SISTEMA: A large and standardized collection of transcriptome data sets for human pluripotent stem cell research
Margot Jarrige, Hélène Polvèche, Alexandre Carteron, Stéphane Janczarski, Marc Peschanski, Didier Auboeuf, Cécile Martinat

To cite this version:
Margot Jarrige, Hélène Polvèche, Alexandre Carteron, Stéphane Janczarski, Marc Peschanski, et al.. SISTEMA: A large and standardized collection of transcriptome data sets for human pluripotent stem cell research. iScience, 2021, 24 (7), pp.102767. 10.1016/j.isci.2021.102767 . inserm-04385458

HAL Id: inserm-04385458
https://inserm.hal.science/inserm-04385458
Submitted on 22 Jul 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License
SISTEMA: a large and standardized collection of transcriptome datasets for human pluripotent stem cell research

Margot Jarrige¹, Hélène Polvèche¹, Alexandre Carteron¹, Stéphane Janczarski³, Marc Peschanski¹, Didier Auboeuf³, Cécile Martinat²*

¹CECS/AFM, I-STEM, 91100 Corbeil-Essonnes, France
²INSERM/ UEVE UMR 861, Paris Saclay Univ I-STEM 91100 Corbeil-Essonnes, France
³LBMC, Univ Lyon, ENS de Lyon, Univ Claude Bernard, CNRS UMR 5239, INSERM U1210, 46 Allée d'Italie Site Jacques Monod, F-69007, Lyon, France

# Contributed equally

* Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Cécile Martinat (cmartinat@istem.fr).
SUMMARY

Human pluripotent stem cells have ushered in an exciting new era for disease modelling, drug discovery and cell therapy development. Continued progress toward realizing the potential of human pluripotent stem cells will be facilitated by robust datasets and complementary resources that are easily accessed and interrogated by the stem cell community.

In this context, we present SISTEMA, a quality-controlled curated gene expression database, built on a valuable catalogue of human pluripotent stem cell lines and their derivatives for which transcriptomic analyses have been generated using a single experimental pipeline. SISTEMA functions as a one-step resource that will assist the stem cell community to easily evaluate the expression level for genes of interest, while comparing them across different hPSC lines, cell types, pathological conditions or after pharmacological treatments.
INTRODUCTION

In the last decades, the isolation of human pluripotent stem cells, first from human embryos (Thomson et al., 1998), and more recently from the genetic conversion of somatic cells (Takahashi et al., 2007), represents a major breakthrough to study lineage-specific mechanisms underlying development or pathological condition and to broaden our capacity for biological therapeutics. Based on their unique capacities for self-renewal and differentiation into specialized cell types, human pluripotent stem cells have become the subject of intense research not only for developmental studies but also in the field of cell therapy, modeling of human diseases and drug discovery and testing (Bassez et al., 2018; Ben M’Barek et al., 2020; Ben M’Barek et al., 2017; Rowe & Daley, 2019).

Genome-wide expression profiling has been a common endpoint for these studies for analyzing pluripotency, differentiation and the function of genetic variants associated with monogenic diseases. A growing body of transcriptomic datasets for various hPSC lines and their specific differentiated counterparts has been deposited in public data repositories (Assou et al., 2007; Au & Sebastiano, 2014; Ghosh & Som, 2020; Godoy et al., 2018; Mair et al., 2019; Mallon et al., 2013; Streeter et al., 2017). Despite this wealth of data, exploring gene expression profiles from different experiments is not trivial. Different challenges still remain such as the integration and interpretation of massive amount of available data, the heterogeneity in original data processing as well as the quality datasets (Pinto et al., 2018).

To overcome these challenges, we provide here a large well-characterized collection of human pluripotent stem cell lines and their derivatives for which transcriptomic analysis has been generated using a single experimental pipeline. Moreover, we have developed a user-friendly search engine to facilitate comparisons of gene expression data between these different biological samples. This open resource presented in SISTEMA web portal is designed for the community to examine individual or multiple genes for expression under
different states from pluripotency to differentiated conditions and under normal or pathological conditions, with 8 different monogenic diseases already represented in this growing database.
RESULTS

Data collection and dataset curation

All the biological samples were collected from the same laboratory and included 45 hPSC samples from 11 hPSC lines (2 hESCs and 9 hiPSC) at either undifferentiated state or one of eight differentiated states representing the different germ layers (Guenou et al., 2009; Maury et al., 2015; Mournetas et al., 2019; Nicoleau et al., 2013) (Figure 1A). This collection included 23 healthy and 24 clones from donors with a rare genetic disease which represented a total of 8 different pathological conditions comprising neuromuscular, neurodegenerative and skin disorders, as well as some isogenic lines generated either using CRISPR/Cas 9 technology or lentiviral transduction (Figure 1B and 1C). Samples treated with different chemical compounds have also been included (Figure 1D). These small molecules were evaluated in the context of different pathologies such as lithium for the autism spectrum disorder caused by SHANK3 haplo-insufficiency, MG132 for Hutchinson-Gilford progeria syndrome, Thiostrepton and Bortezomib for Limb-girdle muscular dystrophy type 2D, VPA for Wolfram syndrome or Kenpaullone for Spinal muscular atrophy (Darville et al., 2016; Harhouri et al., 2017; Hoch et al., 2019). All the transcriptomic data were generated in comparison to DMSO treated controls. Of interest, some of the data provided by the database have not been published yet all the scientific community to have access to a larger spectrum of data. All the cell lines used for the creation of this database as well as the metadata files are also available upon request and following the French legal policy.

