nature biotechnology

A protein activity assay to measure global transcription factor activity reveals determinants of chromatin accessibility

Bei Wei[1], Arttu Jolma[1], Biswajyoti Sahu[2], Lukas M Orre[3], Fan Zhong[1], Fangjie Zhu[1], Teemu Kivioja[2], Inderpreet Sur[1], Janne Lehtiö[3], Minna Taipale[1] & Jussi Taipale[1,2,4]

No existing method to characterize transcription factor (TF) binding to DNA allows genome-wide measurement of all TF-binding activity in cells. Here we present a massively parallel protein activity assay, active TF identification (ATI), that measures the DNA-binding activity of all TFs in cell or tissue extracts. ATI is based on electrophoretic separation of protein-bound DNA sequences from a highly complex DNA library and subsequent mass-spectrometric identification of the DNA-bound proteins. We applied ATI to four mouse tissues and mouse embryonic stem cells and found that, in a given tissue or cell type, a small set of TFs, which bound to only ~10 distinct motifs, displayed strong DNA-binding activity. Some of these TFs were found in all cell types, whereas others were specific TFs known to determine cell fate in the analyzed tissue or cell type. We also show that a small number of TFs determined the accessible chromatin landscape of a cell, suggesting that gene regulatory logic may be simpler than previously appreciated.

Transcription factors are proteins that read the gene regulatory information in DNA and determine which genes are expressed. There have been many efforts to determine the most important TFs in different cell types. Genetic analyses have revealed that cell identity is determined by relatively few TFs that regulate each other and often bind to specific regions of the genome together[1–3]. Cell identity can often be reprogrammed by exogenous expression of one to five TFs; for example, differentiated cells can be transformed into induced pluripotent stem (iPS) cells by expression of several different subsets of the TFs OCT4 (POU5F1), SOX2, KLF4, c-MYC, and ESRRB[4–6]. These results indicate that the cellular regulatory system can be controlled by a relatively small subset of TFs.

Analyses based on gene expression profiling have, however, revealed that most tissues express hundreds of TFs[7]. Similarly, total proteomic analyses have indicated that tissues commonly express >40% of all ~1,500 known human TF proteins[8]. Experiments based on chromatin immunoprecipitation followed by sequencing (ChIP-seq) have also suggested that a large number of TFs are active in individual cell lines[9,10]. Taken together, these results suggest that downstream of the master regulators, gene regulatory logic inside cells seems to be extremely complex, and that the cellular state could potentially be defined by a very large number of regulatory interactions. However, little information exists on which TFs have the strongest activities in a given cell type. This is because previous analyses have either only analyzed RNA or protein levels[7,8] or measured individual TF activities using methods that cannot compare activity levels between TFs (for example, ChIP-seq[9–11]).

To determine the relative activities of TFs in cells, here we developed a massively parallel protein activity assay, ATI, that determines the absolute number of TF binding events from cells or tissues. This information can then be used to derive relative DNA-binding activity of different TFs in the same sample. The word 'activity' is used here in the same sense as in enzymology, where activity represents total enzyme activity (specific activity × molar amount). Because ATI measures TF activity, and not occupancy at specific sites, it can be used to build models of TF binding on the basis of biochemical principles, which will contribute to our understanding of how DNA sequence determines when and where genes are expressed. Here we have used ATI to determine the relative distribution of biochemical TF activities in different cells, as well as the relationship between TF DNA-binding activities and overall chromatin architecture. We found that only few TFs displayed strong DNA-binding activity in any cell or tissue type from all of the tested organisms. The strongly active TFs can be used to predict transcript start positions in yeast and chromatin accessibility in mammalian cells.

RESULTS

De novo discovery of motifs bound by transcription factors from cell extracts

In ATI, a library of double-stranded oligonucleotides containing a 40-bp random sequence is incubated with a nuclear extract from different cell or tissue types. The oligonucleotides bound by TFs are then separated from the unbound fraction by electrophoretic mobility

[1]Division of Functional Genomics and Systems Biology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden. [2]Genome-Scale Biology Program, University of Helsinki, Helsinki, Finland. [3]Department of Oncology–Pathology, Science for Life Laboratory, Karolinska Institutet, Stockholm, Sweden. [4]Department of Biochemistry, University of Cambridge, Cambridge, UK. Correspondence should be addressed to J.T. (ajt208@cam.ac.uk).

Received 8 December 2017; accepted 23 March 2018; published online 21 May 2018; doi:10.1038/nbt.4138

shift assays (EMSAs) (**Fig. 1a**). The bound DNA fragments are eluted from the gel and amplified by PCR, and the entire process is repeated three more times. Comparison of millions of sequences derived from the input and the selected libraries then allows for identification of enriched binding motifs that correspond to the TFs present in the nuclear extract. Given that the binding motifs identified are relatively short as compared to the 40-bp random sequence, the sequence flanking the motif can also be used as a unique molecular identifier[12], allowing absolute quantification of the number of proteins bound to each type of motif.

