Stem Cell Reports
Resource
ISSCR
OPEN ACCESS

Automatic identification of small molecules that promote cell conversion and reprogramming

Francesco Napolitano,[1,2,9] Trisevgeni Rapakoulia,[2,3,9] Patrizia Annunziata,[4] Akira Hasegawa,[5] Melissa Cardon,[5] Sara Napolitano,[1] Lorenzo Vaccaro,[4] Antonella Iuliano,[1] Luca Giorgio Wanderlingh,[1] Takeya Kasukawa,[5] Diego L. Medina,[1,6] Davide Cacchiarelli,[4,6,*] Xin Gao,[2,*] Diego di Bernardo,[1,7,*] and Erik Arner[5,8,*]

[1]Telethon Institute of Genetics and Medicine (TIGEM), Pozzuoli (NA) 80078, Italy
[2]Computational Bioscience Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia
[3]Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73, 14195 Berlin, Germany
[4]Telethon Institute of Genetics and Medicine (TIGEM), Armenise/Harvard Laboratory of Integrative Genomics, Pozzuoli (NA) 80078, Italy
[5]RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa 230-0045 Japan
[6]Department of Translational Medicine, University of Naples Federico II, Naples, Italy
[7]Department of Chemical, Materials and Industrial Production Engineering, University of Naples Federico II, 80125 Naples, Italy
[8]Graduate School of Integrated Sciences for Life, Hiroshima University, Kagamiyama, Higashi-Hiroshima, 739-8528 Japan
[9]These authors contributed equally
*Correspondence: d.cacchiarelli@tigem.it (D.C.), xin.gao@kaust.edu.sa (X.G.), dibernardo@tigem.it (D.d.B.), erik.arner@riken.jp (E.A.)
https://doi.org/10.1016/j.stemcr.2021.03.028

SUMMARY

Controlling cell fate has great potential for regenerative medicine, drug discovery, and basic research. Although transcription factors are able to promote cell reprogramming and transdifferentiation, methods based on their upregulation often show low efficiency. Small molecules that can facilitate conversion between cell types can ameliorate this problem working through safe, rapid, and reversible mechanisms. Here, we present DECCODE, an unbiased computational method for identification of such molecules based on transcriptional data. DECCODE matches a large collection of drug-induced profiles for drug treatments against a large dataset of primary cell transcriptional profiles to identify drugs that either alone or in combination enhance cell reprogramming and cell conversion. Extensive validation in the context of human induced pluripotent stem cells shows that DECCODE is able to prioritize drugs and drug combinations enhancing cell reprogramming. We also provide predictions for cell conversion with single drugs and drug combinations for 145 different cell types.

INTRODUCTION

Controlling cell fate has enormous potentials for regenerative medicine (Cohen and Melton, 2011), drug discovery (Avior et al., 2016), and cell-based therapy (Kikuchi et al., 2017). A milestone discovery by Yamanaka and colleagues, who induced human stem cells via genetic reprogramming of mature somatic cells using four transcription factors (TFs) (Takahashi and Yamanaka, 2006) (Takahashi et al., 2007), has recently revolutionized the field of stem cell biology. To date, numerous studies have revealed distinct sets of TFs that achieve or promote cell reprogramming (Soufi et al., 2015) and transdifferentiation (Rosa et al., 2018; Sekiya and Suzuki, 2011). However, these methods often suffer from low efficacy due to partly unknown barriers that need to be overcome for complete conversion (Smith et al., 2016).

Optimizing the reprogramming system using non-invasive approaches, such as small-molecule treatment is a promising strategy that may increase the reprogramming potential. The cellular effects of small-molecule treatment are often rapid, dose dependent, and reversible (Zhang et al., 2012), and have potential for *in situ* regeneration therapeutic interventions (Biswas and Jiang, 2016). Recently, several methods relying fully or partially on drug treatment to enhance cell conversion have emerged (Federation et al., 2014). Many of these use fibroblasts as the starting cell type, reprogrammed toward pluripotency (Hou et al., 2013; Zhu et al., 2010) or transdifferentiated to specialized cell types, including neurons (Ladewig et al., 2012), endothelial cells (Sayed et al., 2015), pancreatic like cells (Zhu et al., 2016), cardiomyocytes (Cao et al., 2016), hepatocytes (Lim et al., 2016), or other cell types (Cheng et al., 2015; Li et al., 2017; Wang et al., 2016). Such studies provide a proof of principle for drug-based reprogramming, the exact mechanisms of which, however, are often poorly understood, making extensive trial-and-error unavoidable. Indeed, methods for identifying small-molecule candidates include either exhaustive screenings of drug libraries followed by marker gene readout (Li et al., 2013) or application of drugs known to modulate specific pathways involved in the desired lineage commitment (Federation et al., 2014). While these methods are promising, they are laborious and do not scale.

Whereas computational approaches to identify novel combinations of TFs to facilitate cell reprogramming have been developed and validated (Cahan et al., 2014; Rackham et al., 2016), no similar tools exist for small molecules.

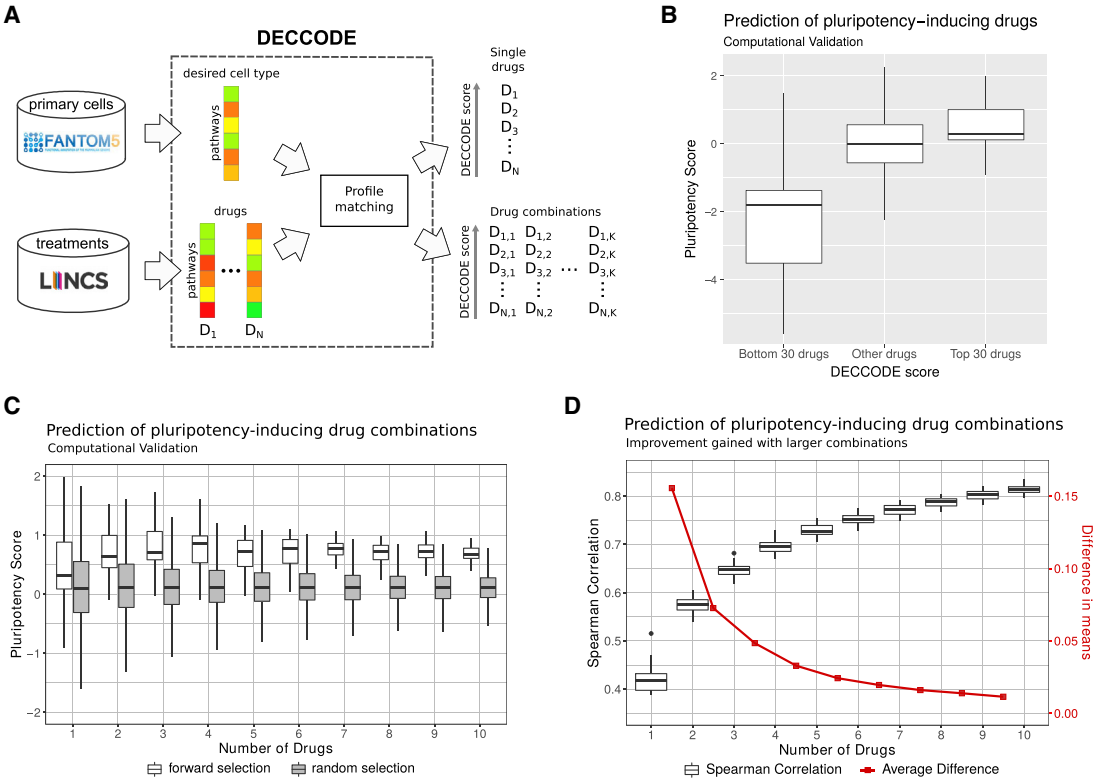


Figure 1. Computational identification of drugs facilitating cell conversion

(A) Workflow of the DECCODE approach. Target cell profiles are constructed from the FANTOM5 collection of human primary cell samples. Drug-induced consensus profiles are created for each of the treated cell lines included in the LINCS database. Single drugs or drug combinations are then prioritized based on their similarity with the target cell-type profile.

(B) *In silico* validation of single drugs facilitating conversion to hiPSCs. Drugs are grouped by their DECCODE scores and the Pluripotency scores (PSs) of the drug-induced gene expression profiles within each group are computed and represented as a boxplot.

(C) *In silico* validation of drug combinations of increasing size facilitating conversion to hiPSCs. PSs for drugs within each of the top 30 combinations are compared with random sets of the same size (see supplemental experimental procedure for the details). Random selection was repeated 100 times.

(D) Improvement obtained when using drug combinations of increasing size. The Spearman correlation between predicted drug combination profiles and hiPSC profile as more drugs are added is reported in the boxplot. The red line highlights the difference between the means of subsequent sets.

Here, we present a methodology to automatically identify small molecules that either alone or in combination enhance cell reprogramming and cell conversion. We analyzed 447 genome-wide expression profiles of untreated primary cells from the FANTOM5 project (Forrest et al., 2014) together with 107,404 transcriptional responses to small-molecule treatment from the LINCS project (Keenan et al., 2018) to identify small molecules that drive the cell transcriptional program toward the one of the desired lineage. We make the results available in an online tool named DECCODE (Drug Enhanced Cell COnversion using Differential Expression), that, when queried, returns the top compounds predicted to enhance conversion toward the desired cell type. We extensively validated DECCODE to identify single or combined small molecules enhancing reprogramming of human fibroblasts toward human induced pluripotent stem cells (hiPSCs). DECCODE is unbiased, as it does not rely on previous knowledge, and it can scale up to identify drugs to enhance cell conversion to any desired cell type. We make the results available in an online tool (available at the following: https://fantom.gsc.riken.jp/5/cellconv/), which, when queried, returns the top compounds predicted to enhance conversion toward a large collection of primary cell types.

RESULTS

The DECCODE approach

A schematic representation of our approach is illustrated in Figure 1A. Given a target cell type, we constructed its

cell-type-specific differential gene expression profile using the FANTOM5 database (Noguchi et al., 2017), the most extensive atlas of gene expression profiles across primary human cells (Forrest et al., 2014), thus obtaining 447 expression profiles corresponding to 145 different cell types (see the supplementary methods). Specifically, the target cell-type expression profile was compared against the profiles of all the remaining cell types to detect differentially expressed genes specific to the target cell type. We then compared the target cell-type profile with drug-induced transcriptional profiles obtained from the LINCS database (GEO: GSE70138). LINCS contains 107,404 differential gene expression profiles corresponding to the transcriptional responses of 41 cell lines to 1,768 different drugs spanning different concentrations and time points, far exceeding any other publicly available resource of cellular perturbations (Keenan et al., 2018).

Since FANTOM5 and LINCS use different expression profiling technologies, we first converted differential gene expression profiles in both datasets to differential pathway-based expression profiles (DPEPs) (Napolitano et al., 2018) (supplemental experimental procedure) to enable an integrative analysis over the two datasets. Subsequently, we generated a consensus profile for each drug by merging together DPEPs across different time points and dosages. Finally, given a cell type of interest, we searched among the 1,768 drugs that induce a transcriptional response similar to the expression profile of the target cell type. The underlying hypothesis is that the selected drugs will induce a change in gene expression in the starting cell type by making it more transcriptionally similar to the target cell type, and thus facilitating the cell conversion process.