The processing of all these biological samples were performed using a unique platform following a standardized protocol from RNA extraction to biological analysis. To overcome the current limitations associated with traditional whole-transcriptome sequencing technology, e.g., the requirements of a significant amount of input RNA, we chose to process the samples by using the Ion Ampliseq technology (Li et al., 2015). This technology has been designed for targeted amplification of over 20,000 distinct human RNA targets simultaneously and is
described to present the advantage of amplifying a short amplicon (~ 150 bp) for each targeted gene allowing a much smaller amount of raw reads (only 10 Million reads sequenced by sample) required for accurate gene expression quantification (Li et al., 2015). To evaluate the reproducibility of the Ampliseq technology, one RNA sample was sequenced twice independently leading to a percentage of correlation of 97.8, which indicated a consistent performance between technical replicates (Figure S1A). This experiment also confirms the robustness of this technology allowing us to study and compare the gene expression of several samples sequenced at different times on the platform. Then we validated the sequencing coverage offered by the Ampliseq technology. For this purpose, each amplicon sequence was BLASTed against a transcriptome of reference (GRCh38.92). This analysis revealed that 99.89% of the amplicons panel were correctly identified and represented 21,080 genes. (Figure S1B). Based on this technology, a total of 443 transcriptomes were generated by following a unique and standardized protocol which led to the consistent deep sequencing coverage and percentage of alignment (Figure S2A and S2B). The sequencing process was followed by a standardized data analyzing pipeline as described in Supplementary Figure 2A. For each sample, reads were aligned to the Amplicons sequence reference using bowtie2 local (v2.3.4.1). Reads below a mapping score of 10 or multi-mapped were filtered using samtools (v0.1.19). Samples were normalized by RPM (reads per Million) and the expression values were converted into logarithm scale (log10(x+1)) using R (v3.6.0).

**Web interface and graphical visualization**

All expression profiles from the 443 experimental datasets have been implemented in the form of a web-resource called SISTEMA ([http://sistema.ens-lyon.fr](http://sistema.ens-lyon.fr)). It was designed as a user-friendly interface in which a set of tools has been integrated for providing a rapid overview of the data and for querying, analyzing and displaying data online. With the purpose of continuously updating this resource with new datasets, SISTEMA enables to compare the
expression levels of particular genes of interest or a combination of genes across different lineages and conditions.

The query can be specified by first selecting the datasets pertaining to the cell lineage(s) of interest, the genotype(s), the cell line(s), the time of differentiation, the effect of a chemical treatment (Figure 2A). Then, the users can define the gene(s) of interest (Figure 2A). As an example of querying for a single gene, the expression profile of \textit{NANOG}, one of the key pluripotent markers (Chambers et al., 2007; Silva et al., 2009), has been analyzed in different samples of undifferentiated hESC samples or hiPSC-derived Keratinocytes. Once the queried data is processed, detailed information regarding the chromosomal localization, the gene structure, its interacting partners are first displayed as SISTEMA interface is linked to other databases such as Genecards, NCBI, MIM, HGNC, Ensembl, Vega and STRINGDB (Figure 2B). For better visualization, the results are presented as a histogram and a downloadable table in which data are expressed as the normalized counts in logarithmic scale observed in each sample. As expected, \textit{NANOG} was found to be highly expressed in pluripotent cells whereas its expression is absent in the hiPSC-derived keratinocytes (Figure 2C-2D). For each sample, a list of different characteristics such as the cell type, genotype, the phenotype, the time and condition in culture, is provided (Table1) and the user can customize the graph and the table in function of the cell types, genotypes, cell lines or time of differentiation (Figure 2D).

SISTEMA also allows a multi-gene analysis. An option is proposed to visualize the expression profile of up to 10 genes on a selected set of samples. As an example, the expression profiles of \textit{NANOG, OCT4} and \textit{SOX2}, the triad of transcription factors known to be critical for self-renewal and pluripotency (Chambers & Tomlinson, 2009), has been analyzed in hESC and hiPSC-derived keratinocytes (Figure 3A). As for the single gene analysis, detailed information regarding each gene is provided and their expression profile is
presented as a histogram and a downloadable table (Figure 3B-3D). In addition, results are also presented as a heatmap in which the retrieved expression data for the selected genes and cell types are indicated and the samples are hierarchically clustered (Figure 3C).

**Example applications**

To illustrate the utility and the potential applications of SISTEMA, we first sought to evaluate the assignments of the samples from clustering the entire database in function of the expression of selected markers. For that purpose, the expression of a set of 10 genes of pluripotency, extracted from the database KEGG pathway, were analyzed across all the biological samples. Hierarchical clustering allowed the correct distinction between pluripotent stem cells and differentiated cells (Figure 4A). Conversely, the expression profile of a set of 10 genes associated with the GO term “spinal cord motoneuron differentiation” led to the distinction between pluripotent stem cells and their motoneuronal derivatives, confirming the relevance of the SISTEMA Database (Figure 4B).

We also evaluated the possibility to define the temporal expression profile of 10 genes known to be associated with the development of the melanocyte lineage (from GO) at different time points of the differentiation protocol from human pluripotent stem cells. As expected, the expression profile of this set of markers allowed to accurately make the distinction between hPSCs and hPSC-derived melanocytes. In addition, it revealed profiles of gene expression that evolved over time along with the differentiation (Figure 4C). Thus, *CITED1* and *ENDRB*, previously described as two important regulators of melanogenesis (Howlin et al., 2015; Nair et al., 2001; Park et al., 2015), appeared progressively expressed during the differentiation process (Figure 4C). The expression of *GLI3* was in contrast more pronounced in the early specification of the melanocyte lineage, as previously described in a mouse model (Matera et al., 2008). As a control, we also chose *GNAQ* which appears to be the only one with no modification of its expression with the differentiation. Concordant with our observations, this
gene has been previously shown to be involved in the regulation of pigmentation rather than differentiation (Kusters-Vandevelde et al., 2010; Van Raamsdonk et al., 2004). Altogether, these analyses validated the relevance of the SISTEMA database both to evaluate the expression profiles of different gene sets across a large battery of biological samples and to visualize the expression profiles of gene sets along lineage specification.