As an initial test, we performed ATI using nuclear extract from mouse embryonic stem (mES) cells. *De novo* motif discovery using Autoseed[13] revealed motifs that were characteristic of TF families, such as NFI, RFX, KLF and POU, and of subfamily-specific motifs for MIT–TFE basic helix-loop-helix (bHLH) proteins, class I ETS factors, ZIC zinc fingers and ERR-type nuclear receptors (**Fig. 1b**; for motif similarity, see **Supplementary Fig. 1a**). Many motifs were similar to motifs bound by known lineage-determining factors for ES cells[14–18], such as KLF4, POU5F1 (OCT4), ZIC3 and ESRRB, suggesting that the ATI assay was able to detect the DNA-binding activity of such TFs. Some, but not all, ATI motifs were also detected by MEME motif mining of DNase I hypersensitive sites (DHSs) from ES cells, suggesting that ATI could also detect activities of TFs that contributed to chromatin accessibility (**Fig. 1b**).

Most recovered motifs have previously been identified

Although *de novo* motif discovery using Autoseed is relatively sensitive and can identify motifs that represent ~5 p.p.m. of all sequences, it cannot identify very rare events. To detect such events, we also analyzed the enrichment of known motif matches during the ATI process (**Fig. 1c** and **Supplementary Table 1**). This method yielded results similar to those for the *de novo* method, but it also identified an enrichment of motif matches for additional TFs, including homeodomain motifs similar to the motif bound by the pluripotency regulator NANOG, indicating that it has higher sensitivity than the *de novo* motif discovery method. The relatively low enrichment of the homeodomain motif was consistent with low abundance of NANOG in the mass spectrometry (MS) analyses (**Supplementary Table 2**). However, even using known motif enrichment, it is hard to unambiguously assign a weak activity to a specific TF. This is because it is difficult to determine whether the motif enrichment that is too weak to be detected by *de novo* methods represents specific enrichment of the tested motif or whether it is a consequence of stronger enrichment of a related motif for another TF.

Notably, all of the motifs recovered from ES cells were known before this study, suggesting that few strong TF activities remain to be discovered. Among the motifs found, there were both monomeric motifs, which are bound by one single TF, and dimeric motifs, which are bound by dimers formed by two TFs from the same structural family (**Fig. 1b**). However, some motifs representing dimers formed on DNA[19], such as the well-known SOX2–POU5F1 motif, were absent, which implied that the ATI method might be biased against such DNA-dependent dimers, as a much larger number of monomeric sites exist in random sequences than dimeric sites (**Supplementary Fig. 1b**). To address this issue, we performed two additional experiments using a synthetic library consisting of known motifs and a library derived from mouse genomic sequences. These targeted analyses were more sensitive than the method based on a random library and resulted in identification of many additional motifs, including that for CTCF (**Supplementary Table 3**). However, analysis of the data did not reveal DNA-dependent dimeric motifs, suggesting that the activity of DNA-dependent TF dimers was lower than that of the corresponding monomeric TFs. This is likely because specific formation of DNA-dependent dimers requires that the binding activity of at least one of the TFs be relatively low, as its concentration must be less than the individual K_d values toward the motif and the partner but greater than the K_d toward the combination of the dimer motif and the TF partner. Owing to the fact that proteins are more concentrated in the nucleus than in the nuclear extract, such dimers are difficult to detect using ATI.

Identification of specific transcription factors by mass spectrometry

The ATI assay can be used to identify active motifs, but analysis of motif activity alone cannot, in most cases, identify the specific TF that is active in the tissue or cell types, due to the fact that many related proteins can bind to the same sequence motif. To address this, we captured DNA-binding proteins from nuclear extracts of mES cells by using a biotinylated version of the control and enriched ATI DNA ligand libraries. After washing and elution, we identified proteins that were bound to the ATI ligands by using MS (Online Methods). This analysis revealed the TFs that bound to the ligand. In most cases, a motif could be assigned to a specific protein or a group of paralogous proteins (**Supplementary Table 4**).