We also developed an extension of this method to predict drug combinations that synergize to enhance cell conversion. In previous work, we showed that combinatorial drug treatment is effectively described by a linear combination of the individual drug responses (Rapakoulia et al., 2017) at the transcriptional level. The same finding has also been proven at the protein level, where protein dynamics in drug combinations can be explained by a linear superposition of their responses to individual drugs (Geva-Zatorsky et al., 2010). After confirming that the linear relationship also holds at the pathway level (supplemental experimental procedure), we used a multivariable linear regression model to describe the combined effect of drug combinations. First, for each drug, we selected the profile having the highest DECCODE score across the treated cell lines, thus obtaining a single profile for each drug. Then we used forward selection to pick out the drug subsets yielding the most significant correlation with the target cell profile (supplementary methods).

Application of DECCODE to hiPSC conversion

We first applied DECCODE in the single-drug mode to identify drugs enhancing cell reprogramming to hiPSCs. We thus selected hiPSCs as the target cell type and DECCODE returned the list of all 1,768 drugs ranked according to their predicted efficacy in enhancing cell reprogramming. We performed Drug Set Enrichment Analysis (Napolitano et al., 2016) of the first 25 drugs in the ranking to identify those pathways that are consistently modulated by most of the drugs. As a result, we observed a consistent enrichment of pathways associated with pluripotency, such as differentiation and proliferation (Table S1).

To further assess the validity of the DECCODE score, we devised an *in silico* validation method based on assigning a Pluripotency score (PS) to each drug according to the upregulation of pluripotency-specific genes and downregulation of somatic-specific genes (supplemental experimental procedure). We then compared the DECCODE scores with the PSs. A clear trend can be observed with top-ranked (higher DECCODE scores) drugs exhibiting higher PSs, and bottom-ranked (lower DECCODE scores) drugs exhibiting lower PSs, whereas no obvious correlation existed in the middle-ranked profiles (Figure 1B). The full distribution of the DECCODE scores is reported in Figure S1A.

We then applied DECCODE in the drug combination mode to identify drugs that can jointly enhance reprogramming to hiPSCs. PS and predicted similarity to the hiPSC profile of the top 30 drug combinations significantly improved when increasing the number of drugs, as assessed by Spearman correlation and adjusted R^2 values (Figures 1C, 1D, and S2A). However, we observed that the most significant improvement was achieved when adding just one additional drug and gradually decreased as we kept adding more drugs, eventually reaching a plateau. Akaike information criterion further confirmed that the relative goodness of fit increased more than what would be expected by chance as more drugs were added to the single-drug models (Figure S2B). The distribution of the transcriptional similarities between the two drug profiles in each of the top 30 drug pair combinations was compared with randomly chosen drug pairs (Figure S2C). The results indicated that the two selected drugs in each combination tend to be transcriptionally different. Taken together, these *in silico* results suggest that drug combinations may offer an increased capacity to promote reprogramming compared with single-drug administration.

Experimental validation of DECCODE for conversion to hiPSCs

We set out to experimentally validate predictions of DECCODE in both single-drug and drug combination mode applied to the hiPSCs reprogramming problem. As a biological model of reprogramming, we used human secondary

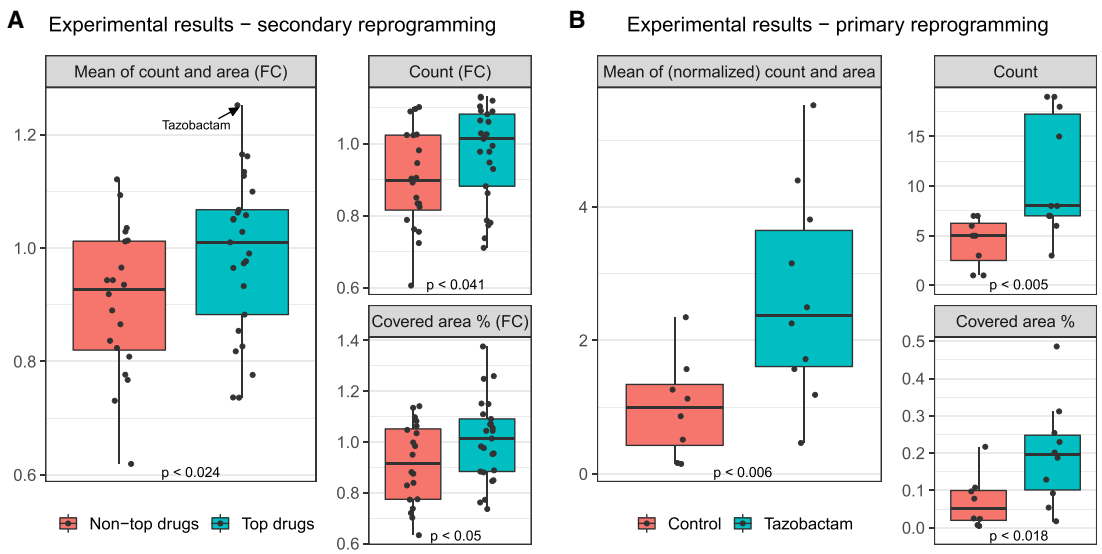


Figure 2. Experimental validation of DECCODE to identify drugs that enhance conversion to hiPSCs

(A) Secondary reprogramming: fold change (FC) relative to controls of the number of colonies and their total area following treatment either with the 25 drugs ranked by DECCODE at the top of the ranking (green boxes) or with 20 drugs ranked in the bottom half of the ranking (red boxes). Dots represent the effect of individual drugs in terms of the average FC for three replicate experiments against controls. Main panel shows the combined average of both the number of colonies (count) and their area expressed as FCs; smaller panels report counts and areas separately.

(B) Primary reprogramming: number of colonies and percentage of their total area following treatment with tazobactam and OSKM compared with OSKM alone (control). Dots represent the single values for each replicate.

All p-values were obtained through single-tailed Mann–Whitney U test.

fibroblasts harboring a doxycycline-inducible OSKM (OCT4, SOX2, KLF4, C-MYC) gene cassette (hiF-T cells) (Cacchiarelli et al., 2015). For the single-drug case, we selected for experimental validation the top-ranked 25 drugs present in either of 2 widely used chemical screening libraries (supplemental experimental procedure) that were predicted by DECCODE to enhance the pluripotency transcriptional program. In addition, we selected 20 drugs with scores in the lower half of the total ranking for comparison. hiF-T cells were treated with a total of 45 drugs in triplicate. After 21 days of treatment, cells were stained for the pluripotency marker TRA-1-60, imaged, and cell colony number and area were quantified for each well (see example in Figure S1D). Increases in efficacy were difficult to detect by visual inspection due to the already highly efficient secondary reprogramming system used in the screening; however, the top-ranked drugs performed significantly better than the lower-ranked drugs, either when considering the number of colonies or the total area covered by the colonies (Figure 2A). Although some drugs with predicted low efficacy performed well in the screen, when ranked, the 45 tested drugs significantly correlated with the observed efficacy based on colony count and covered areas combined (Spearman $\rho = 0.32$, $p = 0.033$, see Figure S1B). Several top-ranked drugs have been already associated with enhancement of reprogramming (Table S2), including tranylcypromine, which we previously identified as a new positive regulator of the reprogramming process (Cacchiarelli et al., 2015). Another set of experimentally validated small molecules relevant for hiPSC generation reported in (Chen et al., 2020) (Table S3), which we analyzed as a whole, was also ranked significantly high by DECCODE (Kolmogorov-Smirnov $D = 0.33$, $p = 9.36^{-3}$, see Figure S1E).

Tazobactam, an antibiotic of the beta-lactamase inhibitor class previously unexamined in the context of cell reprogramming, achieved the highest performance when considering the area covered by the colonies and the second highest performance when considering the number of colonies (Figures 2A and S1C), thus ranking first when considering both area and colony number together. Tazobactam was further validated by performing primary reprogramming of human primary foreskin fibroblasts through OSKM transduction either in the presence or absence of tazobactam. Both the number of colonies and the total area covered by the colonies confirmed the ability of tazobactam to enhance reprogramming to hiPSCs (Figure 2B).

For validation of the drug combinations, we focused on the best drug pairs as ranked by DECCODE. In particular, from the ranked list of drug pairs, we chose the top eight combinations, including a sufficiently variant set of drugs

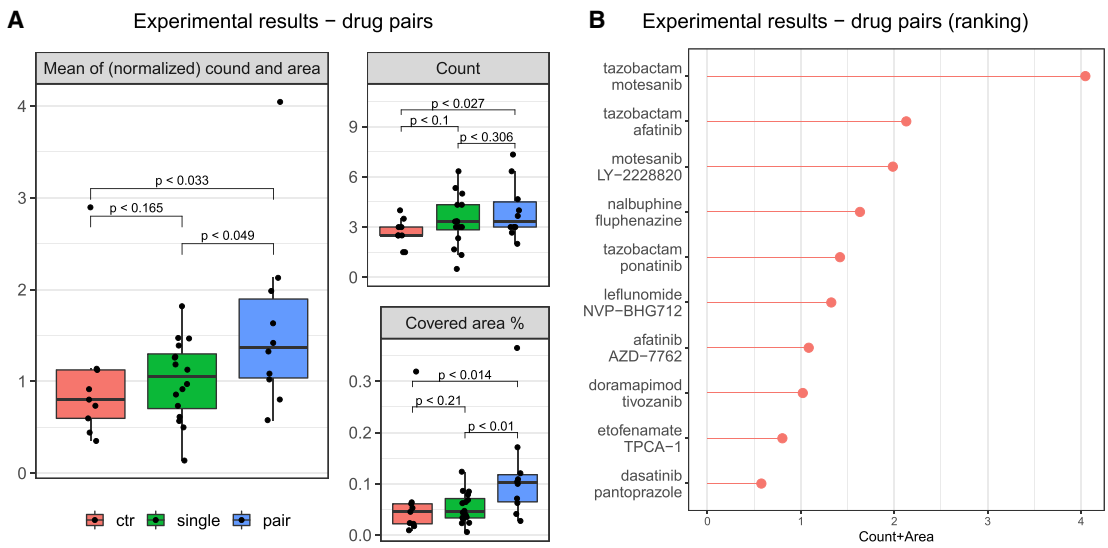


Figure 3. Experimental validation of DECCODE to identify drug combinations that enhance conversion to hiPSCs

(A) Number of colonies and total area obtained following treatment with the top 10 drug pairs as ranked by DECCODE (blue boxes), with each of the 16 drugs included in the 10 pairs (green boxes), and without drug treatment (red boxes). Dots represent the effects of individual samples for controls or average over triplicates for treatments. The main panel shows the combined outcome of both the number of colonies and area normalized against the respective control average; smaller panels report counts and areas separately.

(B) Reprogramming efficacy (colonies count and area) of the drug combinations tested.

All p-values were obtained through single-tailed Mann–Whitney U test.

and excluding those already proven not to be effective in single-drug experiments (supplemental experimental procedure). Finally, we experimentally tested the top eight pairs plus two additional top-ranked drug combinations which included tazobactam. The 10 final pairs included 16 different drugs, which were tested individually and in combinations using the same experimental setting as described previously. As shown in Figure 3A, the experimental results demonstrate a clear trend of increasing reprogramming efficacy from the untreated drugs to the single-drug treatments and from the single-drug treatments to the drug pair treatments. The two best performing drug pairs included tazobactam (Figure 3B), highlighting again the efficiency of the drug in reprogramming. The top pair, which is a combination of tazobactam and motesanib, showed a 4-fold performance improvement as compared with untreated cells.

