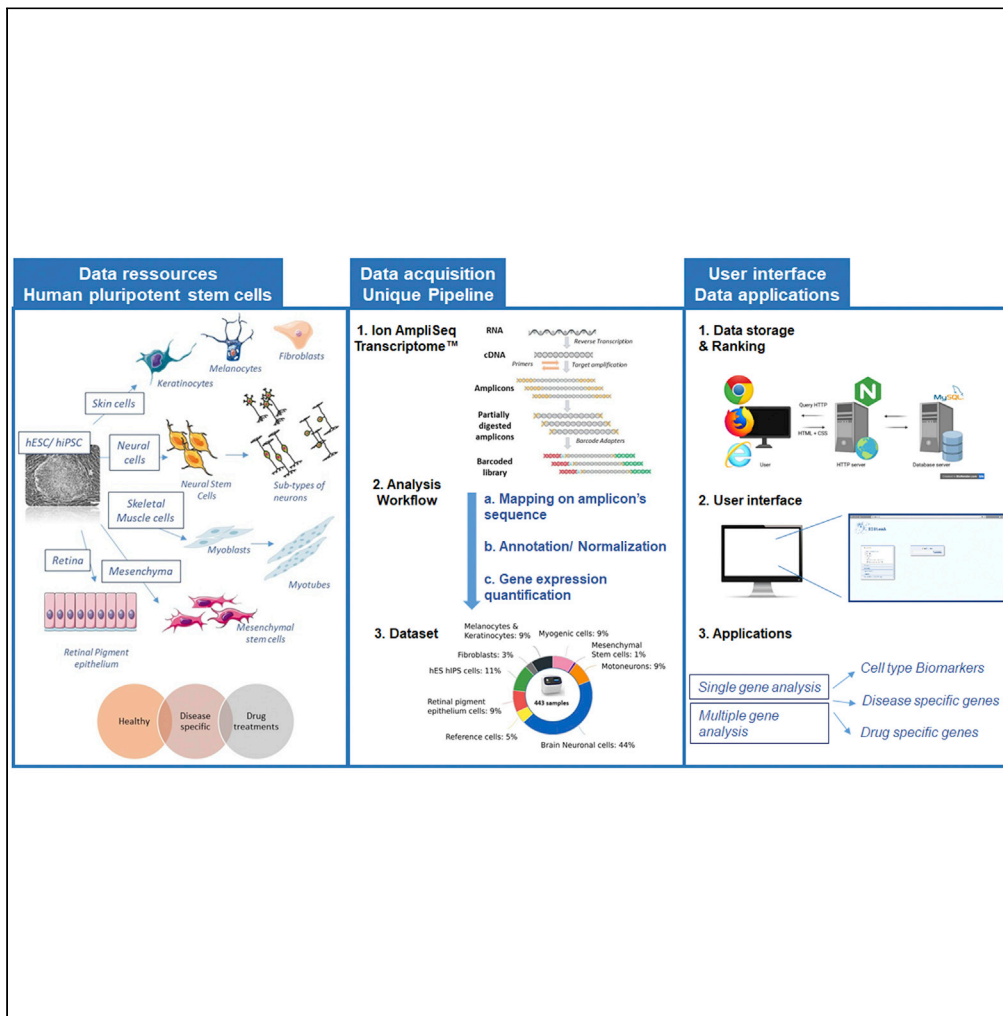


Article

SISTEMA: A large and standardized collection of transcriptome data sets for human pluripotent stem cell research



Margot Jarrige,
Hélène Polvêche,
Alexandre
Carteron,
Stéphane
Janczarski, Marc
Peschanski, Didier
Auboeuf, Cécile
Martinat

cmartinat@istem.fr

Highlights
SISTEMA is a curated
gene expression database
using human pluripotent
stem cell lines

A single experimental and
analytical pipeline are
used for the
transcriptomic analyses

SISTEMA is a user friendly
Web portal designed for
the stem cell community



Article

SISTEMA: A large and standardized collection of transcriptome data sets for human pluripotent stem cell research

Margot Jarrige,^{1,4} Hélène Polvêche,^{1,4} Alexandre Carteron,¹ Stéphane Janczarski,³ Marc Peschanski,¹ Didier Auboeuf,³ and Cécile Martinat^{2,5,*}

SUMMARY

Human pluripotent stem cells have ushered in an exciting new era for disease modeling, drug discovery, and cell therapy development. Continued progress toward realizing the potential of human pluripotent stem cells will be facilitated by robust data sets 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 catalog 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 and Daley, 2019).

Genome-wide expression profiling has been a common end point 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 and Sebastiano, 2014; Ghosh and 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 data sets (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 data set 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

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, 69007 Lyon, France

⁴These authors contributed equally

⁵Lead contact

*Correspondence:

cmartinat@istem.fr

<https://doi.org/10.1016/j.isci.2021.102767>



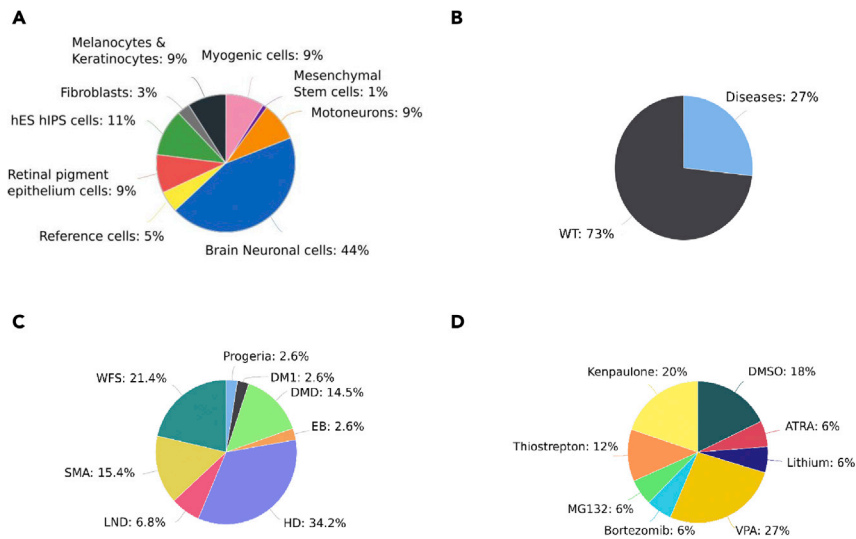


Figure 1. Representation of the phenotype and genotype diversity currently described in the database