Highly active transcription factors can be classified into three categories

We next applied ATI to the identification of TFs that were active in mES cells and four adult mouse tissues, including the heart, spleen, brain and liver. *De novo* motif discovery followed by motif match counting revealed that limited sets of TFs were highly active in different tissue types (**Fig. 2a** and **Supplementary Table 5**). In the five samples, only few (two to seven) TFs displayed activities that were >10% of that of the most active TF. The identified motifs could be broadly classified into three groups: common, shared and specific. Five common motifs were found in all cell and tissue types tested. They represented an extended E-box site (gGTCACGTGACc) that was bound by the MIT–TFE family of bHLH TFs, a GGTCAaaGGTCA motif that was bound by a subfamily of nuclear receptors (NRs) and canonical sites that were bound by the NFI, NRF1 and basic leucine zipper (bZIP; such as CREB) family of TFs (**Fig. 2a**; see **Supplementary Fig. 2a** for comparison to known motifs). Even within this common set, there were large differences in quantitative TF binding activities between the cell types, suggesting that the relative activities of the common TFs may have contributed to cell lineage determination (**Fig. 2b** and **Supplementary Fig. 2b**).

In addition to the common TF motifs, we also found three motifs—which corresponded to RBPJ[20], class I ETS family TFs[21], and YY1 and YY2 (ref. 9)—that were shared by more than two different tissue types, suggesting that members of these families of TFs have important roles in many different contexts (**Fig. 2a**; see **Supplementary Fig. 2a** for comparison to known motifs). By contrast, there were some other motifs that were specific for only one or two tissue types (**Fig. 2a**; see **Supplementary Fig. 2c** for comparison to known motifs); some of their corresponding TFs have previously been shown to be crucial for the particular cell identities. For example, the binding motif of the NR THRA that is important for heart function[22] was found only in the heart, whereas the motifs for the PAX and RFX TF families were found only in the spleen, where it is known that members of these families, such as PAX5 and RFX5, contribute to development[23,24] and MHC class II expression[25] of B cells, respectively. In the brain, we detected motifs for the EGR, SCRT, and POU families of TFs, members of which are known to be specifically expressed, and have important