To create a comprehensive resource for drug-assisted cell conversions, we applied DECCODE to the whole FANTOM5 set of primary cells. We observed that closely related cell types exhibit high transcriptional similarity leading to tautological and nonspecific drug predictions. To reduce redundancy, we thus used a two-level hybrid clustering of cell types taking into consideration both knowledge-driven and data-driven similarities (Figure 4). In the first level, we applied the affinity propagation (Frey and Dueck, 2007) algorithm to cluster primary cells using either an ontological similarity (supplemental experimental procedures) or a transcriptional similarity, thus obtaining two different clusterings (Figure 4). For visualization purposes, we also performed a hierarchical clustering for both similarity measures (Figure S3). In the second level, cell types that were grouped together by both ontological and transcriptional clustering, were kept in one cluster, otherwise they were separated into distinct clusters. Finally, DPEPs of cell types in the same cluster were merged together to create a single consensus profile. We thus obtained 69 consensus DPEPs corresponding to distinct "meta-cells," i.e., an ensemble of different cell types very similar in both ontological and transcriptional terms (the 69 meta-cell clusters are reported in Table S4).

We applied DECCODE to the 69 meta-cells profiles and found several drugs experimentally proved to facilitate cell conversions among the top 5% of ranked drugs (Table S5). Of note, 2 small molecules (Y-27632 and PD0325901) that were ranked among the top 20 candidates for conversion into the neuronal cell type were previously experimentally proven to promote neural conversion. Y-27632, a ROCK inhibitor that assists in neuron survival, was used in combination with six other small molecules to convert human fibroblasts into neuronal-like cells (Hu et al., 2015). PD0325901, a MEK-ERK inhibitor, facilitated the direct conversion of somatic cells to induced neuronal cells in a chemical cocktail of six compounds (Dai et al., 2015).

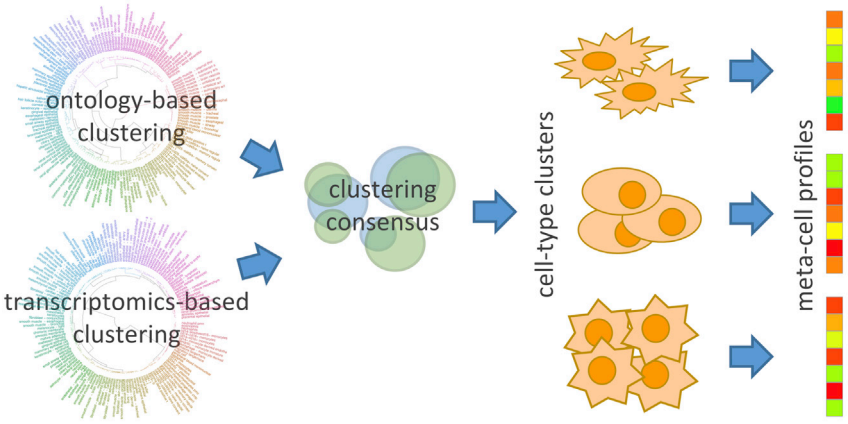


Figure 4. Two-level clustering to obtain cell-type-specific consensus profiles

Our method has been pre-computed on 69 meta-cells representing all the primary cell types included in the FANTOM5 database, and the top single- and multi-drug predictions are publicly available through the DECCODE website (https://fantom.gsc.riken.jp/5/cellconv/) to provide an extensive resource that may support and complement future chemical enhanced cell conversion studies.

DISCUSSION

Therapies based on cell reprogramming and conversion are becoming a reality (Mandai et al., 2017; Stoddard-Bennett and Reijo Pera, 2019) along with the need for methods that improve efficacy and safety of these processes. In the field of reprogramming and in general cell conversion, the genomic integrity and safety of the approach are often overlooked and difficult to globally evaluate. Maximizing the efficiency and speed is not only a means to improve the generation throughput of the cell of interest (particularly important in the case of conversion to non-replicating cells, such as neurons and hepatocytes) but also a means to bypass clonal expansion of subtypes resulting from of possible gain-of-function selections. The use of small molecules, rather than genetic factors, is a promising approach to address these issues. In addition to the ultimate goal of finding combinations that can fully convert one cell type to another, the identification of small molecules that can perform a partial conversion, or make the conversion more effective, also represents an important improvement on current methods. Here, we developed an unbiased method, DECCODE, which does not rely on expert knowledge of lineage-specific genes and pathways, scales to large numbers of cell conversions and drugs, and relies on publicly available data, thus not requiring massive screening efforts. Our method is the first validated computational approach for prioritizing small molecules promoting cell reprogramming. We have applied DECCODE to 145 human primary cell types from FANTOM5 and made the results available, providing a comprehensive resource, including both single drugs and combinations of two drugs predicted to facilitate conversion to a variety of cell types. Together with such resources, we also released the full automated pipeline used for colony quantification, including source code and high-resolution plate scans (Napolitano et al., 2020) (https://doi.org/10.5281/zenodo.3732772).

Although the number of gene expression profiles following drug treatment available in public databases is substantial, the number of unique small molecules profiled is not. Indeed, only a subset of small molecules that have been experimentally validated to facilitate cell conversions, were transcriptionally profiled in LINCS. Moreover, considering that many of the profiled agents are kinase inhibitors, relevant for cancer therapy, publicly available drug profile resources represent a small portion of the "druggable genome" in cell conversion applications. Future profiling efforts that include additional libraries of small molecules will increase the utility of our approach.

Since our method relies on the analysis of transcriptomic data, there are some restrictions regarding the small molecules that can be captured. For example, epigenetic modifiers, such as HDAC inhibitors are extensively used in cell conversions to tackle the epigenetic barriers between different types of cells. The broad action and the nonspecific transcriptional behavior (Chen et al., 2015) of these drugs limits their identifiability by DECCODE. In contrast, our approach gives priority to compounds exhibiting strong transcriptional regulation toward the target cell type. Looking forward, results from our validation experiments could be used in future studies to identify and weigh the most relevant but less transcriptionally evident pathways. On the other hand, it may be advisable to combine compounds identified in this study with treatment with epigenetic modifiers to further increase the efficacy of cell conversion.

Experiments on human fibroblasts confirmed the ability of DECCODE to predict single or combined small molecules facilitating cell reprogramming to iPSCs. Although we see no reason to believe that adding small molecules predicted by DECCODE to the reprogramming protocol would affect pluripotency in a negative way compared with other compounds in the reprogramming cocktails already present, full understanding of the efficacy of DECCODE for conversion to iPSCs, and other cell types, will require deep characterization of converted cells in terms of karyotype, pluripotency, and other relevant parameters. Deeper examination of reprogrammed cells may also reveal additional insights, as our validation experiments showed that increased efficacy was sometimes due to increased colony counts and at other times due to increased colony sizes, which may in turn be due to different drug treatments having different effects on processes, such as proliferation or reprogramming kinetics. The efficiency of reprogramming when treating cells with the best-ranked drugs was increased when compared with the control case and even more when using drug pairs. In the screening of high- and low-ranking drugs, a secondary reprogramming system highly optimized for hiPSC generation was used and, consequently, although statistically significant, the biological effects from drug treatment on colony formation were moderately strong and in many cases hardly visible by eye. Indeed, in the follow-up primary reprogramming experiment with tazobactam, the biological and statistical effects were considerably higher as this assay is less efficient.

The purpose of DECCODE is to provide a completely unbiased, data-driven approach to prioritize small molecules facilitating cell conversion. Alternative methods specialized for the case of specific cell types may also have utility. We used a knowledge-based approach, the PS, as an indication of the DECCODE performance, and indeed the PS itself, or other cell-type-specific transcriptional scoring approaches, may be useful in identifying suitable drug treatments for distinct target lineages. An unbiased method like DECCODE may identify different sets of molecules than more targeted methods. As an example, tazobactam, which proved to be effective in reprogramming, had a high DECCODE score and a low PS and consequently would have been missed by PS prioritization. Although this does not prove that the DECCODE score is superior to the PS in the particular case of iPSC reprogramming, it does provide evidence that DECCODE is complementary to knowledge-based approaches, and thus more generally applicable.

In summary, our work demonstrates that DECCODE is able to distinguish and prioritize small molecules based on their potential to promote reprogramming and its usefulness in facilitating cell conversion should not be underestimated. We identified the core reprogramming chemicals for each lineage commitment, which could aid in establishing the role of various small molecules in different cell fates. We made our results available via a user-friendly interface to facilitate the design of cell conversion experiments involving chemical compounds. Our method provides a significant head start toward the development of systematic chemical-based reprogramming strategies.

EXPERIMENTAL PROCEDURES

Gene expression data

We used untreated primary cell profiles from the FANTOM5 collection and drug-induced gene expression profiles from the LINCS collection. We selected all primary human cells having at least two biological replicates from the FANTOM5 database (http://fantom.gsc.riken.jp/5/data), resulting in 447 samples, which correspond to 145 different cell types. Expression tables of robust CAGE peaks for these samples were processed as follows: we kept only the promoters located within 500 bp of known RefSeq transcripts (87,400 promoters). We added read counts of all the promoters sharing the same Entrez id annotation, resulting in 18,980 genes (Figure S4A). Read counts of samples were converted in counts per million values and averaged across the same cell type. Z score normalization was applied to each gene across cell types to obtain differential expression profiles for each cell type. Subsequently, genes in every cell type were ranked according to their expression, from the most expressed to the least expressed gene.

LINCS database is available as gene-based expression profiles. We downloaded the fifth level of differential gene expression signatures released on the GEO website (GEO: GSE70138), which includes 107,404 profiles corresponding to 1,768 different drugs in 41 cell lines, 83 concentrations, and 4 treatment durations. Data access was performed through the cmapR package (Enache et al., 2019). Genes in each drug profile were ranked according to their expression, from the most upregulated to the most downregulated gene.

Single- and multi-drug DECCODE scores

We converted LINCS and target cell-type PEPs to ranks based on their enrichment score (from the most enriched to the least enriched pathway). Then, we ranked each LINCS PEP by computing its L_1 distance from the target cell-type PEP:

$$D(d_i,\ t) = \sum_{p=1}^{250} |d_{i,p} - t_p|$$

where d_i is the LINCS PEP for drug i, $i = 1, \ldots, 17{,}259$ (number of drug-induced PEPs), t is the target cell type PEP, $p = 1, \ldots, 250$ (number of pathways in C2 collection). We finally converted the distance to the target cell type to a similarity measure:

$$S(d_i,\ t) = 1 - D(d_i,\ t)$$

For further analysis, we considered only the top-ranked profile for each small molecule, resulting in 1,768 profiles (number of unique small molecules). The DECCODE score ranges from 0 to

1, where a score close to 1 signifies a strong predicted similarity to the target profile, whereas a score close to 0 means no predicted similarity. Figure S1F shows the distribution of the top 30 DECCODE drug scores against all the 1,768 DECCODE scores for all meta-cell cluster profiles.

To produce DECCODE scores for drug combinations, we considered only the top-ranked PEP for each drug based on its L_1 distance to the target cell type. For each drug PEP (1,768), we fitted a simple linear regression model and we ranked drugs based on the Spearman correlation between fitted and observed values. We picked the top 30 drug PEPs and searched through the remaining drugs to find out which one should be added to the current models to best improve the Spearman correlation. Repeated occurrences of the same drug sets in different order were discarded. We continued to add variables to the top 30 models until we reached 10 predictors. DECCODE multi-drug score in each step is the Spearman correlation between fitted and target PEPs.

Human cellular reprogramming

All the reprogramming experiments and procedures were performed as described previously (Cacchiarelli et al., 2015). In summary, secondary reprogramming was performed by seeding, on a confluent irradiated mouse embryonic fibroblast (MEF) feeder layer, clonal TERT-immortalized secondary fibroblasts (hiF-T) harboring a doxycycline-inducible OSKM cassette. The day after seeding reprogramming was initiated by doxycycline supplementation and protracted for 21 days. For primary reprogramming, BJ foreskin fibroblasts were infected with a lentivirus harboring the constitutive OSKM cassette (pLM-fSV2A) (Papapetrou et al., 2011) split onto an irradiated MEF feeder layer and reprogrammed for 15 days.