(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 fetal 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).

(D) Graphical representation of pharmacological treatments performed on some biological samples.

See also [Figures S1](#) and [S2](#).

representing the different germ layers ([Guenou et al., 2009](#); [Maury et al., 2015](#); [Mourmetas 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 ([Figures 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 haploinsufficiency, MG132 for Hutchinson-Gilford progeria syndrome, thiostrepton and bortezomib for limb-girdle muscular dystrophy type 2D, VPA for Wolfram syndrome or Kenpaulone for spinal muscular atrophy (SMA) ([Darville et al., 2016](#); [Harhoury 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 was 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

A

Select from 6 sample(s). Selected items: hPSC-derived Keratinocyte.

<input checked="" type="checkbox"/>	All	Name	Replicate	Cell type	Phenotype	Cell line	Time of differentiation	Treatment / Time (derived)	GEO
<input checked="" type="checkbox"/>		478_hPSC-derived Keratinocyte_WT	1	hiPSC-derived Keratinocyte	WT	PC085	NA	NA	NA
<input checked="" type="checkbox"/>		479_hPSC-derived Keratinocyte_WT	2	hiPSC-derived Keratinocyte	WT	PC1432	NA	NA	NA
<input checked="" type="checkbox"/>		480_hPSC-derived Keratinocyte_WT	3	hiPSC-derived Keratinocyte	WT	PC15	NA	NA	NA
<input checked="" type="checkbox"/>		481_hPSC-derived Keratinocyte_EB	1	hiPSC-derived Keratinocyte	EB	9991	NA	NA	NA
<input checked="" type="checkbox"/>		482_hPSC-derived Keratinocyte_EB	2	hiPSC-derived Keratinocyte	EB	10295	NA	NA	NA
<input checked="" type="checkbox"/>		483_hPSC-derived Keratinocyte_EB	3	hiPSC-derived Keratinocyte	EB	NAMAD	NA	NA	NA

B

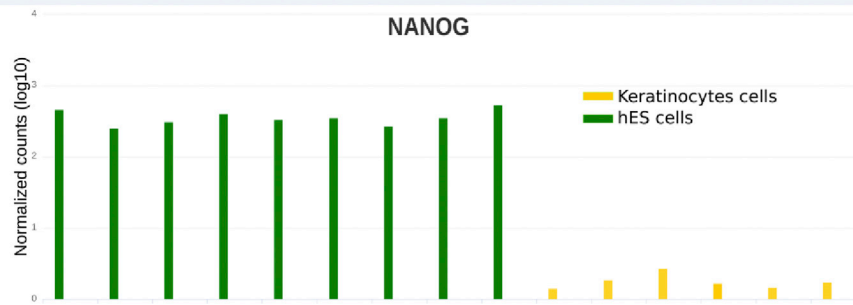
NANOG : *Nanog homeobox* (protein-coding)

Localization : 12p13.31 chr12:7789395-7796060 (+) 4 exons

Synonyms :

References : GeneCards NCBI : 79923 MIM : 607937 HGNC : 20857 Ensembl : ENSG00000111704 Vega : OTIHUMG00000166019

C



D

Counting scale: 0, 0 - 1,40, 1,41 - 1,70, 1,71 - 2, 2 - 2,70, 2,71 - 3, 3 - 4, 4 - 4,65. Export (xlsx)

#	NGS	Name	Counts	Replicate	Cell type	Phenotype	Cell line	Time of differentiation	Derived Treatment / Time	GEO
19	NGS13	19_hESC_WT	2.64	1	hESC	WT	SA001	NA	NA	NA
23	NGS13	23_hESC_WT	2.4	1	hESC	WT	SA001	NA	NA	NA
22	NGS13	22_hESC_WT	2.49	1	hESC	WT	SA001	NA	NA	NA
28	NGS14	28_hESC_WT	2.6	1	hESC	WT	SA001	NA	NA	NA
29	NGS14	29_hESC_WT	2.52	2	hESC	WT	SA001	NA	NA	NA
30	NGS14	30_hESC_WT	2.54	3	hESC	WT	SA001	NA	NA	NA
412	NGS29	412_hESC_WT	2.43	1	hESC	WT	SA001	NA	NA	NA
428	NGS29	428_hESC_WT	2.73	3	hESC	WT	SA001	NA	NA	NA
420	NGS29	420_hESC_WT	2.54	2	hESC	WT	SA001	NA	NA	NA
482	NGS55	482_hPSC-derived Keratinocyte_EB	0.17	2	hiPSC-derived Keratinocyte	EB	10295	NA	NA	NA
483	NGS55	483_hPSC-derived Keratinocyte_EB	0.24	3	hiPSC-derived Keratinocyte	EB	NAMAD	NA	NA	NA
478	NGS55	478_hPSC-derived Keratinocyte_WT	0.15	1	hiPSC-derived Keratinocyte	WT	PC085	NA	NA	NA
479	NGS55	479_hPSC-derived Keratinocyte_WT	0.27	2	hiPSC-derived Keratinocyte	WT	PC1432	NA	NA	NA
480	NGS55	480_hPSC-derived Keratinocyte_WT	0.43	3	hiPSC-derived Keratinocyte	WT	PC15	NA	NA	NA
481	NGS55	481_hPSC-derived Keratinocyte_EB	0.22	1	hiPSC-derived Keratinocyte	EB	9991	NA	NA	NA

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: $\log_{10}(\text{cpm} + 1)$).
 (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 S1](#).