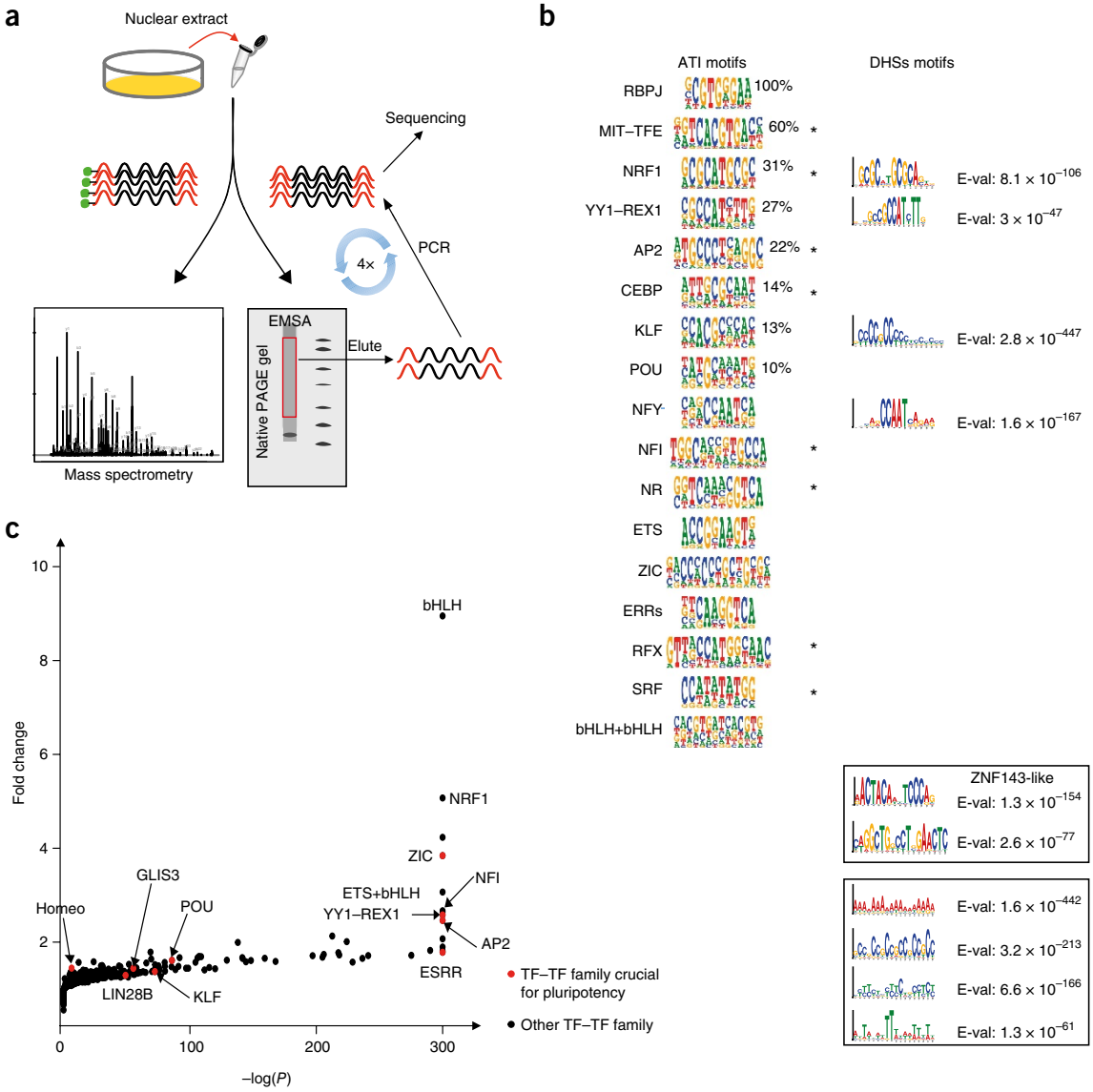

Figure 1 Active transcription factor identification (ATI) assay. (**a**) General description of the process. After incubation of proteins extracted from cells or tissues with double-stranded DNA oligonucleotides containing 40-bp random sequences, oligos bound to the proteins are separated from unbound DNA by native PAGE gel purification (EMSA; right) and amplified by PCR. The process is repeated three more times, resulting in an enriched DNA oligonucleotide pool that reflects activities of the TFs in the cell lysate. The ATI assay consists of two parts: motif analysis and MS identification of the TFs. The sequencing data for the original input and enriched DNA oligos was used for motif analysis; the MS-based identification of the TFs was performed by capturing proteins from the nuclear extract by using a biotinylated oligonucleotide pool followed by MS (left). (**b**) Logos of motifs discovered by using the *de novo* motif discovery approach from ~10 million ATI sequence reads (ATI motifs, based on the Autoseed program[13]) and from the top ~1,000 DHS regions (DHS motifs, based on MEME) in mES cells. In the category of 'ATI motifs', the names of the TFs are based on motif similarity to known motifs (see **Supplementary Fig. 1a** for details). The motifs were quantified with background-corrected absolute molecular counts[12] at cycle 4; the highest count was normalized to 100%. Motifs with counts greater than 10% of the maximum were considered 'strong' motifs; for these motifs, the relative molecular counts are indicated in the upper right-hand corner of the corresponding motif. An asterisk indicates dimeric motifs bound by TFs from the same family. In the 'DHS motifs' category, the top ten motifs with the lowest E values (estimation of the number of motifs one would expect to find by chance if the letters in the input sequences are shuffled) are shown; motifs not detected by ATI are boxed. Notably, *de novo* motif discovery analysis of both ATI and DHS data failed to detect DNA-dependent dimer signals[19], suggesting that either such dimers are not commonly strongly active or that many different dimers contribute to the opening of chromatin at different DHSs. (**c**) Known motif enrichment analysis of ATI data from mES cells. Known motifs of TFs or TF families were matched to the unselected (read count = 16,919,406) and ATI-enriched (read count = 14,306,415) DNA sequences, and *P* values (*x* axis, log scale, calculated by winflat[51]; due to the precision of calculation, many *P* values were set to a minimum of 10^{-300}). *y*-axis: and fold changes calculated for each known motif.

roles, in the brain[26–28]. In the liver, motifs bound by CEBP family TFs, HNF1A–HNF1B, and PAR-domain-containing bZIP family TFs DBP/TEF/HLF were specifically enriched, and these TFs have been verified to be crucial for liver functionality, as well as circadian control of metabolism[29–32]. Moreover, it has been shown that HNF1A–HNF1B and CEBPA, together with other factors, can be used to reprogram fibroblasts into induced hepatocytes (iHeps)[33,34], which indicated the significance of these factors for hepatocyte identity.