The secondary reprogramming experiment to test drug combinations was performed in sub-optimal conditions to allow easier identification of gain-of-function events. As the split of hiF-T at confluence above 60% deeply impinge the reprogramming efficiency, sub-optimal conditions were obtained by splitting cells only upon semi-confluency before reprogramming them, thus creating a reduction of reprogramming efficiency of at least 10 times, as evident by the reduced colony number. At the end of each reprogramming, quantitative analysis of colony number and area was performed using a TRA-1-60 chromogenic staining in bright field. All candidate drugs for reprogramming were tested for the entire duration of the reprogramming process at a final concentration of 10 nM in several technical or biological replicates, as indicated.

Data and code availability

DECCODE top single- and multi-drug predictions for 69 meta-cells are publicly available through the DECCODE website (https://fantom.gsc.riken.jp/5/cellconv/). The full automated pipeline used for colony quantification, including source code and high-resolution plate scans are available on (doi: 10.5281/zenodo.3732772).

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.stemcr.2021.03.028.

AUTHOR CONTRIBUTIONS

Data analysis was performed by F.N. and T.R. and supervised by X.G., D.d.B., and E.A. Additional analysis was performed by S.N. and A.I. Validation experiments were performed by P.A., L.V., and L.G.W., supervised by D.C. and D.L.M. The database and website were implemented by M.C. and A.H., and supervised by T.K. The manuscript was drafted by T.R., F.N., D.d.B., and E.A. with additional input from all other co-authors. The study was conceived by E.A. and jointly supervised by E.A., D.d.B., X.G., and D.C.

ACKNOWLEDGMENTS

E.A. was supported by a Research Grant from MEXT to the RIKEN Center for Integrative Medical Sciences. X.G. was supported by funding from King Abdullah University of Science and Technology (KAUST), under award number FCC/1/1976-18-01, FCC/1/1976-23-01, FCC/1/1976-25-01, FCC/1/1976-26-01, FCS/1/4102-02-01, REI/1/4216-01-01, and REI/1/4437-01-01. D.L.M. was supported by the Italian Telethon Foundation under project number TMDMCBX16TT. D.C. was supported by Fondazione Telethon Core Grant, Armenise-Harvard Foundation Career Development Award, European Research Council (grant agreement 759154, CellKarma), and the Rita-Levi Montalcini program from MIUR.

Received: May 13, 2020
Revised: March 24, 2021
Accepted: March 25, 2021
Published: April 22, 2021

REFERENCES

Avior, Y., Sagi, I., and Benvenisty, N. (2016). Pluripotent stem cells in disease modelling and drug discovery. Nat. Rev. Mol. Cell Biol. *17*, 170–182.

Biswas, D., and Jiang, P. (2016). Chemically induced reprogramming of somatic cells to pluripotent stem cells and neural cells. Int. J. Mol. Sci. *17*, 226.

Cacchiarelli, D., Trapnell, C., Ziller, M.J., Soumillon, M., Cesana, M., Karnik, R., Donaghey, J., Smith, Z.D., Ratanasirintrawoot, S., Zhang, X., et al. (2015). Integrative analyses of human reprogramming reveal dynamic nature of induced pluripotency. Cell *162*, 412.

Cahan, P., Li, H., Morris, S.A., Lummertz da Rocha, E., Daley, G.Q., and Collins, J.J. (2014). CellNet: network biology applied to stem cell engineering. Cell *158*, 903–915.

Cao, N., Huang, Y., Zheng, J., Spencer, C.I., Zhang, Y., Fu, J.-D., Nie, B., Xie, M., Zhang, M., Wang, H., et al. (2016). Conversion of human fibroblasts into functional cardiomyocytes by small molecules. Science *352*, 1216–1220.

Chen, G., Guo, Y., Li, C., Li, S., and Wan, X. (2020). Small molecules that promote self-renewal of stem cells and somatic cell reprogramming. Stem Cell Rev. Rep. *16*, 511–523.

Chen, H.P., Zhao, Y.T., and Zhao, T.C. (2015). Histone deacetylases and mechanisms of regulation of gene expression. Crit. Rev. Oncog. *20*, 35–47.

Cheng, L., Gao, L., Guan, W., Mao, J., Hu, W., Qiu, B., Zhao, J., Yu, Y., and Pei, G. (2015). Direct conversion of astrocytes into neuronal cells by drug cocktail. Cell Res. *25*, 1269–1272.

Cohen, D.E., and Melton, D. (2011). Turning straw into gold: directing cell fate for regenerative medicine. Nat. Rev. Genet. *12*, 243–252.

Dai, P., Harada, Y., and Takamatsu, T. (2015). Highly efficient direct conversion of human fibroblasts to neuronal cells by chemical compounds. J. Clin. Biochem. Nutr. *56*, 166–170.

Enache, O.M., Lahr, D.L., Natoli, T.E., Litichevskiy, L., Wadden, D., Flynn, C., Gould, J., Asiedu, J.K., Narayan, R., and Subramanian, A. (2019). The GCTx format and cmap{Py, R, M, J} packages: resources for optimized storage and integrated traversal of annotated dense matrices. Bioinformatics *35*, 1427–1429.

Federation, A.J., Bradner, J.E., and Meissner, A. (2014). The use of small molecules in somatic-cell reprogramming. Trends Cell Biol. *24*, 179–187.

Forrest, A.R.R., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, M.J.L., Lassmann, T., Itoh, M., Summers, K.M., Suzuki, H., Daub, C.O., et al. (2014). A promoter-level mammalian expression atlas. Nature *507*, 462–470.

Frey, B.J., and Dueck, D. (2007). Clustering by passing messages between data points. Science *315*, 972–976.

Geva-Zatorsky, N., Dekel, E., Cohen, A.A., Danon, T., Cohen, L., and Alon, U. (2010). Protein dynamics in drug combinations: a linear superposition of individual-drug responses. Cell *140*, 643–651.

Hou, P., Li, Y., Zhang, X., Liu, C., Guan, J., Li, H., Zhao, T., Ye, J., Yang, W., Liu, K., et al. (2013). Pluripotent stem cells induced from mouse somatic cells by small-molecule compounds. Science *341*, 651–654.

Hu, W., Qiu, B., Guan, W., Wang, Q., Wang, M., Li, W., Gao, L., Shen, L., Huang, Y., Xie, G., et al. (2015). Direct conversion of normal and Alzheimer's disease human fibroblasts into neuronal cells by small molecules. Cell Stem Cell *17*, 204–212.

Keenan, A.B., Jenkins, S.L., Jagodnik, K.M., Koplev, S., He, E., Torre, D., Wang, Z., Dohlman, A.B., Silverstein, M.C., Lachmann, A., et al. (2018). The library of integrated network-based cellular signatures NIH program: system-level cataloging of human cells response to perturbations. Cell Syst. *6*, 13–24.

Kikuchi, T., Morizane, A., Doi, D., Magotani, H., Onoe, H., Hayashi, T., Mizuma, H., Takara, S., Takahashi, R., Inoue, H., et al. (2017). Human iPS cell-derived dopaminergic neurons function in a primate Parkinson's disease model. Nature *548*, 592–596.

Ladewig, J., Mertens, J., Kesavan, J., Doerr, J., Poppe, D., Glaue, F., Herms, S., Wernet, P., Kögler, G., Müller, F.-J., et al. (2012). Small molecules enable highly efficient neuronal conversion of human fibroblasts. Nat. Methods *9*, 575–578.

Li, J., Casteels, T., Frogne, T., Ingvorsen, C., Honoré, C., Courtney, M., Huber, K.V.M., Schmitner, N., Kimmel, R.A., Romanov, R.A., et al. (2017). Artemisinins target $GABA_A$ receptor signaling and impair α cell identity. Cell *168*, 86–100.e15.

Li, W., Li, K., Wei, W., and Ding, S. (2013). Chemical approaches to stem cell biology and therapeutics. Cell Stem Cell *13*, 270–283.

Lim, K.T., Lee, S.C., Gao, Y., Kim, K.-P., Song, G., An, S.Y., Adachi, K., Jang, Y.J., Kim, J., Oh, K.-J., et al. (2016). Small molecules facilitate single factor-mediated hepatic reprogramming. Cell Rep. *15*, 814–829.

Mandai, M., Watanabe, A., Kurimoto, Y., Hirami, Y., Morinaga, C., Daimon, T., Fujihara, M., Akimaru, H., Sakai, N., Shibata, Y., et al. (2017). Autologous induced stem-cell-derived retinal cells for macular degeneration. N. Engl. J. Med. *376*, 1038–1046.

Napolitano, F., Sirci, F., Carrella, D., and Di Bernardo, D. (2016). Drug-set enrichment analysis: a novel tool to investigate drug mode of action. Bioinformatics *32*, 235–241.

Napolitano, F., Carrella, D., Mandriani, B., Pisonero-Vaquero, S., Sirci, F., Medina, D.L., Brunetti-Pierri, N., and di Bernardo, D. (2018). gene2drug: a computational tool for pathway-based rational drug repositioning. Bioinformatics *34*, 1498–1505.

Napolitano, F., Rapakoulia, T., Annunziata, P., Hasegawa, A., Cardon, M., Napolitano, S., Vaccaro, L., Iuliano, A., Wanderlingh, L.G., Kasukawa, T., et al. (2020). Automatic Identification of Small Molecules that Promote Cell Conversion and Reprogramming—Plate Scans, Colony Quantification Scripts, and DECCODE Ranking https://doi.org/10.5281/zenodo.3672708.

Noguchi, S., Arakawa, T., Fukuda, S., Furuno, M., Hasegawa, A., Hori, F., Ishikawa-Kato, S., Kaida, K., Kaiho, A., Kanamori-Katayama, M., et al. (2017). FANTOM5 CAGE profiles of human and mouse samples. Sci. Data *4*, 170112.

Papapetrou, E.P., Lee, G., Malani, N., Setty, M., Riviere, I., Tirunagari, L.M.S., Kadota, K., Roth, S.L., Giardina, P., Viale, A., et al. (2011). Genomic safe harbors permit high β-globin transgene expression in thalassemia induced pluripotent stem cells. Nat. Biotechnol. *29*, 73–81.

Rackham, O.J.L., Firas, J., Fang, H., Oates, M.E., Holmes, M.L., Knaupp, A.S., Suzuki, H., Nefzger, C.M., Daub, C.O., Shin, J.W., et al. (2016). A predictive computational framework for direct reprogramming between human cell types. Nat. Genet. *48*, 331–335.

Rapakoulia, T., Gao, X., Huang, Y., De Hoon, M., Okada-Hatakeyama, M., Suzuki, H., and Arner, E. (2017). Genome-scale regression analysis reveals a linear relationship for promoters and enhancers after combinatorial drug treatment. Bioinformatics *33*, 3696–3700.

Rosa, F.F., Pires, C.F., Kurochkin, I., Ferreira, A.G., Gomes, A.M., Palma, L.G., Shaiv, K., Solanas, L., Azenha, C., Papatsenko, D., et al. (2018). Direct reprogramming of fibroblasts into antigen-presenting dendritic cells. Sci. Immunol. *3*, eaau4292.

Sayed, N., Wong, W.T., Ospino, F., Meng, S., Lee, J., Jha, A., Dexheimer, P., Aronow, B.J., and Cooke, J.P. (2015). Transdifferentiation of human fibroblasts to endothelial cells: role of innate immunity. Circulation *131*, 300–309.