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 ([Figures S2A and S2B](#)). The sequencing process was followed by a standardized data analyzing pipeline as described in [Figure S2A](#). 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 multimapped were filtered using samtools (v0.1.19). Samples were normalized by RPM (reads per million) and the expression values were converted into logarithm scale ($\log_{10}(x+1)$) using R (v3.6.0).

Web interface and graphical visualization

All expression profiles from the 443 experimental data sets have been implemented in the form of a Web-resource called SISTEMA (<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 data sets, 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 data sets pertaining to the cell lineage(s) of interest, the genotype(s), the cell line(s), the time of differentiation, and 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 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 are processed, detailed information regarding the chromosomic localization, the gene structure, and its interacting partners is 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, NANOG was found to be highly expressed in pluripotent cells whereas its expression is absent in the hiPSC-derived keratinocytes ([Figures 2C and 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 ([Table S1](#)), 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 multigene 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 NANOG, OCT4, and SOX2, the triad of transcription factors known to be critical for self-renewal and pluripotency ([Chambers and Tomlinson, 2009](#)), have 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 ([Figures 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, was 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

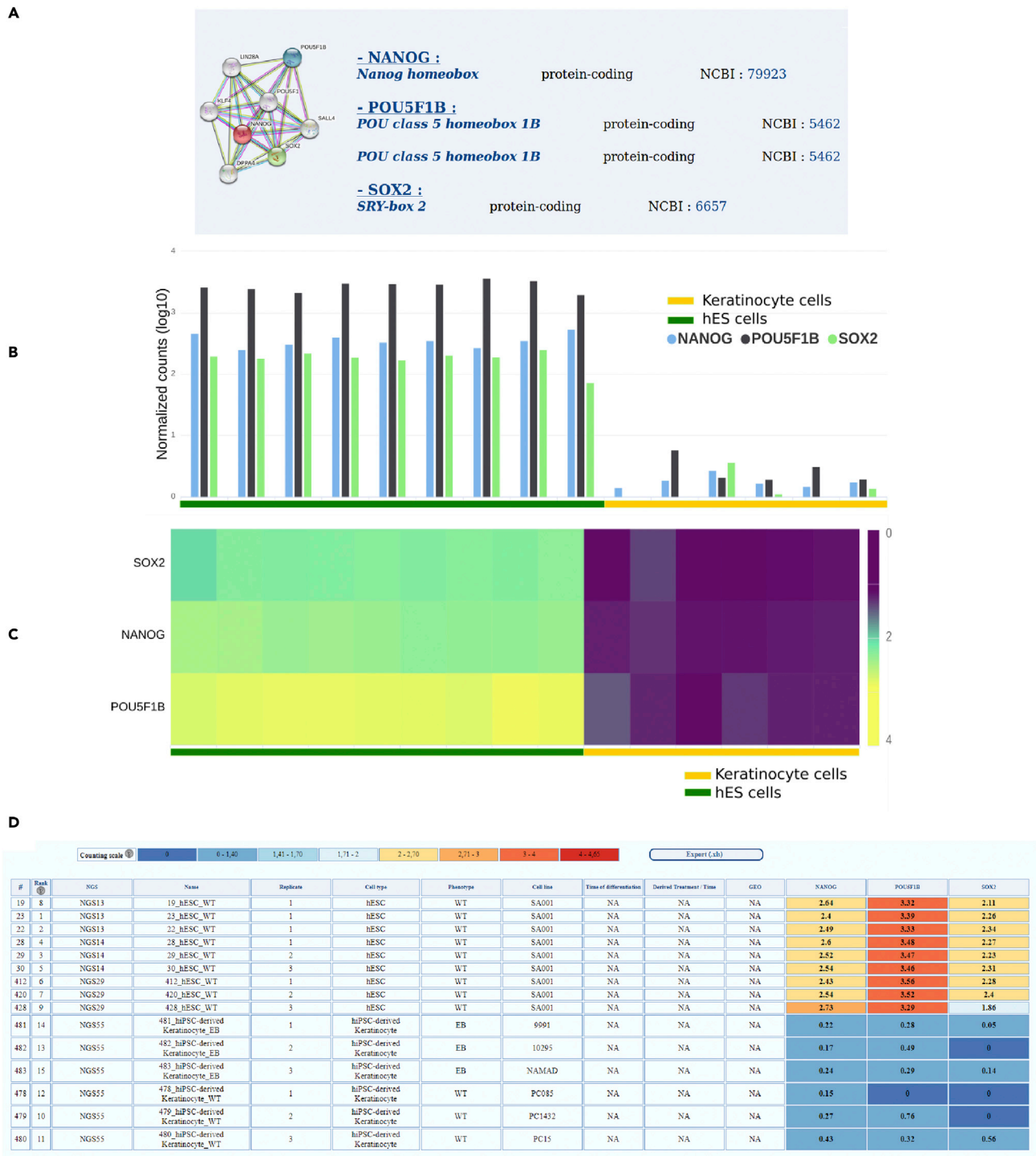


Figure 3. Visualization of the Web Interface for analysis of multiple genes

(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: $\log_{10}(\text{cpm} + 1)$).
 (D) Downloadable expression table and detailed information about individual biological sample. Figures 3C and 3D correspond to the representative results obtained with 9 independent hPSC lines.