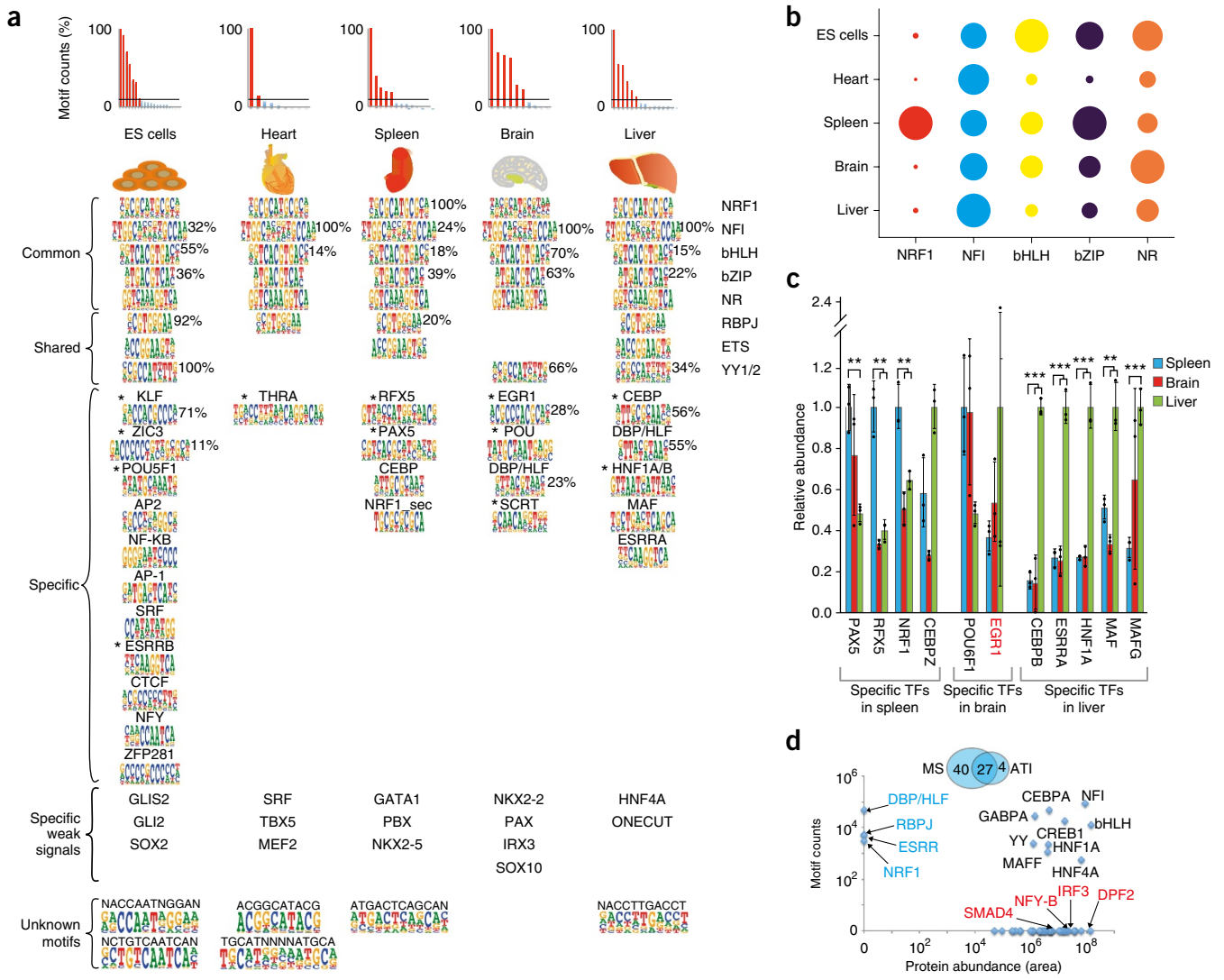

Figure 2 Deciphering the most strongly active TFs in different cell types. (**a**) Top, histogram showing background-corrected absolute molecular counts[12] (*y* axis, motif counts) of all discovered motifs at cycle 4; the highest count was normalized to 100%. Counts >10% of the maximum are indicated by red bars, with the relative activities of corresponding motifs shown on the right-hand corner of the sequence logos. Bottom, sequence logos and the corresponding TFs identified by *de novo* motif discovery from the indicated samples (see **Supplementary Table 8** for PWM models of all logos). The TFs known to contribute to lineage determination in the analyzed samples are indicated by asterisks. Examples of TFs known to be important for the specific tissues whose motifs were identified by only using the known motif discovery pipeline are also indicated (specific weak signals; see **Supplementary Table 6** for details). Some detected but unknown motifs are also shown. The names of the TFs are based on the motifs (see **Supplementary Fig. 2a**,**c**). In cases for which multiple TFs are known to bind the same motif, the motif was assigned to the specific TF on the basis of mRNA expression levels and functional data from previous studies (see references in **Supplementary Fig. 2c**). (**b**) Levels of DNA-binding activities of the common TFs vary between tissues. The sizes of the circles indicate the activities of the five common motifs in the indicated tissues, based on increase of absolute molecular counts[12] of each motif between the original library (cycle 0) and the selected library after the last cycle (cycle 4). The activities of each TF were normalized by setting its highest activity in any of the tissues to 1. (**c**) Relative protein levels of the 'specific' TFs, as measured by TMT-labeled MS from triplicate samples ($n = 3$), correlate partially with the corresponding TF activities, as measured by ATI motif analysis. The relative abundance of each TF protein was normalized by setting the highest abundance to 1. The bars indicate the mean of triplicate samples, and error bars represent the s.d. $^{**}P < 0.01$; $^{***}P < 0.001$; by two-sided Student's *t*-test. The abundance of EGR1 varies in the triplicate liver samples (marked red, see **Supplementary Table 9**), suggesting that its abundance varies naturally in mouse liver. (**d**) Label-free MS analysis showing that many TFs that were strongly active (black font; *y* axis, ATI motif counts) but also revealing high protein abundance (*x* axis) of TFs that were not strongly active in mouse liver, including DPF2, IRF3, NFY-B and SMAD4 (red font). The MS analysis also failed to detect some TFs whose motifs were recovered by ATI (blue font; RBPJ, DBP, ESRR and NRF1). Note that the same motif detected in ATI could be correlated with several different TFs detected in the MS analysis.