Sekiya, S., and Suzuki, A. (2011). Direct conversion of mouse fibroblasts to hepatocyte-like cells by defined factors. Nature *475*, 390–393.

Smith, Z.D., Sindhu, C., and Meissner, A. (2016). Molecular features of cellular reprogramming and development. Nat. Rev. Mol. Cell Biol. *17*, 139–154.

Soufi, A., Garcia, M.F., Jaroszewicz, A., Osman, N., Pellegrini, M., and Zaret, K.S. (2015). Pioneer transcription factors target partial

DNA motifs on nucleosomes to initiate reprogramming. Cell *161*, 555–568.

Stoddard-Bennett, T., and Reijo Pera, R. (2019). Treatment of Parkinson's disease through personalized medicine and induced pluripotent stem cells. Cells *8*, 26.

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell *126*, 663–676.

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.

Wang, Y., Qin, J., Wang, S., Zhang, W., Duan, J., Zhang, J., Wang, X., Yan, F., Chang, M., Liu, X., et al. (2016). Conversion of human gastric epithelial cells to multipotent endodermal progenitors using defined small molecules. Cell Stem Cell *19*, 449–461.

Zhang, Y., Li, W., Laurent, T., and Ding, S. (2012). Small molecules, big roles—the chemical manipulation of stem cell fate and somatic cell reprogramming. J. Cell Sci. *125*, 5609–5620.

Zhu, S., Li, W., Zhou, H., Wei, W., Ambasudhan, R., Lin, T., Kim, J., Zhang, K., and Ding, S. (2010). Reprogramming of human primary somatic cells by OCT4 and chemical compounds. Cell Stem Cell *7*, 651–655.

Zhu, S., Russ, H.A., Wang, X., Zhang, M., Ma, T., Xu, T., Tang, S., Hebrok, M., and Ding, S. (2016). Human pancreatic beta-like cells converted from fibroblasts. Nat. Commun. *7*, 10080.

Stem Cell Reports, Volume *16*

Supplemental Information

Automatic identification of small molecules that promote cell conversion and reprogramming

Francesco Napolitano, Trisevgeni Rapakoulia, Patrizia Annunziata, Akira Hasegawa, Melissa Cardon, Sara Napolitano, Lorenzo Vaccaro, Antonella Iuliano, Luca Giorgio Wanderlingh, Takeya Kasukawa, Diego L. Medina, Davide Cacchiarelli, Xin Gao, Diego di Bernardo, and Erik Arner

Automatic identification of small molecules that promote cell conversion and reprogramming

Francesco Napolitano[1,2]*, Trisevgeni Rapakoulia[2,3]*, Patrizia Annunziata[4], Akira Hasegawa[5], Melissa Cardon[5], Sara Napolitano[1], Lorenzo Vaccaro[4], Antonella Iuliano[1], Luca Giorgio Wanderlingh[1], Takeya Kasukawa[5], Diego L. Medina[1,6], Davide Cacchiarelli[4,6#], Xin Gao[2#], Diego di Bernardo[1,7#], Erik Arner[5,8#]

[1]Telethon Institute of Genetics and Medicine (TIGEM), Pozzuoli (NA) 80078, Italy
[2]Computational Bioscience Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia.
[3]Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73, 14195 Berlin, Germany
[4]Telethon Institute of Genetics and Medicine (TIGEM), Armenise/Harvard Laboratory of Integrative Genomics, Pozzuoli (NA) 80078, Italy.
[5]RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa, 230-0045 Japan
[6]Department of Translational Medicine, University of Naples Federico II, Naples, Italy
[7]Department of Chemical, Materials and Industrial Production Engineering, University of Naples Federico II, 80125 Naples, Italy.
[8]Graduate School of Integrated Sciences for Life, Hiroshima University, Kagamiyama, Higashi-Hiroshima, 739-8528 Japan

[*]Contributed equally to this work.
[#]Correspondence: d.cacchiarelli@tigem.it, xin.gao@kaust.edu.sa, dibernardo@tigem.it, erik.arner@riken.jp

Supplementary Materials

Supplementary Methods

Conversion to pathway-based profiles

To harmonize the two datasets, we converted the ranked lists of genes from both cell-types (FANTOM5) and drug treatments (LINCS) into *pathway-based expression profiles* (PEPs). A PEP is a transcriptomic profile expressed in terms of pathways as opposed to genes. PEPs were introduced in our previous work (Napolitano et al., 2016) and their efficacy for drug discovery applications was also proved (Napolitano et al., 2018). To convert FANTOM5 and LINCS Gene Expression Profiles (GEPs) to PEPs, we applied the gep2pep Bioconductor package (Napolitano et al., 2019) using all the 14,645 gene sets from 16 different gene set collections included in the MsigDB v6.1 (Liberzon et al., 2015). The gep2pep package iteratively performs Gene Set Enrichment Analysis (GSEA)(Subramanian et al., 2005) to compute Enrichment Scores for each gene set and each expression profile. A PEP is then defined as a ranked list of pathways, each of which is associated with an Enrichment Score (and the corresponding p-value). Once FANTOM5 and LINCS GEPs are converted to PEPs, they can be directly compared (**Supplementary Figure 4A**).

Various pathway-based profiles for the same gene expression profile can be obtained based on the chosen pathway database. In our case, as previously mentioned, we tried 16 different pathway collections available at the MSigDB database. We then evaluated which one out of these 16 collections best captured cell-type similarities, with respect to the Cell Ontology (Bard et al., 2005). To this aim, we used the Cell Ontology annotation of cell-types created by the FANTOM5 consortium (Lizio et al., 2015). In order to obtain a numerical score for each pair of cell-types *i* and *j* in the ontology, we used the Jaccard Index as follows:

$$D_{CO}(i,j)=1-\frac{|C_i \cap C_j|}{|C_i|+|C_j|-|C_i \cap C_j|} \text{ (Jaccard index)}$$

where C_i are the ontology ancestors of cell type *i*, C_j are the ontology ancestors of cell type *j*, $1 \leq i, j \leq 145$, $i \neq j$. Then we defined the PEP-based distance between cell types *i* and *j* using the Manhattan distance as follows:

$$D_P(i,j) = |P_i - P_j|$$

P_i is the PEP of cell type *i*, P_j is the PEP of cell type *j*, $1 \leq i, j \leq 145, i \neq j$.

Finally, we compared the cell distances computed on the PEPs with the same cell distances based on the Cell Ontology (**Supplemental Figure 4B**). The PEPs based on the C2 collection (Canonical Pathways) achieved the highest agreement with the ontology-based similarities, capturing more accurately the known cell hierarchy, even when compared to a previously developed gene-based approach (Iorio et al., 2010).Thus, pathway-based profiles obtained with C2 collection, which includes 250 pathways, were chosen for all further analyses.

Merging of Pathway-based Expression Profiles

As previously proposed (Iorio et al., 2010) we merged multiple expression profiles elicited by the same drug treatment in order to obtain a single "consensus-profile" for each drug, thus enhancing drug-specific effects while reducing unrelated ones. The gep2pep package (Napolitano et al., 2019) supports this operation by averaging the Enrichment Scores over multiple profiles and applying the Fisher method to aggregate their p-values. Using this approach, we merged together all the LINCS profiles induced by the same drug in the same cell line across different concentrations and treatment durations. An additional profile for each drug was generated by averaging all conditions, including different cell lines (termed "independent"). We used both approaches to obtain both cell-specific and cell-independent meta-profiles. We ended up with 17,259 drug-induced PEPs.

Additivity of the drugs at the pathway level

We showed in previous work that the transcriptional response to combinatorial drug treatment at promoters and enhancers is effectively described by a linear combination of the responses of the individual drugs (log2FC values) (Rapakoulia et al., 2017). We used our previous dataset to test if this additive relationship also applies to PEPs. Accordingly, we performed multivariable linear regression analysis, where PEPs of individual drugs were considered as explanatory variables and the PEP of combinatorial drug action as the response variable. We applied our analysis to five pathway databases, Biological Process (BP), Molecular Function (MF), Cellular Component (CC), Transcription Factor Targets (TFT) and Canonical Pathways (C2_CP). **Supplementary Table 6** demonstrates the performance of the linear regression model after ten-fold cross-validation in all the three drug combinations and the four pathway collections. The results show that the linear model using PEPs can describe the relation between single and combinatorial treatment.

To validate whether both single drug PEPs contribute to the model, we performed the same regression analysis 100,000 times with random permutations of one of the single drug PEP. The Pearson

correlation between the observed and predicted values after the permutations was significantly lower for all combinations compared to the regression model based on the non-permuted individual drug PEPs (**Supplementary Figure 2D-F**).

In silico validation with the Pluripotency Score

While the DECCODE framework is based on an unbiased, data-driven approach, we devised a pluripotency-specific method to score gene expression profiles based on prior knowledge about genes involved in the conversion to hIPSCs. We then compared these scores with DECCODE scores to validate the predictions. The pluripotency score (PS) is based on genes that were identified as differentially expressed during reprogramming. In particular, we used the "early pluripotency", "late pluripotency", "early somatic", and "late somatic" gene sets previously identified (Cacchiarelli et al., 2015) that characterize gene expression dynamics in the corresponding stages of conversion from human fibroblasts (HiF-T) to hIPSCs. The original study included also other six sets from the same context, which we used as statistical background. For each of the ten sets and for each drug-induced gene expression profile, we computed an Enrichment Score (ES) using the gep2pep tool. We then ranked them from 1st to 10th according to their ESs, thus obtaining a PEP profile. Finally, we computed the Pluripotency Score (PS) for each profile *p* as:

$$PS(p) = \log\left(\frac{R_{early\ pluripotent}(p)+R_{late\ pluripotent}(p)}{R_{early\ somatic}(p)+R_{late\ somatic}(p)}\right),$$

where $R_x(p)$ is the rank of the gene set *x* within the profile *p*. The score is positive (negative) when genes associated with pluripotency stages tend to be more up-regulated (down-regulated) than genes associated with somatic stages.

Computational validation of the obtained combinations was assessed using PSs (**Figure 1B**). In particular, for any drug combination the median of the corresponding PSs was used. Moreover, the top 30 solutions were considered for a given drug combination size, thus obtaining 30 median PS values. In order to obtain a corresponding null distribution, the same calculation was performed also for random drug combinations of equal size. The random selection was repeated 100 times for each size, thus obtaining 100 times 30 median PSs. **Figure 1B** summarizes this analysis by reporting the obtained 30 versus 300 median PSs for drug combinations of size 1 to 10.

Selection of the drugs for the experimental validation

In order to validate the method experimentally, we selected two lists of drugs: the first using the single-drug approach and the second using the combined approach.

For single drugs, we selected 25 drugs from the top of the DECCODE ranking, plus 20 non-top drugs for comparison. In particular, to build the set of non-top drugs, we chose 10 drugs from the middle of the ranking and 10 drugs from the bottom. In case of an overlap between top and non-top drugs due to the same drugs being profiled across multiple cellular contexts in the LINCS database, we removed the repeated drugs from the non-top sets and chose the next one in the ranking. In all cases, only the drugs included in the Prestwick-FDA library or in the SelleckChem Kinase inhibitors library were considered.