Since their first isolation, human pluripotent stem cells have sparked widespread enthusiasm for the development of new models of human disease and have enhanced platforms for drug discovery. To evaluate the potential of SISTEMA database for these applications, we focused on the expression profile of the survival motor neuron 1 (SMN1) gene, the homozygous deletions/mutations of which causes spinal muscular atrophy (SMA), a severe genetic neuromuscular disease (Lefebvre et al., 1995). As expected, SMN1 expression was significantly decreased only in the biological samples derived from SMA patients (Figure 5A). Next, we sought to explore SISTEMA database to identify a potential small molecule capable to increase the level of SMN1, which therefore would present a therapeutic interest. Thanks to an unbiased analysis including all the pharmacologically-treated biological samples present in SISTEMA database, we identified three proteasome inhibitors, Bortezomib, Thiostrepton and MG132 that lead to an increased expression of SMN1 compared to mock-treated controls. The results were concordant with previous studies that showed at least for two of these proteasome inhibitors, Bortezomib and MG132, a beneficial effect on SMN1 expression both in vitro and in vivo (Kwon et al., 2011; Locatelli et al., 2015). In contrast, and in agreement with previous studies (Brichta et al., 2003; Dachs et al., 2013; Elshafay et al., 2019; Piepers et al., 2011), neither VPA nor lithium treatment showed a significant effect (Figure 5B).
DISCUSSION

Here we present the data access and presentation services that SISTEMA database provides, which enable the community to both discover and to reuse the cell lines and data included in this project.

SISTEMA further complements and outspreads the two gene expression databases dedicated to pluripotent stem cells, StemCellDB for human embryonic stem cells (Mallon et al., 2013) and HipSci for human induced pluripotent stem cells (Streeter et al., 2017). Whereas these online resources are based on either microarrays or whole transcriptome RNA sequencing technologies, the main strength of SISTEMA lies on the use of a unique pipeline to generate transcriptomic analysis providing a collection of quality-controlled curated data. Most of the other data resources from public repository, i.e. EBI expression atlas or NCBI short read archive, are also useful tools to explore gene expression dataset but cover a large set of biological field from microbiology to stem cells biology which makes difficult to easily and rapidly analyze a specific research. One of the objectives of Sistema is to provide a rapid overview of the gene expressing by providing a comprehensive and limited number of hESC and hIPSCs cells and their derivatives. Additionally, SISTEMA relies on the use of a recent technology, named AmpliSeq, designed for targeted amplification of over 20,000 distinct human RNA targets simultaneously using a highly multiplexed amplification method (Li et al., 2015). Because of this targeted nature and small amplicon size, the amount of raw data required for accurate gene expression quantification is much smaller with the AmpliSeq technology than classical RNA sequencing. This clearly represents an advantage since for large scale of RNA sequencing experiments involving hundreds or even thousands of samples the bioinformatics analysis as well as the data storage needs can rapidly become strenuous (Schmidt & Hildebrandt, 2017). All the data included in SISTEMA have thus been generated
through this unique experimental workflow and on the same platform, giving rise to
standardized operative procedures and the production of homogenous samples in term of
quality and coverage depth.

Since the derivation of the first human embryonic stem cell line more than two decades ago (Thomson et al., 1998), human pluripotent stem cells have ushered in an exciting new era for
the fields of stem cell biology and biomedical research. Nowadays, human pluripotent stem
cells are widely used for disease modelling, drug discovery and cell therapy development. In
this context, SISTEMA provides a valuable catalogue of human pluripotent stem cell lines
with a representative snapshot of these cells in their undifferentiated state as well as in a wide
range of differentiated cell types. SISTEMA functions as a one-step resource that will assist
the stem cell community to easily evaluate the expression level for their genes of interest,
while comparing them across different hPSC lines, cell types, pathological conditions or after
pharmacological treatments (Figure 6). It therefore opens up a wide spectrum of applications
(Figure 6), such as validating the co-expression of different genes along a differentiation
process or under a pathological condition, assessing their cell-type or temporal specificities of
expression, evaluating the effect of a drug treatment on a set a genes of interest. Most of the
data presented on Sistema are already integrated on largest databases such as Gene Expression
Omnibus. Of interest, our database also includes none published results which provide an
opportunity for the scientific community to access to larger not yet published datasets.

LIMITATIONS OF THE STUDY

Currently, SISTEMA contains data generated from the ampliseq technology and does not
include other omics. Our objective for the future is to continuously feed SISTEMA with new
data generated on our platform without ruling out the possibility of including expression
profiles generated by whole transcriptome RNA sequencing or single cell RNA sequencing,
also generated by the same platform as these approaches provide complementary snapshots (Li et al., 2015).

ACKNOWLEDGEMENTS

I-Stem is part of the Biotherapies Institute for Rare Diseases (BIRD) supported by the Association Française contre les Myopathies (AFM-Téléthon). This project was supported by grants from Laboratoire d’Excellence Revive (Investissement d’Avenir; ANR-10-LABX-73) and the program “investissement d’avenir” INGESTEM. We thank the groups of Anselme Perrier, Laetitia Aubry, Christine Baldeschi, Xavier Nissan, Alexandra Benchoua, Christelle Monville, Christian Pinset, for providing biological resources and Lina El-Kassar for karyotyping the cell lines. We gratefully acknowledge support from the PSMN (Pôle Scientifique de Modélisation Numérique) of the ENS de Lyon for computing resource. The authors would like to thank Jean-Baptiste Claude and Laurent Modolo for the useful discussions on the analysis protocols and the development of the web interface.

AUTHOR CONTRIBUTIONS

Conceptualization: M.J., H.P., D.A., C.M. and M.P.; Methodology: M.J., H.P., D.A. and C.M.; Investigation: M.J., H.P., A.C.; Data Curation: M.J., H.P., A.C.; Software: H.P. and S.J.; Resources: M.J., H.P., A.C.; Writing – Original Draft & Review & Editing: M.J., H.P., D.A., and C.M. Funding Acquisition: C.M., M.P and D.A

DISCLOSURE OF POTENTIAL CONFLICTS OF INTEREST: The authors declare no conflicts of interest.