Transcription factors that determine cell fate are strongly active in cells

We next analyzed the relative enrichment of known TF motifs across the tissues by using motif enrichment analysis. This analysis confirmed the enrichment of the *de novo*–discovered motifs and revealed additional TFs whose motifs were specifically enriched in the different cell types or tissues (**Fig. 2a** and **Supplementary Table 6**). For instance, from the mES cells we detected motifs for additional pluripotency factors, such as GLIS2, although with a relatively low level of enrichment. From adult mouse liver, we also detected motifs specific for TFs such as ONECUT

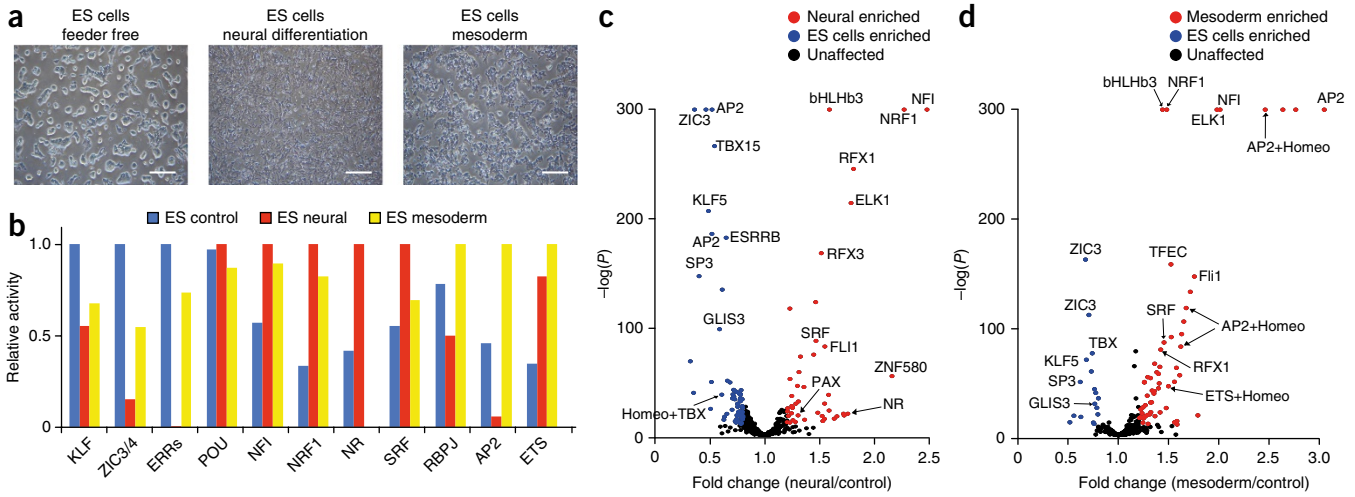

Figure 3 ATI analysis of transcription factor activities in differentiating ES cells. (**a**) Representative images showing the morphology of control mES cells (left) and of mES cells 2 d after they were induced to differentiate toward neural (middle) and mesodermal (right) lineages. The experiment was done in technical duplicates. Scale bars, 400 µm. (**b**) Comparison of the motifs detected by the *de novo* motif discovery method in the control and differentiated mES cells. Data from a single ATI experiment, which consisted of four enrichment cycles, is shown. Bars indicate the relative activities of the indicated TFs, based on increase of the absolute molecular counts[12] of each motif between the first cycle and the fourth cycle. The activities of each TF were normalized by setting the highest activity in any of the three conditions to 1. (**c**,**d**) Comparison of motif enrichment between the neural (read counts = 16,544,923) (**c**) or mesoderm (read counts = 16,720,589) (**d**) differentiated ES cells and the control ES cells (read counts = 14,306,415). The *y* axis indicates the *P* value (log scale, calculated by winflat[51]; due to the precision of calculation, many *P* values were set to a minimum of 1×10^{-300}); the *x* axis indicates fold change. The motifs with $P < 1 \times 10^{-10}$ and greater than 20% change are indicated in red (enriched in neural- or mesoderm-differentiated ES cells) or blue (enriched in control ES cells), respectively, with the names of representative motifs indicated. The black dots represent motifs that did not change less than 20% and/or did not pass the p-value threshold.

and HNF4A. Taken together, ATI analysis of mouse tissues revealed that in addition to five common TF activities, each tissue displayed strong activity of key regulators for the respective cell identities.