For the drug pairs, we applied further filtering in order to obtain a heterogeneous collection also taking into account the results from the single drug experiments. First, we reran the DECCODE algorithm for drug combinations directly considering only the available drugs in Prestwick-FDA library or in the SelleckChem Kinase inhibitors library. For each drug PEP, we fitted a linear regression model as previously described. We then picked the top 30 drugs and for each of them we selected 20 drugs whose addition to the linear models improves best the Spearman correlation with the hIPSCSs profile. Duplicated solutions were removed, resulting in 522 unique drug combinations. From the remaining pairs, we excluded those containing at least one drug that had already been tested in single drug experiments and showed negative outcome (FC < 1 for number of colonies or covered area). We then selected for experimental validation the top eight combinations having the highest DECCODE ranking and passing the above filters. Since Tazobactam showed particularly encouraging results in primary and secondary reprogramming, we considered two additional drug combinations that included Tazobactam (one such pair was already among the top eight, Tazobactam+Motesanib). Therefore, we finally obtained a list of ten drug pairs covering 16 drugs with Tazobactam included in three different pairs and another two drugs appearing twice (Motesanib and Afatinib).

Colony quantification

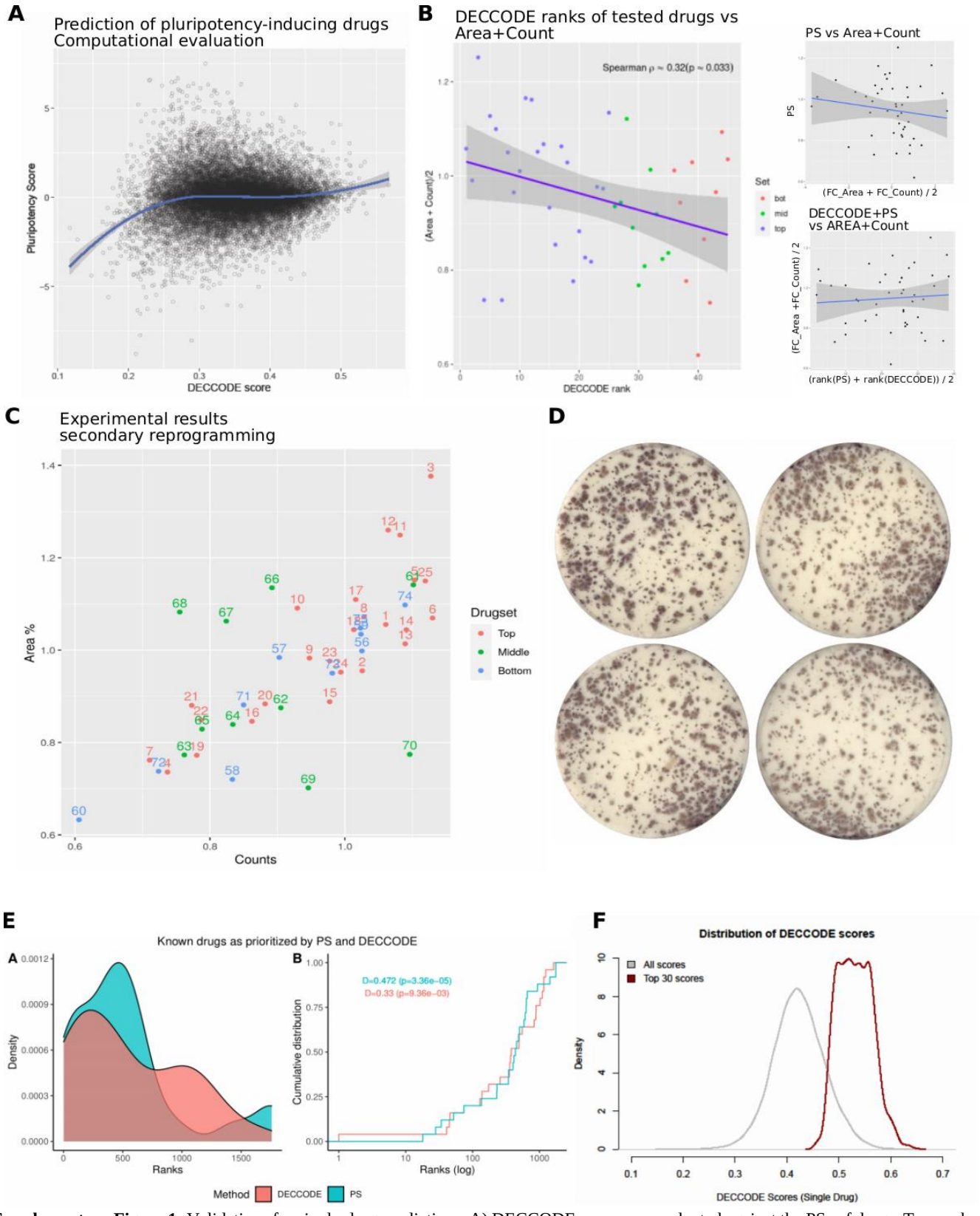
To quantify colony number and size in an unbiased and reproducible way, a completely automated procedure was developed, which is divided in two phases. The first phase was performed through a Matlab script which identifies each well inside all the plate scans, applies a 3X contrast, and saves each of them to a separate image file. The second phase was performed by an ImageJ macro that loads the well images produced by the previous step and performs the final counting and area estimation on each of them. Both Matlab and ImageJ source code, together with the high resolution plate scan images, are available online (Napolitano et al., 2020) (DOI: 10.5281/zenodo.3732772).

In secondary reprogramming experiments, colony count and area values were averaged across the three replicates of the same treatment and across the two controls on the same plate. Average fold change of treatments versus controls were then obtained accordingly (**Figure 2A**, small panels). In order to summarize both count and covered area values together, the corresponding fold changes were averaged (**Figure 2A**, main panel). In the primary reprogramming experiment, counts and areas for tazobactam treatment and controls were compared directly (**Figure 2B**, small panels). Two controls were excluded according to the Bonferroni Outlier Test ($p < 0.0118$ and $p < 0.0106$ respectively). In the case of primary reprogramming results, in order to summarize counts and covered areas together, all the absolute values were normalized dividing by the corresponding mean of the controls (**Figure 2B**, main panel). The same was done for drug combination experimental results (**Figure 3A**).

Computation of DECCODE scores for all the FANTOM5 cell types

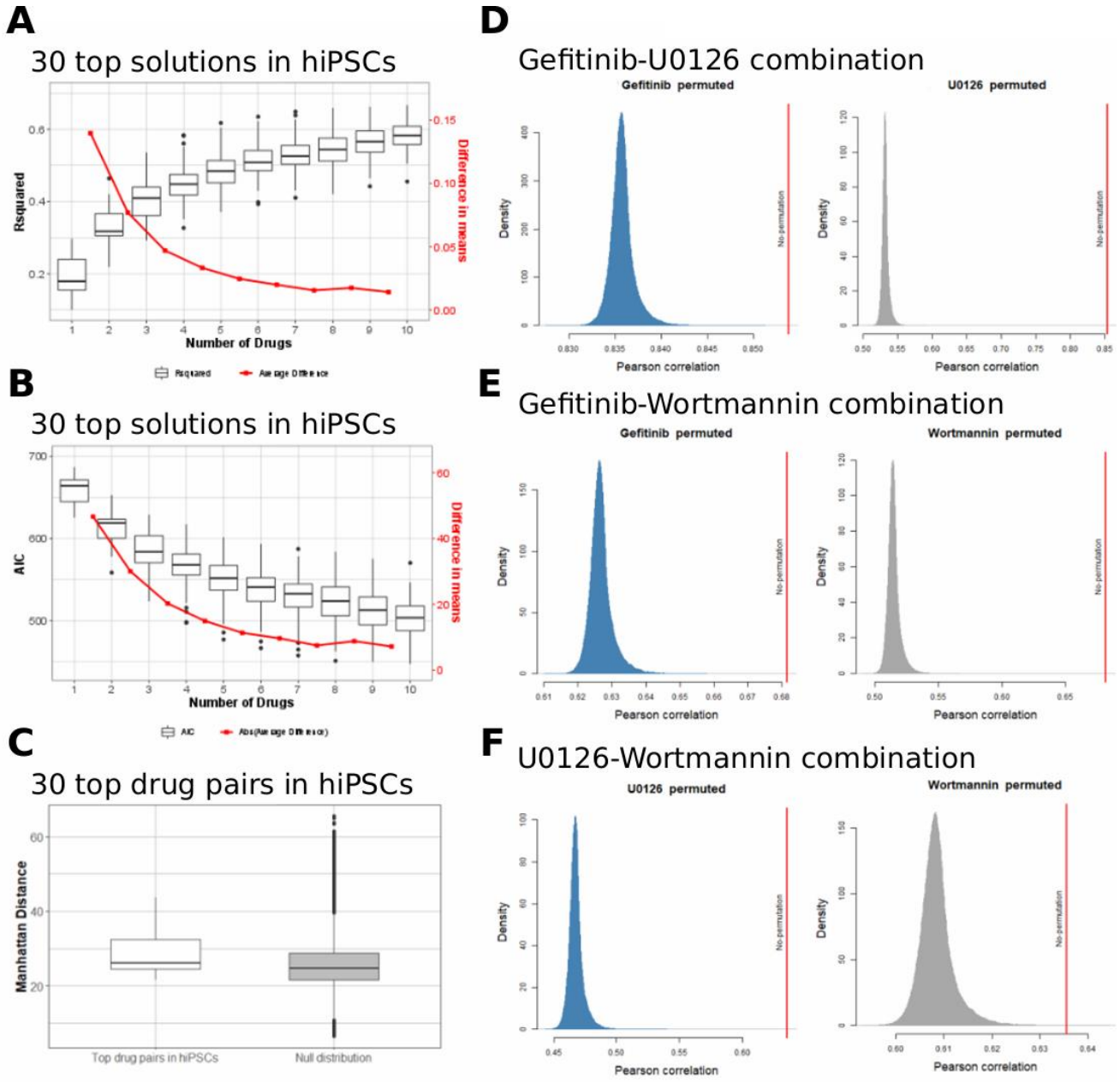
The FANTOM5 cell types include sub-types that are very similar, thus the corresponding expression profiles are not different enough to produce sub-type specific predictions. Therefore, we merged similar cell types to form a single meta-cell profile (see methods subsection “Merging of Pathway-based Expression Profiles”). In order to systematically select which cell-type profiles to merge, we took advantage of the previously computed PEP-based and ontology-based cell type distances (refer to subsection “Conversion to pathway-based profiles”). We applied the Affinity Propagation algorithm (Frey and Dueck, 2007) individually to each of the two pairwise distances to obtain two different clusterings of the same cell types (**Supplementary Figure 3**). Affinity Propagation clustering was performed using the "apcluster" R package (Bodenhofer et al., 2011). Finally, we built a consensus clustering by assigning two cell types to the same cluster if and only if they were assigned to the same cluster by both the ontology-based and PEP-based clusterings. Meta-cell profiles are obtained by averaging all the profiles included in the same cluster. We then computed single-drug and multiple-drug DECCODE scores for all the meta-cell profiles.