INCLUSION AND DIVERSITY: While citing references scientifically relevant for this work, we actively worked to promote gender balance in our reference list.
FIGURE LEGENDS

Figure 1. Representation of the phenotype and genotype diversity currently described in the database. (A). Graphical representation in percentage distribution of the 10 principal cell types that are present in SISTEMA database. Reference cells means RNA from commercial cell lines (primary cells) or total RNA extracted from human tissues; i.e. human Brain or Human foetal Brain. (B). Graphical repartition between biological samples that carry a causal mutation associated with a rare disease and healthy non affected (WT) biological samples. (C). Detailed representation of the different monogenic diseases presented in SISTEMA (EB: Epidermolysis bullosa, HD: Huntington, LND: Lesch Nyhan, SMA: Spinal Muscular Atrophy, WFS: Wolfram Syndrome, DMD: Duchenne Muscular Dystrophy, DM1: Myotonic Dystrophy type 1 and Progeria). (E). Graphical representation of pharmacological treatments performed on some biological samples. See also Figure S1 and S2.

Figure 2. Visualization of the Web Interface for single gene analysis. (A). View of the query page with the different filters that can be used for analysis such as cell type, pharmacological treatments, time of differentiation as well as the visualization of the results obtained for NANOG as the gene of interest. (B). Visualization of the general information provided for each gene as well as the different web links. (C). Level of expression of NANOG in hES cells and hES-derived keratinocytes (normalized counts: log10,cpm + 1)). (D). Visualization of the expression table gathering the different information concerning each cellular sample. This table is downloadable as an excel file. Figures C-D represent the results obtained with 9 different hESC lines. See also Table 1.

Figure 3. Visualization of the Web Interface for analysis of multiple genes. (A). for each gene analyzed, a link to NCBI database is associated as well as the possible relation between the different genes analyzed (by using STRING). (B) Histogram representing the comparative expression value for each gene analyzed in hES cells and hES-derived keratinocytes. (C). Hierarchical clustering of the genes SOX2, NANOG and POU5F1B expressed in undifferentiated hESC and in hPSC-derived keratinocytes (normalized counts: log10,cpm + 1), (D) Downloadable expression table and detailed
information about individual biological sample. Figures 3C-D correspond to the representative results obtained with 9 independent hPSC lines.

**Figure 4. Pairwise dissimilarities of markers identified from specific differentiation status** (“complete” method and Euclidean distance) Hierarchical clustering for 10 genes extracted from (A) KEGG “Signaling pathways regulating pluripotency of stem cells” (hsa04550) on all the biological samples included in SISTEMA database. These genes were selected on the basis of the different quality controls performed in the laboratory to validate the pluripotency of the different hiPSC and hES lines. Other genes have been selected based on the literature such as the example of ID1 gene that belongs to a family called “inhibitors of the differentiation” and previously described to be specific of human pluripotent stem cell lines (Aloia et al., 2015; Hong et al., 2011). (B) Level of expression of 10 genes selected from GO term “Spinal Cord Motor Neuron differentiation (GO 0021522) in hPS cells and hPS-derived spinal motoneurons. The genes ISL1/2 and MNX1 (also known as HB9) are two genes well known to be expressed in spinal motoneurons (Maury et al., 2015). We also used the genes referenced in the GO 0021522 and selected the ones that appear the most relevant. As two examples, Pax6 and Olig2 are two factors we routinely used to control the differentiation of human pluripotent stem cells into spinal motoneurons (Maury et al., 2015). (C). Level of expression of 10 genes selected from GO melanocyte differentiation (GO 0030318) in hPS cells and hPS-derived melanocytes.

**Figure 5. Application examples of SISTEMA** (A). SMN1 gene expression in different biological samples clustered in function of their phenotypes or genotypes. Each bar indicates a unique sample. (B). SMN1 expression in fibroblasts (Hoch et al., 2019) and hESC-derived neurons after treatment with different small compounds. Data have been extracted from SISTEMA and are represented as the mean of normalized count ± SEM and were statistically analyzed with DESeq2 R package using Negative binomial generalized linear model and Wald's test; *p < 0.05, **p < 0.01. Chart generated from GraphPad Prism.

**Figure 6. Large spectrum of applications offered by SISTEMA**

**Table 1:** Metadata concerning samples experiments (Related to Figure 2 and STAR Methods).
THE STAR METHODS TEXT

RESOURCE AVAILABILITY

Lead contact

Further information and requests for biological resources or metafiles should be addressed to the Lead Contact, Cécile Martinat (cmartinat@istem.fr) or Margot Jarrige (mjarrige@istem.fr).

Materials availability

This study did generate new web site available at http://sistema.ens-lyon.fr/

This study did not generate new reagents

Data and Code Availability

• Some data sets used in this work are available from publicly available sources and have been deposited at GEO (GSE 104091 and GSE119841). Accession numbers are listed in the key resources table. RNAseq data are publicly available on a user friendly interface and freely available at http://sistema.ens-lyon.fr

• All original code in this paper is available from the lead contact upon request.

• Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

*Please omit this section if your study does not use experimental models typical in the life sciences (e.g., if your study is computational or physical science research). This section should also include information related to cell lines/strains used for in vitro experiments.

RNA collection from biological samples.

RNA samples were extracted from human embryonic stem cell lines and human induced pluripotent stem cell lines collected from Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM). All the Human embryonic stem cell lines indicated in SISTEMA database are referenced on hPSCreg and listed here:
The collected information regarding the cell culture conditions can be found at http://sistema.ens-lyon.fr/Echantillons.php. For the Human induced pluripotent stem cell lines not all are referenced yet but will be integrated when the program of research they are involved in will be published. The cell lines used on the study are listed on Key resources table.