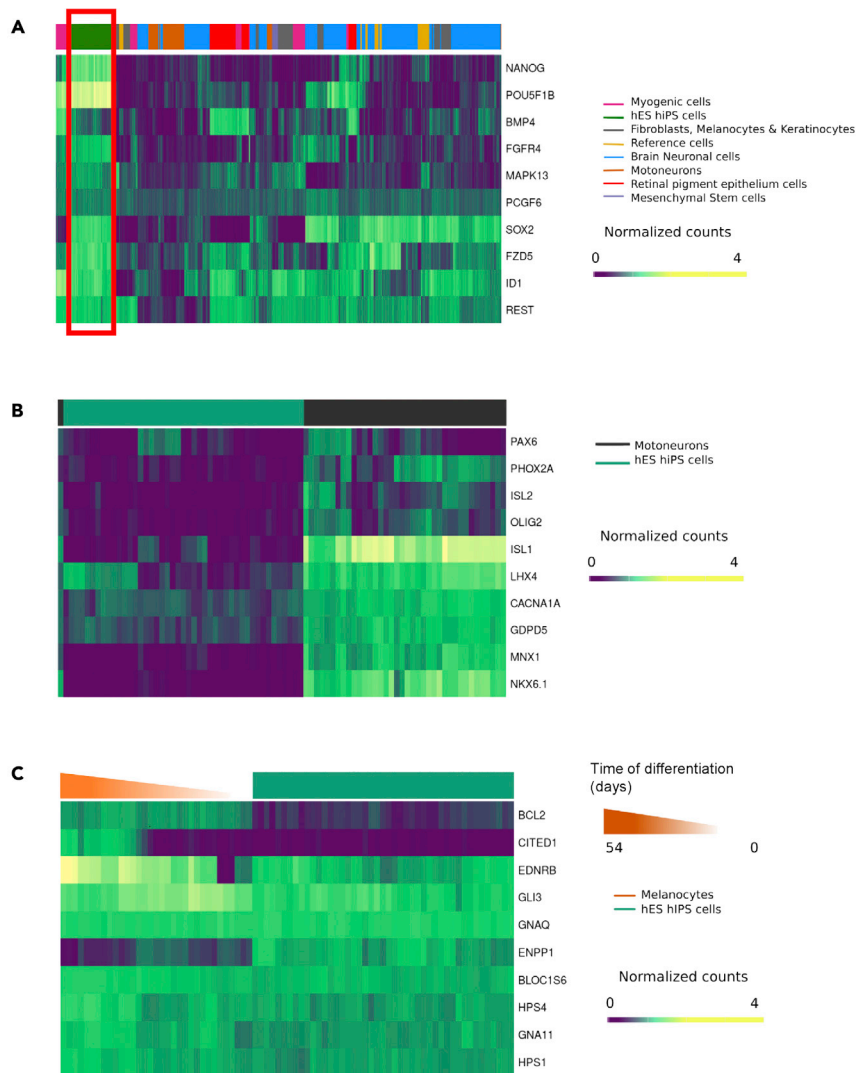


Figure 4. Pairwise dissimilarities of markers identified from specific differentiation status (“complete” method and Euclidean distance)

(A) 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.

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-Vandeveldt 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 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, thioistrepton, 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). These online resources are based on either microarrays or whole transcriptome RNA sequencing technologies, whereas 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 data set 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 iPSCs 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 and 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 homogeneous 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