Supplementary Figures

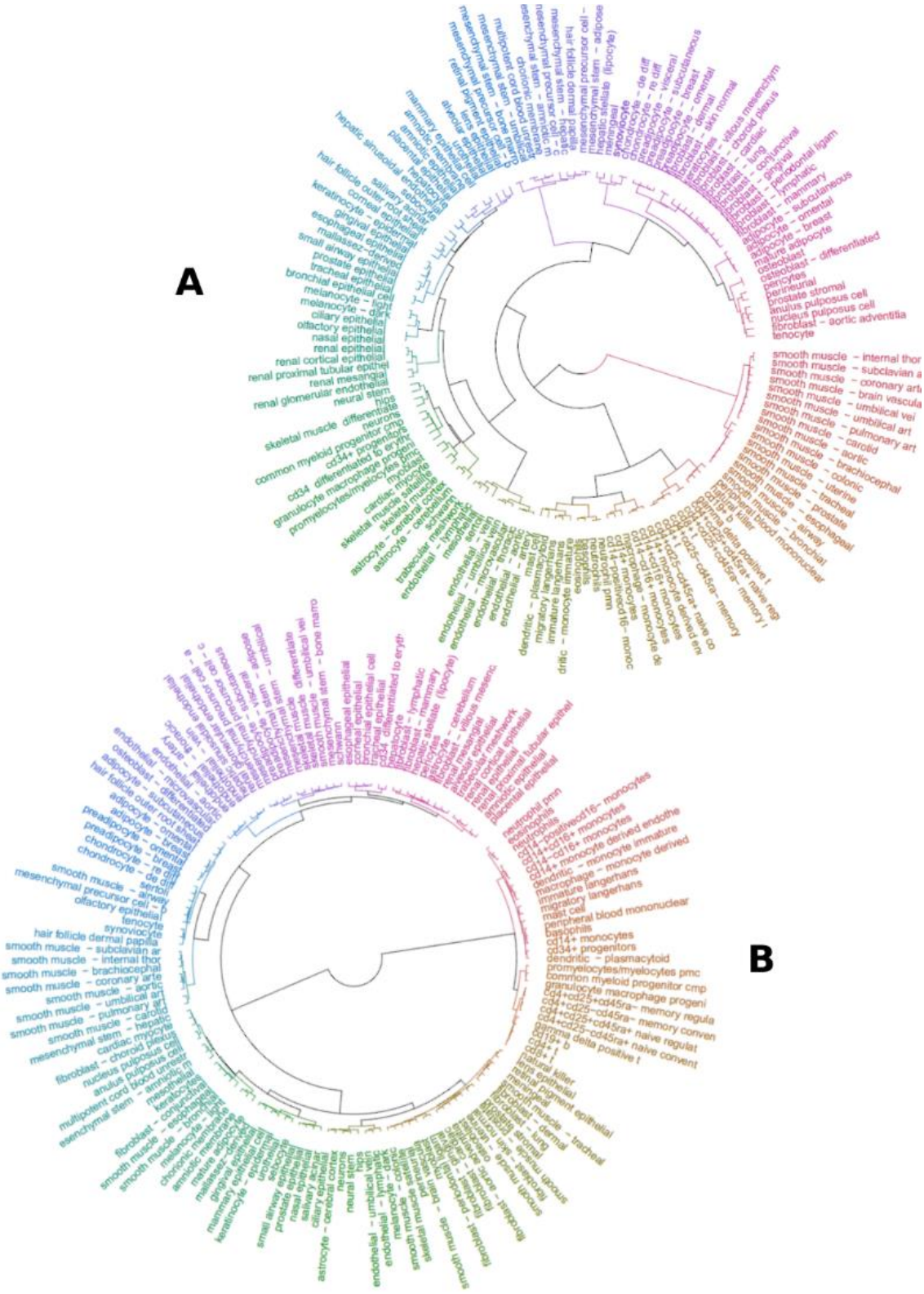


Supplementary Figure 1: Validations for single-drug predictions. A) DECCODE scores are evaluated against the PSs of drugs. Top-ranked (higher DECCODE scores) drugs exhibit higher PSs while bottom-ranked drugs exhibit lower PSs. B) Efficacy of the 45 single drug treatments experimentally tested (colony area and counts) versus DECCODE ranking (left), PS score (top-right) and combined DECCODE

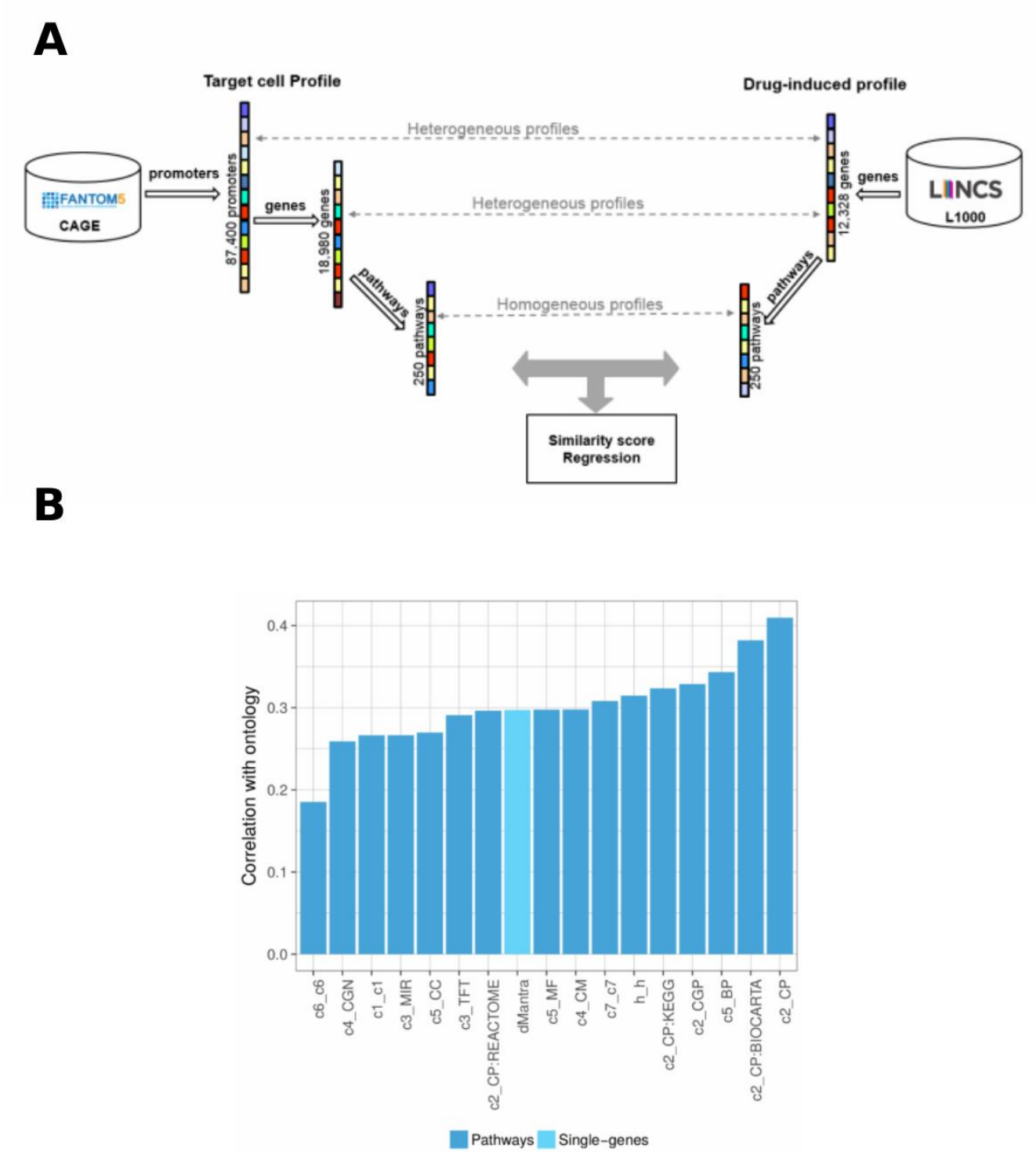
and PS ranking (bottom right). C) Experimental validation of drugs enhancing conversion to hIPSCs: number of colonies formed versus % of covered area. Tazobactam (ID: 3) shows the highest performance for covered. All ID codes are explained in Supplementary Table 2. D) Imaging of wells treated with tazobactam and OSKM (left) against the controls (only OSKM) for the specific plate (right). The fold change in number of colonies and area covered by the colonies for each drug treatment was computed against the control experiments of the corresponding plate. E) Density and Cumulative distribution of the ranks assigned in the small molecules reported in (Chen et al., 2020) based on PS and DECCODE scores. F) Distribution of the top 30 DECCODE drug scores and distribution of all the 1,768 DECCODE drug scores for all meta-cell cluster profiles.



Supplementary Figure 2: Validations for drug combinations. A) Rsquared and B) AIC criterion of the top 30 regression solutions for the hIPSC target profile as more drugs are added to the regression models. Red line highlights the average incremental improvement. C) Distribution of the distances between the drug profiles for the top 30 selected drug pairs in hiPSCs by DECCODE. Null distribution was created by random sampling 1000 drug profiles from LINCS dataset and computing their pairwise distances (499500 distances). D-F) Density plots of the Pearson correlation coefficients between observed and predicted values after the permutations of individual profiles for Gefitinib_U0126 (D), Gefitinib_Wortmannin (E), and U0126_Wortmannin (F) drug combinations using the C2_CP PEPs. The Pearson correlation coefficient achieved without permutation is also reported. Similar results were obtained for all the pathway collections of Supplementary Table 6.



Supplementary Figure 3: Hierarchical clustering visualization of cell types based on the ontology distance (A) and pathway distance (B). Affinity Propagation algorithm (Frey and Dueck, 2007) was applied for the used clustering.



Supplementary Figure 4: A) Harmonization of expression profiles. Promoter-based target cell type profiles are converted to gene-based profiles. Gene-based profiles for both primary cells and drug treated cell lines are then converted to pathway-based expression profiles (PEPs). B) Spearman correlation between pathway distances and ontology distance. Pairwise cell similarity obtained by the different pathway collections are evaluated against cell similarity obtained by the Cell Ontology annotation. Mantra distance (Iorio et al., 2010) computed on single gene ranks is also tested against the ontology distance.

Supplementary Tables

Supplementary Table 1: Top-ranked drugs by DECCODE for the hIPSCs target profile. Several drugs have been already associated with enhancement of the reprogramming process.

Drugs	Indication	ID
Motesanib (Chen et al., 2014)	treatment in solid tumors	1
Fluticasone	activating glucocorticoid receptors, inhibiting nuclear factor kappa b and inhibiting lung eosinophilia in rats	2
Tazobactam	bacterial β-lactamase inhibitor	3
Cyclizine	histamine H1 antagonist	4
Etofenamate (Yang et al., 2011)	nonsteroidal anti-inflammatory drug (NSAID), COX inhibitor	5
Pentoxifylline	modulates immunologic activity by stimulating cytokine production.	6
Irsogladine	anti-inflammatory agent	7
Leflunomide	pyrimidine synthesis inhibitor/chemotherapeutic	8
Dexfenfluramine	serotonergic anorectic drug/studied in obesity	9
Paroxetine	selective serotonin reuptake inhibitor (SSRI) drug commonly known as Paxil	10
Afatinib	tyrosine kinase inhibitor /ErbB family blocker	11
Doramapimod	highly potent p38 MAPK inhibitor	12
Nalbuphine	anticonvulsant effect/inhibited breast cancer cell growth and tumorigenesis	13
PIK-93	PI4KIIIβ inhibitor	14
Glycopyrrolate	synthetic anticholinergic agent	15
SGX523	MET receptor tyrosine kinase inhibitor.	16
Dasatinib (Lin and Wu, 2015)	Src family tyrosine kinase inhibitor	17
SB-203580 (Di Stefano et al., 2016)	inhibitor of p38α and p38β	18
Doxycycline (Chang et al., 2014)	antibacterial agent	19
Saracatinib (Zhang et al., 2014)	inhibitor of the Src/abl family	20
Levetiracetam	plays a role in the control of regulated secretion in neural and endocrine cells	21
Tranylcypromine (Di Stefano et al., 2016)	belongs to a class of antidepressants monoamine oxidase inhibitors (MAOIs).	22
HMN-214	PLK inhibitor	23
histamine	immune responses, neurotransmitter	24
dabrafenib	chemotherapeutic, inhibitor of the associated enzyme B-Raf	25

Supplementary Table 2: Small molecules facilitating reprogramming to hIPSCs reported in (Chen et al., 2020), having an available profile in LINCS database, and their ranking based on PS and DECCODE scores. The DECCODE rankings for both hIPSCs cells and meta-cell cluster 34 (see Supplementary Table 4) which includes hIPSCs as a target profile are reported.