**METHOD DETAILS**

**Data collection.**

Except for the RNA reference sample, all the samples have been generated at ISTEM. Human embryonic cell (hESC) lines were used following the recommendation of the French Law of Bioethics and declared at the French Agency of Biomedicine (Number SASB1020178S). For Human induced pluripotent stem cells (hiPSC), tissue samples were collected under specific participant consents. The SISTEMA portal provides all information regarding the generation, the culture conditions of human pluripotent stem cells (hPSC) as well as their differentiation. Such information is also provided in Table 1. A unique standardized name for each sample was generated by combining “cell type”—”Phenotype”

**RNA extraction.**

Total RNA was extracted using the Qiacube instrument and the RNeasy micro or mini kit depending on the quantity of cells. The quality of the RNA was evaluated using the Bio-
analyseur 2100 Agilent[TM] following the protocol of the manufacturer (Qiagen®). Only samples with a RNA integrity number higher than 7 were processed.

**Library generation and sequencing.**

For each sample, 50 ng of total RNA was reverse transcribed using the Ion AmpliSeq Transcriptome Human Gene Expression kit following the protocol of the manufacturer (Thermofisher Scientific®). Briefly, the cDNA libraries were amplified and barcoded using Ion AmpliSeq Transcriptome Human Gene Expression core panel and Ion Xpress Barcode Adapter, named Amplicons (Thermofisher Scientific®). The amplicons were quantified using Agilent High Sensitivity DNA kit before the samples were pooled in sets of eight. Emulsion PCR and Enrichment was performed on the Ion OT2 system Instrument using the Ion PI Hi-Q OT2 200 kit (Thermofisher Scientific). Samples were loaded on an Ion PI v3 Chip and sequenced on the Ion Proton System using Ion PI Hi-Q sequencing 200 kit chemistry to generated around 10,156,774 reads per sample (200 bp read length; Thermofisher Scientific). The fastQ file was generated on the Ion Torrent server and transfer to the PSMN (for scientific pole for numerical modelling at ENS Lyon; www.ens-lyon.fr/PSMN).

**Amplicon annotation.**

The quality control of the sequencing data was evaluated using FastQC. The reads were trimmed using Prinseq (v0.20.4) (--trim-right 20) and filtered by average quality score (--trim-qual 20). Reads were mapped to the Amplicons sequence reference using bowtie2 local (v2.3.4.1). Reads below a mapping score of 10 or multi-mapped were filtered using samtools (v0.1.19). The level of gene expression was quantified with the specific annotation described above. Each sample was normalized by RPM (reads per Million) and the expression values were converted in logarithm scale (log10(x+1)) using R (v3.6.0).

**Data implementation and web interface.**
SISTEMA database was implemented using MySQL (www.mysql.com) and organized on three main tables. Briefly the first table combines all the information extracted from NCBI such as the chromosome coordinate, the gene type and the gene name (present in the seven main genetic databases: NCBI, EnsEMBL, HGNC IMGT/GENE-DB, MIM et Vega (Ashurst et al., 2005; Yates et al., 2017). The second table contains all the data curated for the sample sequence on the platform such as the library and sequencing number, the sequencing barcode, NGS user, sequencing date, FASTQ_name, sample_name, replicate, cell line, cell matrix, the team and the project). The last table corresponds to the expression value counting table transform in logarithm scale. For presentation and interpretation of the data, we chose to provide only reads per million for each gene analyzed rather than a presentation based on fold change analysis.

The web interface was developed using the programming language Php and JavaScript. Highcharts (© 2020 www.highcharts.com) and StringDB (Jeanquartier et al., 2015) APIs (application programming interface) were used. For NCBI viewer, a direct URL link in an embedded window is used. To facilitate the comprehension of the figures 2C, 3B-C and 4 on the manuscript we used color coding to label the cell type using R. The figure 5B was generated using Graphpad Prism 5.

QUANTIFICATION AND STATISTICAL ANALYSIS
For RNAseq quantification, reads were mapped to the Amplicons sequence reference using bowtie2 local (v2.3.4.1). Reads below a mapping score of 10 or multi-mapped were filtered using samtools (v0.1.19). The level of gene expression was quantified with the specific annotation described above. Each sample was normalized by RPM (reads per Million) and the expression values were converted in logarithm scale (log10(x+1)) using R (v3.6.0). For Figure 5, data have been extracted from SISTEMA and are represented as the mean of normalized count ± SEM and were statistically analyzed with DESeq2 R package using Negative binomial generalized linear model and Wald's test; *p < 0.05, **p < 0.01.
**ADDITIONAL RESOURCES**

Further information relevant to the study can be found on http://sistema.ens-lyon.fr/index.php or on http://sistema.ens-lyon.fr/MatAndMeth.php

**REFERENCES**

Aloia L, Gutierrez A, Caballero JM, Di Croce L (2015). Direct interaction between Id1 and Zrf1 controls neural differentiation of embryonic stem cells. EMBO Rep **16**: 63-70. 10.15252/embr.201439560

Ashurst JL, Chen CK, Gilbert JG, Jekosch K, Keenan S, Meidl P, Searle SM, Stalker J, Storey R, Trevanion S, Wilming L, Hubbard T (2005). The Vertebrate Genome Annotation (Vega) database. Nucleic Acids Res **33**: D459-465. 10.1093/nar/gki135

Assou S, Le Carrou T, Tondeur S, Strom S, Gabelle A, Marty S, Nadal L, Pantesco V, Reme T, Hugnot JP, Gasca S, Hovatta O, Hamamah S, Klein B, De Vos J (2007). A meta-analysis of human embryonic stem cells transcriptome integrated into a web-based expression atlas. Stem Cells **25**: 961-973. 10.1634/stemcells.2006-0352

Au KF, Sebastiano V (2014). The transcriptome of human pluripotent stem cells. Curr Opin Genet Dev **28**: 71-77. 10.1016/j.gde.2014.09.012