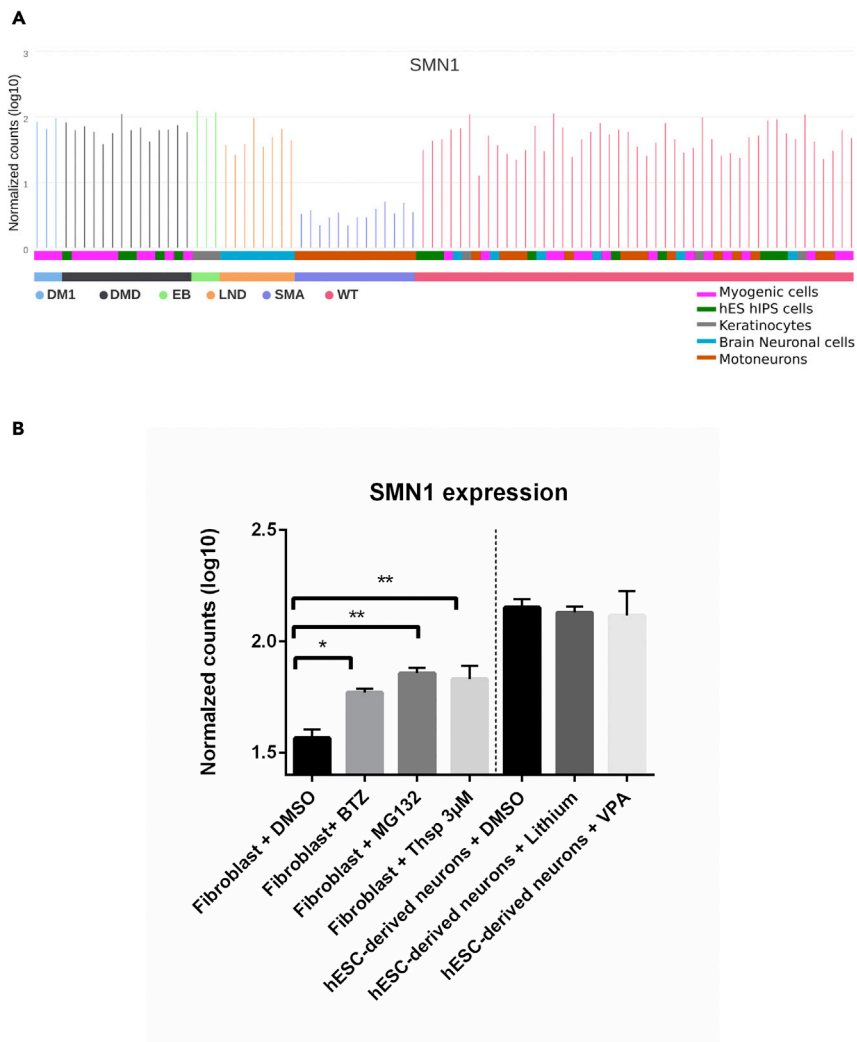


Figure 5. Application examples of SISTEMA

(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 \pm 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.

and biomedical research. Nowadays, human pluripotent stem cells are widely used for disease modeling, drug discovery, and cell therapy development. In this context, SISTEMA provides a valuable catalog 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 coexpression of different genes along a differentiation process or under a pathological condition, assessing their cell type or temporal specificities of expression, and 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.

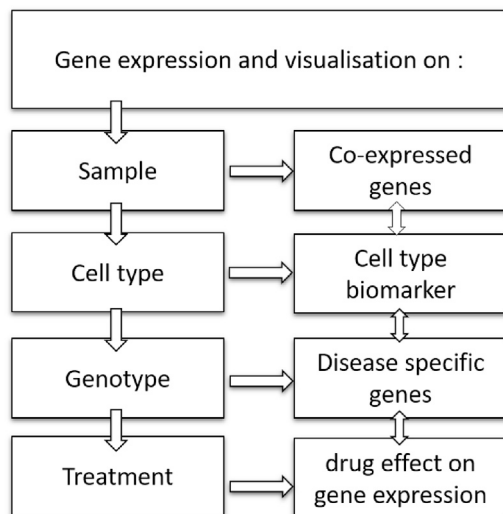


Figure 6. Large spectrum of applications offered by SISTEMA

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).

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
 - Lead contact
 - Materials availability
 - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
 - RNA collection from biological samples
- METHOD DETAILS
 - Data collection
 - RNA extraction
 - Library generation and sequencing
 - Amplicon annotation
 - Data implementation and web interface
- QUANTIFICATION AND STATISTICAL ANALYSIS
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.isci.2021.102767>.

ACKNOWLEDGMENTS

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., and A.C.; Data Curation: M.J., H.P., and A.C.; Software H.P. and S.J.; Resources: M.J., H.P., and 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.

DECLARATION OF INTERESTS

The authors declare no conflicts of interest.

INCLUSION AND DIVERSITY

While citing references scientifically relevant for this work, we also actively worked to promote gender balance in our reference list.