Pert name	**PS ranking**	**DECCODE ranking (hIPSCs)**	**DECCODE ranking (hIPSCs meta-cell)**
BAY-K8644	1434	884	581
BIX-01294	136	1157	783
CHIR-99021	638	1160	368
D-4476	18	369	405
LY-294002	346	494	121
PD-0325901	52	498	53
RG-108	230	554	664
Y-27632	74	378	42
Curcumin	438	258	124
Dexamethasone	589	1022	834
Dovitinib	34	1239	1480
EPZ004777	455	1628	1557
Forskolin	659	129	761
Lenvatinib	1757	355	477
Motesanib	936	1	3
Nintedanib	28	46	168
Pazopanib	633	74	32
Quercetin	613	890	655
Resveratrol	489	41	26
Sorafenib	402	833	710
Sunitinib	415	174	242
Tivozanib	504	138	50
Tranylcypromine	232	45	34
Valproic-acid	1758	1104	1026
Vandetanib	343	359	589
Mean ranking	528.52	553.2	471.4

Supplementary Table 5: Small molecules that were experimentally proved to facilitate various cell conversions and were predicted among the top drug profiles for the corresponding Meta-cells from the DECCODE single and multi-drug approach.

DECCODE Sing Drug Approach		
Target Meta-cell	**Small Molecule**	**Rank**
Astrocyte cells- Cerebral Cortex	Tranylcypromine (Tian et al., 2016)	44
Hepatocyte cells	RG108 (Zhu et al., 2014)	55
hIPS cells - Neural Stem cells	PD0325901 (Lin et al., 2009; Wang et al., 2011; Zhu et al., 2010)	53
hIPS cells - Neural Stem cells	Tranylcypromine (Li et al., 2009; Zhu et al., 2010)	34
Mesenchymal Stem cells - Amniotic membrane - Multipotent Cord Blood Unrestricted Somatic Stem cells	PD0325901 (Lai et al., 2017)	59
Neurons	Y27632 (Hu et al., 2015)	7
Neurons	PD0325901 (Dai et al., 2015)	19

DECCODE Multidrug Approach		
Target Meta-cell	**Small Molecule**	**Rank**
Cardiac Myocyte cells	BIX01294 (Cao et al., 2016)	6
hIPS cells - Neural Stem cells	Tranylcypromine (Li et al., 2009; Zhu et al., 2010)	18
Neurons	PD032590 (Dai et al., 2015)	8
Neurons	Y27632 (Hu et al., 2015)	19

Supplementary Table 6: Pearson and Spearman correlation between fitted and observed PEPs in combinatorial treatment (Rapakoulia et al., 2017). The multivariable linear regression model was applied in five different pathway collections. The values shown in the table are the mean performance after tenfold cross validation.

	BP (4436 pathways)		**MF (901pathways)**		**CC (580 pathways)**		**TFT (615 pathways)**		**C2_CP (250 pathways)**	
	Pearson	**Spearman**	**Pearson**	**Spearman**	**Pearson**	**Spearman**	**Pearson**	**Spearman**	**Pearson**	**Spearman**
Gefitinib-U0126	0.8038	0.7883	0.8685	0.8554	0.8401	0.8401	0.7929	0.8129	0.8538	0.8550
Gefitinib-Wortmannin	0.7460	0.7226	0.7538	0.7203	0.7947	0.6297	0.7276	0.7304	0.6814	0.6751
U0126-Wortmannin	0.6430	0.6211	0.6197	0.6095	0.6826	0.6502	0.6758	0.6336	0.6355	0.6080

References

Bard, J., Rhee, S.Y., and Ashburner, M. (2005). An ontology for cell types. Genome Biol. *6*.

Bodenhofer, U., Kothmeier, A., and Hochreiter, S. (2011). APCluster: an R package for affinity propagation clustering. Bioinformatics *27*, 2463–2464.

Cacchiarelli, D., Trapnell, C., Ziller, M.J., Soumillon, M., Cesana, M., Karnik, R., Donaghey, J., Smith, Z.D., Ratanasirintrawoot, S., Zhang, X., et al. (2015). Integrative analyses of human reprogramming reveal dynamic nature of induced pluripotency. Cell *162*, 412.

Cao, N., Huang, Y., Zheng, J., Spencer, C.I., Zhang, Y., Fu, J.-D., Nie, B., Xie, M., Zhang, M., Wang, H., et al. (2016). Conversion of human fibroblasts into functional cardiomyocytes by small molecules. Science *1502*.

Chang, M.Y., Rhee, Y.H., Yi, S.H., Lee, S.J., Kim, R.K., Kim, H., Park, C.H., and Lee, S.H. (2014). Doxycycline enhances survival and self-renewal of human pluripotent stem cells. Stem Cell Reports *3*, 353–364.

Chen, G., Xu, X., Zhang, L., Fu, Y., Wang, M., Gu, H., and Xie, X. (2014). Blocking autocrine VEGF signaling by sunitinib, an anti-cancer drug, promotes embryonic stem cell self-renewal and somatic cell reprogramming. Cell Res. *24*, 1121–1136.

Chen, G., Guo, Y., Li, C., Li, S., and Wan, X. (2020). Small Molecules that Promote Self-Renewal of Stem Cells and Somatic Cell Reprogramming. Stem Cell Rev. Reports *16*, 511–523.

Dai, P., Harada, Y., and Takamatsu, T. (2015). Highly efficient direct conversion of human fibroblasts to neuronal cells by chemical compounds. J. Clin. Biochem. Nutr. *56*, 166–170.

Frey, B.J., and Dueck, D. (2007). Clustering by Passing Messages Between Data Points. Science (80-.). *315*, 972–976.

Hu, W., Qiu, B., Guan, W., Wang, Q., Wang, M., Li, W., Gao, L., Shen, L., Huang, Y., Xie, G., et al. (2015). Direct Conversion of Normal and Alzheimer's Disease Human Fibroblasts into Neuronal Cells by Small Molecules. Cell Stem Cell *17*.

Iorio, F., Bosotti, R., Scacheri, E., Belcastro, V., Mithbaokar, P., Ferriero, R., Murino, L., Tagliaferri, R., Brunetti-Pierri, N., Isacchi, A., et al. (2010). Discovery of drug mode of action and drug repositioning from transcriptional responses. Proc. Natl. Acad. Sci. *107*, 14621–14626.

Lai, P.-L., Lin, H., Chen, S.-F., Yang, S.-C., Hung, K.-H., Chang, C.-F., Chang, H.-Y., Lu, F.L., Lee, Y.-H., Liu, Y.-C., et al. (2017). Efficient Generation of Chemically Induced Mesenchymal Stem Cells from Human Dermal Fibroblasts. Sci. Rep. *7*, 44534.

Li, W., Zhou, H., Abujarour, R., Zhu, S., Joo, J.Y., Lin, T., Hao, E., Schöler, H.R., Hayek, A., and Ding, S. (2009). Generation of Human Induced Pluripotent Stem Cells in the Absence of Exogenous *Sox2*. Stem Cells *27*, N/A-N/A.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. *1*, 417–425.

Lin, T., and Wu, S. (2015). Reprogramming with Small Molecules instead of Exogenous Transcription Factors. Stem Cells Int. *2015*, 794632.

Lin, T., Ambasudhan, R., Yuan, X., Li, W., Hilcove, S., Abujarour, R., Lin, X., Hahm, H.S., Hao, E., Hayek, A., et al. (2009). A chemical platform for improved induction of human iPSCs. Nat. Methods *6*, 805–808.

Lizio, M., Harshbarger, J., Shimoji, H., Severin, J., Kasukawa, T., Sahin, S., Abugessaisa, I., Fukuda, S., Hori, F., Ishikawa-Kato, S., et al. (2015). Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biol. *16*, 22.

Napolitano, F., Sirci, F., Carrella, D., and Di Bernardo, D. (2016). Drug-set enrichment analysis: A novel tool to investigate drug mode of action. Bioinformatics *32*, 235–241.

Napolitano, F., Carrella, D., Mandriani, B., Pisonero-Vaquero, S., Sirci, F., Medina, D.L., Brunetti-Pierri, N., and di Bernardo, D. (2018). gene2drug: a computational tool for pathway-based rational drug repositioning. Bioinformatics *34*, 1498–1505.

Napolitano, F., Carrella, D., Gao, X., and di Bernardo, D. (2019). gep2pep: a bioconductor package for the creation and analysis of pathway-based expression profiles. Bioinformatics.

Napolitano, F., Rapakoulia, T., Annunziata, P., Hasegawa, A., Cardon, M., Napolitano, S., Vaccaro, L., Iuliano, A., Wanderlingh, L.G., Kasukawa, T., et al. (2020). Automatic identification of small molecules that promote cell conversion and reprogramming - plate scans, colony quantification scripts, and DECCODE ranking.

Rapakoulia, T., Gao, X., Huang, Y., De Hoon, M., Okada-Hatakeyama, M., Suzuki, H., and Arner, E. (2017). Genome-scale regression analysis reveals a linear relationship for promoters and enhancers after combinatorial drug treatment. Bioinformatics *33*.

Di Stefano, B., Collombet, S., Jakobsen, J.S., Wierer, M., Sardina, J.L., Lackner, A., Stadhouders, R., Segura-Morales, C., Francesconi, M., Limone, F., et al. (2016). C/EBPα creates elite cells for iPSC reprogramming by upregulating Klf4 and increasing the levels of Lsd1 and Brd4. Nat. Cell Biol. *18*, 371–381.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. *102*, 15545–15550.

Tian, E., Sun, G., Sun, G., Chao, J., Ye, P., Warden, C., Riggs, A.D., and Shi, Y. (2016). Small-Molecule-Based Lineage Reprogramming Creates Functional Astrocytes. Cell Rep. *16*.

Wang, Q., Xu, X., Li, J., Liu, J., Gu, H., Zhang, R., Chen, J., Kuang, Y., Fei, J., Jiang, C., et al. (2011). Lithium, an anti-psychotic drug, greatly enhances the generation of induced pluripotent stem cells. Cell Res. *21*, 1424–1435.

Yang, C.S., Lopez, C.G., and Rana, T.M. (2011). Discovery of nonsteroidal anti-inflammatory drug and anticancer drug enhancing reprogramming and induced pluripotent stem cell generation. Stem Cells *29*, 1528–1536.

Zhang, X., Simerly, C., Hartnett, C., Schatten, G., and Smithgall, T.E. (2014). Src-family tyrosine kinase activities are essential for differentiation of human embryonic stem cells. Stem Cell Res. 379–389.

Zhu, S., Li, W., Zhou, H., Wei, W., Ambasudhan, R., Lin, T., Kim, J., Zhang, K., and Ding, S. (2010). Reprogramming of Human Primary Somatic Cells by OCT4 and Chemical Compounds. Cell Stem Cell *7*, 651–655.

Zhu, S., Rezvani, M., Harbell, J., Mattis, A.N., Wolfe, A.R., Benet, L.Z., Willenbring, H., and Ding, S. (2014). Mouse liver repopulation with hepatocytes generated from human fibroblasts. Nature *508*, 93–97.