Bassez G, Audureau E, Hogrel JY, Arrouasse R, Baghdoyan S, Bhugaloo H, Gourlay-Chu ML, Le Corvoisier P, Peschanski M (2018). Improved mobility with metformin in patients with myotonic dystrophy type 1: a randomized controlled trial. Brain **141**: 2855-2865. 10.1093/brain/awy231

Ben M’Barek K, Bertin S, Brazhnikova E, Jaillard C, Habeler W, Plancheron A, Fovet CM, Demilly J, Jarraya M, Bejanariu A, Sahel JA, Peschanski M, Goureaud O, Monville C (2020). Clinical-grade production and safe delivery of human ESC derived RPE sheets in primates and rodents. Biomaterials **230**: 119603. 10.1016/j.biomaterials.2019.119603

Ben M’Barek K, Habeler W, Plancheron A, Jarraya M, Regent F, Terray A, Yang Y, Chatrousse L, Domingues S, Masson Y, Sahel JA, Peschanski M, Goureaud O, Monville C (2017). Human ESC-derived retinal epithelial cell sheets potentiate rescue of photoreceptor cell loss in rats with retinal degeneration. Sci Transl Med **9**10.1126/scitranslmed.aai7471

Brichta L, Hofmann Y, Hahnen E, Siebzehnrubl FA, Raschke H, Blumcke I, Eypupoglu IY, Wirth B (2003). Valproic acid increases the SMN2 protein level: a well-known drug as a potential therapy for spinal muscular atrophy. Hum Mol Genet **12**: 2481-2489. 10.1093/hmg/ddg256
Chambers I, Silva J, Colby D, Nichols J, Nijmeijer B, Robertson M, Vrana J, Jones K, Grotewold L, Smith A (2007). Nanog safeguards pluripotency and mediates germline development. Nature 450: 1230-1234. 10.1038/nature06403

Chambers I, Tomlinson SR (2009). The transcriptional foundation of pluripotency. Development 136: 2311-2322. 10.1242/dev.024398

Dachs E, Piedrafita L, Hereu M, Esquerda JE, Caldero J (2013). Chronic treatment with lithium does not improve neuromuscular phenotype in a mouse model of severe spinal muscular atrophy. Neuroscience 250: 417-433. 10.1016/j.neuroscience.2013.07.026

Darville H, Poulet A, Rodet-Amsellem F, Chatrousse L, Pernelle J, Boissart C, Heron D, Nava C, Perrier A, Jarrige M, Coge F, Millan MJ, Bourgeron T, Peschanski M, Delorme R, Benchoua A (2016). Human Pluripotent Stem Cell-derived Cortical Neurons for High Throughput Medication Screening in Autism: A Proof of Concept Study in SHANK3 Haploinsufficiency Syndrome. EBioMedicine 9: 293-305. 10.1016/j.ebiom.2016.05.032

Elshafay A, Hieu TH, Doheim MF, Kassem MAM, MF EL, Holloway SK, Abo-Elghar H, Hirayama K, Huy NT (2019). Efficacy and Safety of Valproic Acid for Spinal Muscular Atrophy: A Systematic Review and Meta-Analysis. CNS drugs 33: 239-250. 10.1007/s40263-019-00606-6

Galvan L, Francelle L, Gaillard MC, de Longprez L, Carrillo-de Sauvage MA, Liot G, Cambon K, Stimmer L, Luccantoni S, Flament J, Valette J, de Chaldee M, Auregan G, Guillermier M, Josephine C, Petit F, Jan C, Jarrige M, Dufour N, Bonvento G, Humbert S, Saudou F, Hantraye P, Merienne K, Bemelmans AP, Perrier AL, Deglon N, Brouillet E (2018). The striatal kinase DCLK3 produces neuroprotection against mutant huntingtin. Brain 141: 1434-1454. 10.1093/brain/awy057

Ghosh A, Som A (2020). RNA-Seq analysis reveals pluripotency-associated genes and their interaction networks in human embryonic stem cells. computational biology and chemistry 85: 1476-9271. https://doi.org/10.1016/j.compbiolchem.2020.107239

Godoy P, Schmidt-Heck W, Hellwig B, Nell P, Feuerborn D, Rahnenfuhrer J, Kattler K, Walter J, Bluthgen N, Hengstler JG (2018). Assessment of stem cell differentiation based on genome-wide expression profiles. Philosophical transactions of the Royal Society of London Series B, Biological sciences 37310.1098/rstb.2017.0221

Guenou H, Nissan X, Larcher F, Feteira J, Lemaitre G, Saidani M, Del Rio M, Barrault CC, Bernard FX, Peschanski M, Baldeschi C, Waksman G (2009). Human embryonic stem-cell derivatives for full reconstruction of the pluristratified epidermis: a preclinical study. Lancet 374: 1745-1753.

Harhouri K, Navarro C, Depetris D, Mattei MG, Nissan X, Cau P, De Sandre-Giovannoli A, Levy N (2017). MG132-induced progerin clearance is mediated by autophagy activation and splicing regulation. EMBO molecular medicine 9: 1294-1313. 10.15252/emmm.201607315
Hoch L, Henriques SF, Bruge C, Marsolier J, Benabides M, Bourg N, Tournois J, Mahe G, Morizur L, Jarrige M, Bigot A, Richard I, Nissan X (2019). Identification of thioestrepton as a pharmacological approach to rescue misfolded alpha-sarcoglycan mutant proteins from degradation. Scientific reports 9: 6915. 10.1038/s41598-019-43399-w

Hong SH, Lee JH, Lee JB, Ji J, Bhatia M (2011). ID1 and ID3 represent conserved negative regulators of human embryonic and induced pluripotent stem cell hematopoiesis. J Cell Sci 124: 1445-1452. 10.1242/jcs.077511

Howlin J, Cirenajwis H, Lettiero B, Staaf J, Lauss M, Saal L, Borg A, Gruvberger-Saal S, Jonsson G (2015). Loss of CITED1, an MITF regulator, drives a phenotype switch in vitro and can predict clinical outcome in primary melanoma tumours. PeerJ 3: e788. 10.7717/peerj.788