Received: December 21, 2020

Revised: March 29, 2021

Accepted: June 21, 2021

Published: July 23, 2021

REFERENCES

- Aloia, L., Gutierrez, A., Caballero, J.M., and Di Croce, L. (2015). Direct interaction between *Id1* and *Zrf1* controls neural differentiation of embryonic stem cells. *EMBO Rep.* 16, 63–70. <https://doi.org/10.15252/embr.201439560>.
- Ashurst, J.L., Chen, C.K., Gilbert, J.G., Jekosch, K., Keenan, S., Meidl, P., Searle, S.M., Stalker, J., Storey, R., Trevanion, S., et al. (2005). The vertebrate genome annotation (Vega) database. *Nucleic Acids Res.* 33, D459–D465. <https://doi.org/10.1093/nar/gki135>.
- Assou, S., Le Carrour, T., Tondeur, S., Strom, S., Gabelle, A., Marty, S., Nadal, L., Pantesco, V., Reme, T., Hugnot, J.P., et al. (2007). A meta-analysis of human embryonic stem cells transcriptome integrated into a web-based expression atlas. *Stem Cells* 25, 961–973. <https://doi.org/10.1634/stemcells.2006-0352>.
- Au, K.F., and Sebastiano, V. (2014). The transcriptome of human pluripotent stem cells. *Curr. Opin. Genet. Dev.* 28, 71–77. <https://doi.org/10.1016/j.gde.2014.09.012>.
- Bassez, G., Audureau, E., Hogrel, J.Y., Arrouasse, R., Baghdoyan, S., Bhugalo, H., Gourlay-Chu, M.L., Le Corvoisier, P., and Peschanski, M. (2018). Improved mobility with metformin in patients with myotonic dystrophy type 1: a randomized controlled trial. *Brain* 141, 2855–2865. <https://doi.org/10.1093/brain/awy231>.
- Ben M'Barek, K., Bertin, S., Brazhnikova, E., Jaillard, C., Habeler, W., Plancheron, A., Fovet, C.M., Demilly, J., Jarraya, M., Bejanariu, A., et al. (2020). Clinical-grade production and safe delivery of human ESC derived RPE sheets in primates and rodents. *Biomaterials* 230, 119603. <https://doi.org/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., et al. (2017). Human ESC-derived retinal epithelial cell sheets potentiate rescue of photoreceptor cell loss in rats with retinal degeneration. *Sci. Transl. Med.* 9, eaai7471. <https://doi.org/10.1126/scitranslmed.aai7471>.
- Brichta, L., Hofmann, Y., Hahnen, E., Siebzehrubl, F.A., Raschke, H., Blumcke, I., Eyupoglu, I.Y., and 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. <https://doi.org/10.1093/hmg/ddg256>.
- Chambers, I., Silva, J., Colby, D., Nichols, J., Nijmeijer, B., Robertson, M., Vrana, J., Jones, K., Grotewold, L., and Smith, A. (2007). Nanog safeguards pluripotency and mediates germline development. *Nature* 450, 1230–1234. <https://doi.org/10.1038/nature06403>.
- Chambers, I., and Tomlinson, S.R. (2009). The transcriptional foundation of pluripotency. *Development* 136, 2311–2322. <https://doi.org/10.1242/dev.024398>.
- Dachs, E., Piedrafita, L., Hereu, M., Esquerda, J.E., and 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. <https://doi.org/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., et al. (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. <https://doi.org/10.1016/j.ebiom.2016.05.032>.
- Elshafay, A., Hieu, T.H., Doheim, M.F., Kassem, M.A.M., ELdoada, M.F., Holloway, S.K., Abo-Elghar, H., Hirayama, K., and Huy, N.T. (2019). Efficacy and safety of valproic acid for spinal muscular atrophy: a systematic Review and meta-analysis. *CNS Drugs* 33, 239–250. <https://doi.org/10.1007/s40263-019-00606-6>.
- Galvan, L., Francelle, L., Gaillard, M.C., de Longprez, L., Carrillo-de Sauvage, M.A., Liot, G., Cambon, K., Stimmer, L., Luccantoni, S., Flament, J., et al. (2018). The striatal kinase DCLK3 produces neuroprotection against mutant huntingtin. *Brain* 141, 1434–1454. <https://doi.org/10.1093/brain/awy057>.
- Ghosh, A., and Som, A. (2020). RNA-Seq analysis reveals pluripotency-associated genes and their interaction networks in human embryonic stem cells. *Comput. Biol. Chem.* 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., and Hengstler, J.G. (2018). Assessment of stem cell differentiation based on genome-wide expression profiles. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 373, 20170221. <https://doi.org/10.1098/rstb.2017.0221>.
- Guenou, H., Nissan, X., Larcher, F., Feteira, J., Lemaire, G., Saidani, M., Del Rio, M., Barrault, C.C., Bernard, F.X., Peschanski, M., et al. (2009). Human embryonic stem-cell derivatives for full reconstruction of the pluristratified epidermis: a preclinical study. *Lancet* 374, 1745–1753.
- Harhour, K., Navarro, C., Depetris, D., Mattei, M.G., Nissan, X., Cau, P., De Sandre-Giovannoli, A., and Levy, N. (2017). MG132-induced progerin clearance is mediated by autophagy activation and splicing regulation. *EMBO Mol. Med.* 9, 1294–1313. <https://doi.org/10.15252/emmm.201607315>.
- Hoch, L., Henriques, S.F., Bruge, C., Marsolier, J., Benabides, M., Bourg, N., Tournois, J., Mahe, G., Morizur, L., Jarrige, M., et al. (2019). Identification of thiothrepton as a pharmacological approach to

rescue misfolded alpha-sarcoglycan mutant proteins from degradation. *Sci. Rep.* 9, 6915. <https://doi.org/10.1038/s41598-019-43399-w>.

Hong, S.H., Lee, J.H., Lee, J.B., Ji, J., and 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. <https://doi.org/10.1242/jcs.077511>.

Howlin, J., Cirenajwis, H., Lettiero, B., Staaf, J., Lauss, M., Saal, L., Borg, A., Gruvberger-Saal, S., and 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. <https://doi.org/10.7717/peerj.788>.

Jeanquartier, F., Jean-Quartier, C., and Holzinger, A. (2015). Integrated web visualizations for protein-protein interaction databases. *BMC Bioinformatics* 16, 195. <https://doi.org/10.1186/s12859-015-0615-z>.

Kusters-Vandeveld, H.V., Klaasen, A., Kusters, B., Groenen, P.J., van Engen-van Grunsven, I.A., van Dijk, M.R., Reifemberger, G., Wesseling, P., and Blok, W.A. (2010). Activating mutations of the GNAQ gene: a frequent event in primary melanocytic neoplasms of the central nervous system. *Acta Neuropathologica* 119, 317–323. <https://doi.org/10.1007/s00401-009-0611-3>.

Kwon, D.Y., Motley, W.W., Fischbeck, K.H., and Burnett, B.G. (2011). Increasing expression and decreasing degradation of SMN ameliorate the spinal muscular atrophy phenotype in mice. *Mol. Genet.* 20, 3667–3677. <https://doi.org/10.1093/hmg/ddr288>.

Lefebvre, S., Burglen, L., Reboullet, S., Clermont, O., Burlet, P., Violette, 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, D.K., and 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. <https://doi.org/10.1186/s12864-015-2270-1>.

Locatelli, D., Terao, M., Kurosaki, M., Zanellati, M.C., Pletto, D.R., Finardi, A., Colciaghi, F., Garattini, E., and Battaglia, G.S. (2015). Different stability and proteasome-mediated degradation rate of SMN protein isoforms. *PLoS One* 10, e0134163. <https://doi.org/10.1371/journal.pone.0134163>.

Mair, B., Tomic, J., Masud, S.N., Tonge, P., Weiss, A., Usaj, M., Tong, A.H.Y., Kwan, J.J., Brown, K.R., Titus, E., et al. (2019). Essential gene profiles for human pluripotent stem cells identify uncharacterized genes and substrate dependencies. *Cell Rep.* 27, 599–615 e512. <https://doi.org/10.1016/j.celrep.2019.02.041>.

