
Q02548	PAX5
A2RU54	HMX2
O43570	CA12
O60928	KCNJ13
Q9BRQ8	AIFM2
Q96M93	ADAD1
Q6P198	INO80C
Q9Y3B7	MRPL11
Q13310	PABPC4
A6NKC9	SH2D7
P20248	CCNA2
O95817	BAG3
O75787	ATP6AP2
A6NI15	MSGN1
Q15653	NFKB1B
Q6QNY0	BLOC1S3
Q6ZMV8	ZNF730
Q9NP70	AMBN
Q86UY5	FAM83A
Q5DX21	IGSF11
Q9NYS0	NKIRAS1

Table_S8 statistically relevant interaction between Mlinc and proteins

Q8NG35	DEFB105A
Q8NG35	DEFB105B
Q9Y3C7	MED31
Q14126	DSG2
P10745	RBP3
Q15125	EFP
Q969G6	RFK
Q96115	SCLY
Q2VPB7	AP5B1
Q8IXB3	TRARG1
Q6VVX0	CYP2R1
Q96M83	CCDC7
Q9EQD9	FYTTD1
Q9NRK6	ABCBL0
Q7L1I2	SV2B
Q86YD3	TMEM25
Q16557	PSG3
P27105	STOM
Q8IZD4	DCP1B
Q8NE01	CNNM3
Q30KQ4	DEFB116
Q95873	C6orf47
Q8NAN2	MIGA1
Q9H1Q7	PCED1A
A6NH0	OTOL1
P59091	LINC00315
Q9H9V9	JMJ4
Q9HBL0	TNS1
P51397	DAP
Q07157	TJP1
Q8IY34	SLC15A3
P30989	NTSR1
Q8NGP4	ORBM3
Q96R84	ORIF2P
Q5T2E6	ARMH3
Q86TN4	TRPT1
Mlinc.128022.2	
Uniprot ID	gene ID
P31994	FCGR2B
Q9Y2T5	GPR52
P14618	PKM
Q9P1W8	SIRPG
Q9P0J1	PDP1
Q8N9I0	SYT2
Q86X59	C17orf82
Q8NC18	TMEM116
Q9P2T1	GMPR2
Q75581	LRP6
Q7L591	DOK3
Q6IF99	OR10K2
Q9UM11	FZR1
Q8WUH6	TMEM263
Q9UKN8	GTF3C4
Q96A47	ISL2
Q9ULJ6	ZMIZ1
Q98829	CPNE1
Q29983	MICA
Q92783	STAM
Q8IYX7	SAXO1
P0DN77	OPN1MW2
P0DN78	OPN1MW3
Q6ZMH5	SLC39A5
Q5TH69	ARFGEF3
P04062	GBA
Q8IVM8	SLC22A9
Q96T88	UHRF1
Q5SWL8	PRAMEF19
Q8N427	NME8
Q8NBP7	PCSK9
Q8TCX1	DYNC2LI1
Q13072	BAGE
Q86Y27	BAGE5
Q96KE9	BTBD6
Q6UXH8	CCBE1
Q14409	GK3P
P56177	DLX1
Q9H190	SDCBP2
P23759	PAX7
O15409	FOXP2
P60520	GABARAPL2
P55809	OXT1
Q9NVH2	INTS7
Q8NOZ8	PUSL1

Table_S8 statistically relevant interaction between Mlinc and proteins

Q7L4S7	ARMCX6
Q9Y5J6	TIMM10B
Q8WUI4	HDAC7
P04259	KRT6B
Q8IYS8	BOD1L2
Q95455	TGDS
Q969P6	TOP1MT
Q3MIP1	ITPR1PL2
Q86V6	MYLK4
P04001	OPN1MW
Q8NH49	OR4X1
P55771	PAX9
Q16656	NRF1
Q5J7Z5	C9orf152
Q95382	MAP3K6
Q2QD12	RPEL1
Q92796	DLG3
P05089	ARG1
Q9Y512	SAMM50
P20265	POU3F2
P04216	THY1
Q96RY7	IFT140
O43187	IRAK2
P21397	MAOA
Q9H2S9	IKZF4
Q8NI99	ANGPTL6