Jeanquartier F, Jean-Quartier C, Holzinger A (2015). Integrated web visualizations for protein-protein interaction databases. BMC bioinformatics 16: 195. 10.1186/s12859-015-0615-z

Kusters-Vandevelde HV, Klaasen A, Kusters B, Groenen PJ, van Engen-van Grunsven IA, van Dijk MR, Reifenberger G, Wesseling P, Blokx WA (2010). Activating mutations of the GNAQ gene: a frequent event in primary melanocytic neoplasms of the central nervous system. Acta neuropathologica 119: 317-323. 10.1007/s00401-009-0611-3

Kwon DY, Motley WW, Fischbeck KH, Burnett BG (2011). Increasing expression and decreasing degradation of SMN ameliorate the spinal muscular atrophy phenotype in mice. Hum Mol Genet 20: 3667-3677. 10.1093/hmg/ddr288

Lefebvre S, Burglen L, Reboullet S, Clermont O, Burlet P, Viollet L, Benichou B, Cruaud C, Millasseau P, Zeviani M, et al. (1995). Identification and characterization of a spinal muscular atrophy-determining gene. Cell 80: 155-165.

Li W, Turner A, Aggarwal P, Matter A, Storvick E, Arnett DK, Broeckel U (2015). Comprehensive evaluation of AmpliSeq transcriptome, a novel targeted whole transcriptome RNA sequencing methodology for global gene expression analysis. BMC genomics 16: 1069. 10.1186/s12864-015-2270-1

Locatelli D, Terao M, Kurosaki M, Zanellati MC, Pletto DR, Finardi A, Colciaghi F, Garattini E, Battaglia GS (2015). Different Stability and Proteasome-Mediated Degradation Rate of SMN Protein Isoforms. PLoS One 10: e0134163. 10.1371/journal.pone.0134163

Mair B, Tomic J, Masud SN, Tonge P, Weiss A, Usaj M, Tong AHY, Kwan JI, Brown KR, Titus E, Atkins M, Chan KSK, Munsie L, Habsid A, Han H, Kennedy M, Cohen B, Keller G, Moffat J (2019). Essential Gene Profiles for Human Pluripotent Stem Cells Identify Uncharacterized Genes and Substrate Dependencies. Cell Rep 27: 599-615 e512. 10.1016/j.celrep.2019.02.041
Mallon BS, Chenoweth JG, Johnson KR, Hamilton RS, Tesar PJ, Yavatkar AS, Tyson LJ, Park K, Chen KG, Fann YC, McKay RD (2013). StemCellDB: the human pluripotent stem cell database at the National Institutes of Health. Stem Cell Res 10: 57-66. 10.1016/j.scr.2012.09.002

Matera I, Watkins-Chow DE, Loftus SK, Hou L, Incao A, Silver DL, Rivas C, Elliott EC, Baxter LL, Pavan WJ (2008). A sensitized mutagenesis screen identifies Gli3 as a modifier of Sox10 neurocristopathy. Hum Mol Genet 17: 2118-2131. 10.1093/hmg/ddn110

Maury Y, Come J, Piskorowski RA, Salah-Mohellihi N, Chevaleyre V, Peschanski M, Martinat C, Nedelec S (2015). Combinatorial analysis of developmental cues efficiently converts human pluripotent stem cells into multiple neuronal subtypes. Nat Biotechnol 33: 89-96.

Mournetas V, Massourides E, Dupont JB, Kornobis E, Polvèche H, Jarrige M, Gosselin M, Manoussopoulou A, Garbis SD, Gorecki D, Pinset C (2019). Myogenesis modelled by human pluripotent stem cells uncovers Duchenne muscular dystrophy phenotypes prior to skeletal muscle commitment. BioRxiv https://doi.org/10.1101/720920

Nair SS, Chaubal VA, Shioda T, Coser KR, Mojamdar M (2001). Over-expression of MSG1 transcriptional co-activator increases melanin in B16 melanoma cells: a possible role for MSG1 in melanogenesis. Pigment cell res 14: 206-209. 10.1034/j.1600-0749.2001.140311.x

Nicoleau C, Varela C, Bonnefond C, Maury Y, Bugi A, Aubry L, Viegas P, Bourgois-Rocha F, Peschanski M, Perrier AL (2013). Embryonic stem cells neural differentiation qualifies the role of Wnt/beta-Catenin signals in human telencephalic specification and regionalization. Stem Cells 31: 1763-1774.

Park PJ, Lee TR, Cho EG (2015). Substance P stimulates endothelin 1 secretion via endothelin-converting enzyme 1 and promotes melanogenesis in human melanocytes. The Journal of investigative dermatology 135: 551-559. 10.1038/jid.2014.423

Piepers S, Cobben JM, Sodaar P, Jansen MD, Wadman RI, Meester-Delver A, Poll-The BT, Lemmink HH, Wokke JH, van der Pol WL, van den Berg LH (2011). Quantification of SMN protein in leucocytes from spinal muscular atrophy patients: effects of treatment with valproic acid. Journal of neurology, neurosurgery, and psychiatry 82: 850-852. 10.1136/jnnp.2009.200253

Pinto JP, Machado RSR, Magno R, Oliveira DV, Machado S, Andrade RP, Braganca J, Duarte I, Futschik ME (2018). StemMapper: a curated gene expression database for stem cell lineage analysis. Nucleic Acids Res 46: D788-D793. 10.1093/nar/gkx921

Rowe RG, Daley GQ (2019). Induced pluripotent stem cells in disease modelling and drug discovery. Nature reviews Genetics 20: 377-388. 10.1038/s41576-019-0100-z

Schmidt B, Hildebrandt A (2017). Next-generation sequencing: big data meets high performance computing. Drug discovery today 22: 712-717. 10.1016/j.drudis.2017.01.014
Silva J, Nichols J, Theunissen TW, Guo G, van Oosten AL, Barrandon O, Wray J, Yamanaka S, Chambers I, Smith A (2009). Nanog is the gateway to the pluripotent ground state. Cell 138: 722-737. 10.1016/j.cell.2009.07.039

Streeter I, Harrison PW, Faulconbridge A, The HipSci C, Flicek P, Parkinson H, Clarke L (2017). The human-induced pluripotent stem cell initiative-data resources for cellular genetics. Nucleic Acids Res 45: D691-D697. 10.1093/nar/gkw928

Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, Yamanaka S (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell 131: 861-872.