Mallon, B.S., Chenoweth, J.G., Johnson, K.R., Hamilton, R.S., Tesar, P.J., Yavatkar, A.S., Tyson, L.J., Park, K., Chen, K.G., Fann, Y.C., and McKay, R.D. (2013). StemCellDB: the human pluripotent stem cell database at the National Institutes of Health. *Stem Cell Res.* 10, 57–66. <https://doi.org/10.1016/j.scr.2012.09.002>.

Matera, I., Watkins-Chow, D.E., Loftus, S.K., Hou, L., Incao, A., Silver, D.L., Rivas, C., Elliott, E.C., Baxter, L.L., and Pavan, W.J. (2008). A sensitized mutagenesis screen identifies Gli3 as a modifier of Sox10 neurocristopathy. *Hum. Mol. Genet.* 17, 2118–2131. <https://doi.org/10.1093/hmg/ddn110>.

Maury, Y., Come, J., Piskowski, R.A., Salah-Mohellibi, N., Chevalyere, V., Peschanski, M., Martinat, C., and 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, J.B., Kornobis, E., Polvêche, H., Jarrige, M., Gosselin, M., Manousopoulou, A., Garbis, S.D., Gorecki, D., and 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, S.S., Chaubal, V.A., Shioda, T., Coser, K.R., and 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. <https://doi.org/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., and Perrier, A.L. (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, P.J., Lee, T.R., and Cho, E.G. (2015). Substance P stimulates endothelin 1 secretion via endothelin-converting enzyme 1 and promotes melanogenesis in human melanocytes. *J. Invest. Dermatol.* 135, 551–559. <https://doi.org/10.1038/jid.2014.423>.

Piepers, S., Cobben, J.M., Sooda, P., Jansen, M.D., Wadman, R.I., Meester-Delver, A., Poll-The,

B.T., Lemmink, H.H., Wokke, J.H., van der Pol, W.L., and van den Berg, L.H. (2011). Quantification of SMN protein in leucocytes from spinal muscular atrophy patients: effects of treatment with valproic acid. *J. Neurol. Neurosurg. Psychiatry* 82, 850–852. <https://doi.org/10.1136/jnnp.2009.200253>.

Pinto, J.P., Machado, R.S.R., Magno, R., Oliveira, D.V., Machado, S., Andrade, R.P., Braganca, J., Duarte, I., and Futschik, M.E. (2018). StemMapper: a curated gene expression database for stem cell lineage analysis. *Nucleic Acids Res.* 46, D788–D793. <https://doi.org/10.1093/nar/gkx921>.

Rowe, R.G., and Daley, G.Q. (2019). Induced pluripotent stem cells in disease modelling and drug discovery. *Nat. Rev. Genet.* 20, 377–388. <https://doi.org/10.1038/s41576-019-0100-z>.

Schmidt, B., and Hildebrandt, A. (2017). Next-generation sequencing: big data meets high performance computing. *Drug Discov. Today* 22, 712–717. <https://doi.org/10.1016/j.drudis.2017.01.014>.

Silva, J., Nichols, J., Theunissen, T.W., Guo, G., van Oosten, A.L., Barrandon, O., Wray, J., Yamanaka, S., Chambers, I., and Smith, A. (2009). Nanog is the gateway to the pluripotent ground state. *Cell* 138, 722–737. <https://doi.org/10.1016/j.cell.2009.07.039>.

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

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

Thomson, J.A., Itskovitz-Eldor, J., Shapiro, S.S., Waknitz, M.A., Swiergiel, J.J., Marshall, V.S., and Jones, J.M. (1998). Embryonic stem cell lines derived from human blastocysts. *Science* 282, 1145–1147.

Van Raamsdonk, C.D., Fitch, K.R., Fuchs, H., de Angelis, M.H., and Barsh, G.S. (2004). Effects of G-protein mutations on skin color. *Nat. Genet.* 36, 961–968. <https://doi.org/10.1038/ng1412>.

Yates, B., Braschi, B., Gray, K.A., Seal, R.L., Tweedie, S., and Bruford, E.A. (2017). Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic Acids Res.* 45, D619–D625. <https://doi.org/10.1093/nar/gkw1033>.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Healthy hiPSC	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Healthy hESC	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Healthy hESC-derived Melanocyte	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Healthy hiPSC-derived Motoneuron	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Healthy hiPSC-derived NSC	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Wolfram Syndrome hiPSC-derived NSC	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Healthy hiPSC-derived Neuron	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Wolfram Syndrome hiPSC-derived Neuron	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Healthy hESC-hiPSC-derived MSC	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Hutchinson-Gilford Syndrome hESC-hiPSC-derived MSC	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Hungtington Disease hiPSC-derived RPE	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Healthy hiPSC-derived Mesoderm precursor	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Duchenne Muscular Dystrophy hiPSC	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Duchenne Muscular Dystrophy hiPSC-derived Mesoderm precursor	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Healthy Primary Myoblast	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php

(Continued on next page)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Healthy Primary Myoblast-derived Myotube	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Duchenne Muscular Dystrophy Primary Myoblast	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Duchenne Muscular Dystrophy Primary Myoblast-derived Myotube	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Healthy hESC-derived RPE	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Healthy hiPSC-derived RPE	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Hungtington Disease hiPSC-derived Cortical Neurons	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Healthy hiPSC-derived Cortical Neurons	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Healthy hESC-derived striatal neurons	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Healthy Fibroblasts	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Healthy hESC-derived Cortical Neuron	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Healthy hiPSC-derived striatal neurons	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Healthy hESC-derived Neuron	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Healthy hESC-derived dopaminergic Neuron	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Healthy hiPSC-derived dopaminergic Neuron	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Lesch Nyhan Disease hiPSC-derived dopaminergic Neuron	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Healthy hiPSC-derived dopaminergic Neuron progenitor	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Lesch Nyhan Disease hiPSC-derived dopaminergic Neuron progenitor	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php