To test the role of ATI-identified TFs in cell fate determination, we transduced human fibroblasts with a combination of constructs encoding nine strongly active TFs from adult mouse liver and investigated the morphology of the cells and expression of the liver-specific marker albumin after 2 weeks of culture. The fibroblasts were converted to iHeps at an efficiency that was similar to the one observed for the most efficient previously described protocol (**Supplementary Fig. 3**), indicating that ATI could identify factors that induced *trans*-differentiation of mammalian cells.

We also detected enrichment of some unknown motifs (**Fig. 2a**, bottom), which we could not assign to a known TF based on current knowledge (HT-SELEX motifs, CIS-BP, TOMTOM[9,13,35,36]). Overall, we recovered 35 motifs, of which only six (17%) were unknown, indicating that specificities for most of the TFs that display strong activity in the tested tissue types have already been determined.

Transcription factor activity is not explained by protein abundance alone

To identify the TFs, we performed MS analyses in three adult mouse tissues: spleen, brain and liver. This analysis was performed by using high-resolution isoelectric focusing liquid chromatography -mass spectrometry (HiRIEF-LC-MS)[37], with relative quantification between samples using isobaric tags (TMT). Comparison of the MS and ATI data revealed that most of the TF proteins whose motifs were specific for one of the tissues were more abundant in that tissue than in the other two tissues (**Fig. 2c**). We also performed label-free MS analysis to estimate the protein levels within the samples. In this analysis, we detected some highly abundant TF proteins, whose motifs were not detected in the ATI assay (**Fig. 2d** and **Supplementary Table 7**). For example, no motif was discovered for DPF2, which is involved in apoptosis[38]. Similarly, SMAD proteins were abundant in all of the tissues, but their signature motif was not detected in any of the cases (**Fig. 2d** and **Supplementary Table 7**). This is consistent with ligand dependence and the low DNA-binding activity of SMADs ($K_d \approx 1 \times 10^{-7}$ M)[39]. Another class of signal-dependent TFs, the interferon regulatory factors (IRFs), were also abundant at the protein level but could not be detected as active by ATI.

In addition, the MS analysis indicated that some subunits of the multimeric TFs, such as NFY, were present at high levels, yet the TFs were not strongly active, which was consistent with the observation that there were lower levels of the other subunits (**Supplementary Table 7**). Of the 67 TFs that were detected in the liver by MS analysis, 40% were known to bind to an ATI-identified motif, 12% represented a known obligate heteromer or ligand-regulated TF, 33% were classified as TFs but did not have a known motif, and 15% had a known motif but were present at a relatively low abundance (**Supplementary Table 7**).

DNA-binding activities change during cell differentiation

To test whether ATI could detect changes in TF activities that were induced during cell differentiation, we induced differentiation of ES cells toward a neural or mesodermal lineage, using standard conditions[40,41] (**Fig. 3a** and Online Methods). This analysis revealed that ATI was able to detect several known quantitative changes in TF binding activities that accompany the neural differentiation process (**Fig. 3b**,**c**). For example, the activities of the pluripotency factors GLIS and ZIC decreased, whereas the activity of RFX and PAX factors, which are known to contribute to neural differentiation, increased. Similarly, the activities of GLIS and ZIC factors decreased after induction of mesodermal differentiation, whereas the activity of the known mesodermal factor AP2 increased dramatically (**Fig. 3b**,**d**). The activation