Thomson JA, Itskovitz-Eldor J, Shapiro SS, Waknitz MA, Swiergiel JJ, Marshall VS, Jones JM (1998). Embryonic stem cell lines derived from human blastocysts. Science 282: 1145-1147.

Van Raamsdonk CD, Fitch KR, Fuchs H, de Angelis MH, Barsh GS (2004). Effects of G-protein mutations on skin color. Nat Genet 36: 961-968. 10.1038/ng1412

Yates B, Braschi B, Gray KA, Seal RL, Tweedie S, Bruford EA (2017). Genenames.org: the HGNC and VGNC resources in 2017. Nucleic Acids Res 45: D619-D625. 10.1093/nar/gkw1033
Figure 1

A

- Melanocytes & Keratinocytes: 9%
- Myogenic cells: 9%
- Mesenchymal Stem cells: 1%
- Motoneurons: 9%
- Fibroblasts: 3%
- hES hIPS cells: 11%
- Retinal pigment epithelium cells: 9%
- Reference cells: 5%
- Brain Neuronal cells: 44%

B

- Diseases: 27%
- WT: 73%

C

- Progeria: 2.6%
- WFS: 21.4%
- DM1: 2.6%
- DMD: 14.5%
- EB: 2.6%
- SMA: 15.4%
- LND: 6.8%
- HD: 34.2%

D

- Kenpaullone: 20%
- DMSO: 18%
- Thioridazine: 12%
- ATRA: 6%
- MG132: 6%
- Lithium: 6%
- VPA: 27%
- Bortezomib: 6%
Figure 2

A

B

C

D

NANOG : Nanog homeobox (protein-coding)

Localization : 12p13.33 chr12:7780395-7789660 (+) 4x cons
Synonyms :

References : GeneCards NCBI : 79923 MIM : 6079337 HGNC : 12087 Ensembl : ENSG00000111704 Vega : OTHEUM00000166019

D

| #   | NGS  | Name             | Counts | Replicates | Cell type | Phenotype | Cell line | Time of differentiation | Derived Treatment / Time | GEO  |
|-----|------|------------------|--------|------------|-----------|-----------|-----------|------------------------|--------------------------|------|
| 19  | NGS1 | H1_NESC_WT       | 2.64   | 1          | hESCs     | WT        | SA001     | NA                     | NA                       | NA   |
| 21  | NGS1 | H1_NESC_WT       | 2.4    | 1          | hESCs     | WT        | SA001     | NA                     | NA                       | NA   |
| 22  | NGS1 | H1_NESC_WT       | 2.49   | 1          | hESCs     | WT        | SA001     | NA                     | NA                       | NA   |
| 23  | NGS1 | H1_NESC_WT       | 2.6    | 1          | hESCs     | WT        | SA001     | NA                     | NA                       | NA   |
| 24  | NGS1 | H1_NESC_WT       | 2.52   | 2          | hESCs     | WT        | SA001     | NA                     | NA                       | NA   |
| 25  | NGS1 | H1_NESC_WT       | 2.54   | 3          | hESCs     | WT        | SA001     | NA                     | NA                       | NA   |
| 26  | NGS1 | H1_NESC_WT       | 2.43   | 1          | hESCs     | WT        | SA001     | NA                     | NA                       | NA   |
| 27  | NGS1 | H1_NESC_WT       | 2.73   | 3          | hESCs     | WT        | SA001     | NA                     | NA                       | NA   |
| 28  | NGS1 | H1_NESC_WT       | 2.54   | 2          | hESCs     | WT        | SA001     | NA                     | NA                       | NA   |
| 29  | NGS1 | H1_NESC_WT       | 0.17   | 2          | hPLSC-derived Keratinocyte | EB | 10295 | NA | NA | NA |
| 30  | NGS1 | H1_NESC_WT       | 0.24   | 3          | hPLSC-derived Keratinocyte | EB | NAMAD | NA | NA | NA |
| 31  | NGS1 | H1_NESC_WT       | 0.15   | 1          | hPLSC-derived Keratinocyte | WT | PC081 | NA | NA | NA |
| 32  | NGS1 | H1_NESC_WT       | 0.37   | 2          | hPLSC-derived Keratinocyte | WT | PC1432 | NA | NA | NA |
| 33  | NGS1 | H1_NESC_WT       | 0.43   | 3          | hPLSC-derived Keratinocyte | WT | PC15   | NA | NA | NA |
| 34  | NGS1 | H1_NESC_WT       | 0.22   | 1          | hPLSC-derived Keratinocyte | EB | 9991  | NA | NA | NA |
Figure 5

A

SMN1

Normalized counts (log10)

DM1  DMD  EB  LND  SMA  WT

Myogenic cells
hES hiPSC cells
Keratinocytes
Brain Neuronal cells
Motoneurons

B

SMN1 expression

Normalized counts (log10)

Fibroblast + DMSO
Fibroblast + BTZ
Fibroblast + MG132
Fibroblast + Tsp 3μM
hESC-derived neurons + DMSO
hESC-derived neurons + Lithium
hESC-derived neurons + VPA
Figure 6

Gene expression and visualisation on:

- Sample
- Genotype
- Treatment
- Co-expressed genes
- Cell type biomarker
- Disease specific genes
- Drug effect on gene expression