(Continued on next page)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Healthy hESC-derived LGE-like progenitors	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Healthy hESC-derived MGE-like progenitors	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
NSC hESC-derived Telencephalic Rosette	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Healthy hESC-derived Cortex progenitors	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Healthy Adult Human Whole Brain tissus	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Healthy hiPSC-derived Myotube	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Healthy hESC-derived neural progenitor	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Spinal Muscular Atrophy hiPSC-derived Motoneuron	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Myotonic Dystrophy type I hiPSC-derived Myotube	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Hungtington Disease hiPSC-derived striatal neurons	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Healthy hiPSC-derived Keratinocyte	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Epidermolysis bullosa simplex hiPSC-derived Keratinocyte	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php

Critical commercial assays

RNeasy Plus Micro Kit	Qiagen	Cat. No. / ID: 74034
RNeasy Plus Mini Kit	Qiagen	Cat. No. / ID:74134
Ion PI™ Hi-Q™ Sequencing 200 Kit	Thermo Fisher Scientific	A26433
Ion AmpliSeq™ Transcriptome Human Gene Expression Kit	Thermo Fisher Scientific	A26325
Ion PI™ Hi-Q™ OT2 200 Kit	Thermo Fisher Scientific	A26434
Ion PI™ Chip Kit v3	Thermo Fisher Scientific	A26771

Deposited data

Analyzed data	This paper	http://sistema.ens-lyon.fr/index.php
Raw and analyzed data	Galvan et al., 2018 https://doi.org/10.1093/brain/awy057.	GEO: GSE104091

(Continued on next page)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Raw and analyzed data	Hoch et al., 2019 https://doi.org/10.1038/s41598-019-43399-w	GEO: GSE119841
Human reference genome NCBI build 38, GRCh38.91	Genome Reference Consortium	http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/
NCBI	National Library of Medicine (US), National Center for Biotechnology Information; [1988	https://www.ncbi.nlm.nih.gov/
STRINGDB	Jean Quartier F, 2015	https://string-db.org/
OMIM	https://omim.org/	https://omim.org/
the HGNC and VGNC resources	Genenames.org: the HGNC and VGNC resources in 2017	https://www.genenames.org/

Experimental models: Cell lines

Human RC17	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	https://hpscereg.eu/cell-line/RCe021-A
Human RC9	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	https://hpscereg.eu/cell-line/RCe013-A
Human SA001	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	https://hpscereg.eu/cell-line/CEBe033-A
Human 14_05	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human 14-27	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human 56c02	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human 1869	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human i90cl16	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human PB12c13	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human WS1	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human WS2	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human WS5	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human delta60	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php

(Continued on next page)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Human HD1	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human M398	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human M418	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human M194	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human M197	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human M180	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human M313	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human V1024	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human M202	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human i90cl17	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human 1869c07	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human PB12c03	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human 23784c05	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human 2227c02	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human 2852c03	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human 2851c15	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human 4603c45	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php

(Continued on next page)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Human VUB01	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human 232-7	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human 13_02	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human 13_01	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human 13_11	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human 14-27	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human 232-11	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human 61c7	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human PC085	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human PC1432	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human PC15	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human 9991	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human 10295	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php
Human NAMAD	Institut for Stem cell Therapy and Exploration of Monogenic diseases (I-STEM)	http://sistema.ens-lyon.fr/index.php

Software and algorithms

FastQC (v0.11.2)	Andrews S. (2010)	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Prinseq (v0.20.4) (–trim-right 20) and filtered by average quality score (–trim-quality 20)	Schmider R, Edwards R (2011)	http://prinseq.sourceforge.net/
Bowtie2 local (v2.3.4.1)	Langmead B, Salzberg S	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml

(Continued on next page)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Samtools (v0.1.19) (Reads below a mapping score of 10 or multi-mapped were filtered)	Li H, Handsaker B (2009)	http://www.htslib.org/
R (v3.6.0)	https://www.r-project.org/	https://www.r-project.org/
Programming language Php	https://www.php.net/	https://www.php.net/
JavaScript	https://developer.mozilla.org/fr/docs/Web/JavaScript	https://developer.mozilla.org/fr/docs/Web/JavaScript
Highcharts	www.highcharts.com	www.highcharts.com
DESeq2	Love MI,2014	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
MySQL	www.mysql.com	www.mysql.com
Other		
Resource website for the publication	This paper	http://sistema.ens-lyon.fr/index.php

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 (GSE104091 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:

- SA001 (hPSCreg Name: CEBE033-A, alternative name SA00), sex Male (XY), <https://hpscereg.eu/cell-line/CEBE033-A>
- RC9 (hPSCreg Name: RCe013-A, alternative name RC9), sex Male (XY), <https://hpscereg.eu/cell-line/RCe013-A>

- RC17 (hPSCreg Name: RCe021-A , alternative name RC17), sex Female(XX), <https://hpscereg.eu/cell-line/RCe021-A>

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 S1](#). 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 of the quantity of cells. The quality of the RNA was evaluated using the Bio-analyseur 2100 Agilent[™] 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 modeling 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 ($\log_{10}(x+1)$) 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 IMG/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, 3C, 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 ($\log_{10}(x+1)$) using R (v3.6.0). For [Figure 5](#), data have been extracted from SISTEMA and are represented as the mean of normalized count \pm 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>