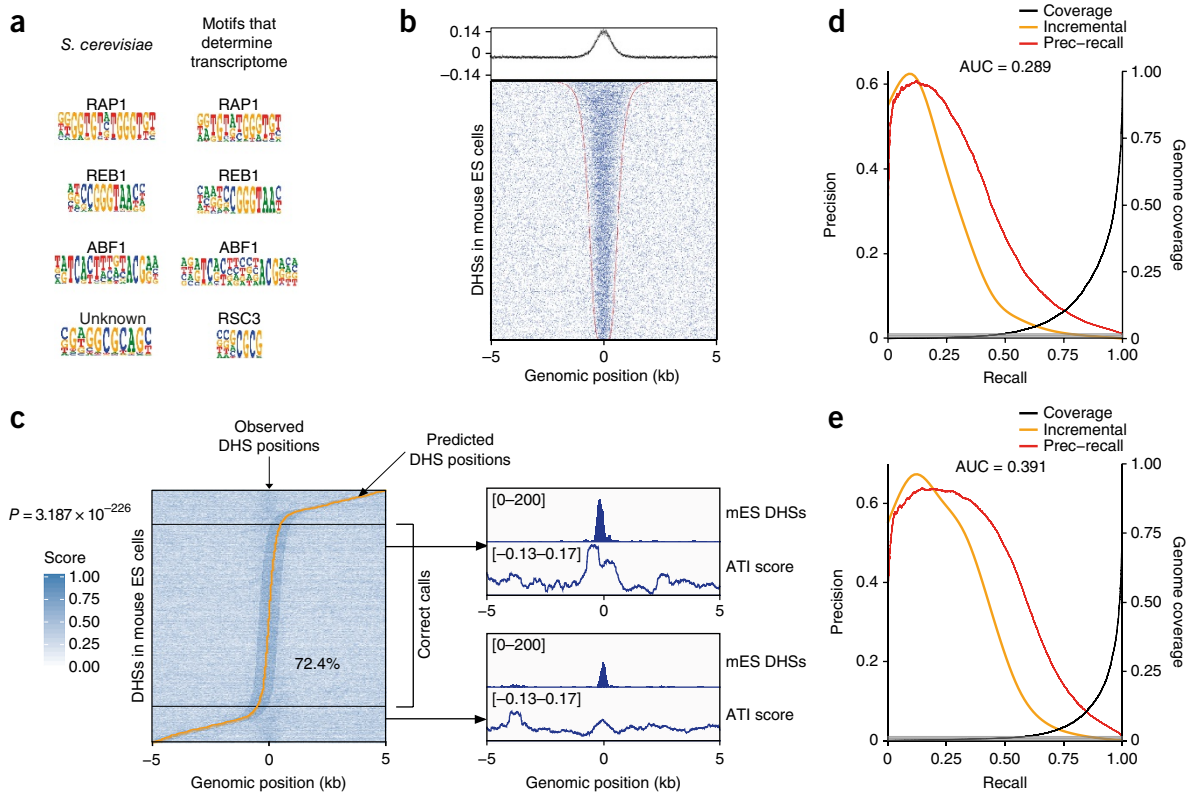

Figure 4 Strongly active TFs explain key features of transcription in yeast and mouse cells. (**a**) Comparison of motifs detected in the ATI assay (*S. cerevisiae*) and the four motifs that can be used to computationally identify yeast transcript start positions[44]. (**b**) ATI-enriched ten-mers from mES cells are also enriched in DHSs from mES cells. In the dot plot, each row indicates one DHS region from the mES cells that is flanked with genomic sequences. Red dots indicate the boundaries of the DHS regions; blue dots indicate the positions of the top 2,000 ATI-enriched ten-mers out of all ~1 million ten-mers. The graph on top shows the average of scores for each ten-mer at each position across the rows. (**c**) Prediction of ES cell DHS regions by using the ten-mer data from the ATI assay. DHSs (n = 3,907) were sorted by position of the prediction call (yellow line) (left). Black horizontal lines separate accurate DHS calls (in the middle of the plot) from calls >500 bp off the known DHS center, which is located at the x axis position 0 in all cases. The fraction of predictions within ±500 bp of the center (72.4%), and the corresponding P value (as determined by Winflat[51]) for the null model in which position calls are randomly distributed is also indicated. Two representative tracks for the DHS and ATI signals are also shown (right). (**d**,**e**) Comparison between the genome-wide predictions of the ES cell DHS regions using ten-mer data from the ATI assay (**d**) and the DHSs themselves (**e**). Precision-recall curves (red lines, cumulative precision) indicate that the ATI ten-mers could be used to predict the DHS positions (area under the curve (AUC) = 0.29) and that the performance of the ATI predictor was relatively close to a predictor that used ten-mer data from the DHSs themselves (AUC = 0.39). Yellow and black lines, respectively, show smoothed non-cumulative (incremental) precision and fraction of the genome selected at the indicated recall level. Gray shading indicates fraction of true DHSs in the genome (0.9%).

of SMAD proteins by BMP4 and activin A, which were used to induce mesodermal differentiation, was not detected, potentially due to the fact that SMAD proteins bind DNA only weakly and often act together with other TFs[39]. In contrast, ATI robustly detected the activation of retinoic acid receptor by the neural inducer retinoic acid (**Fig. 3b**,**c**), indicating that some ligand-gated TFs can be detected by ATI.

Conservation of transcription factor–binding activities between species

One of the notable advantages of the ATI assay is that it can be performed by using any type of protein extract from any species. To analyze how similar active TFs are between organisms, we performed ATI experiments with nuclear extracts from fruit fly (*Drosophila melanogaster*) S2 cells and the yeasts *Saccharomyces cerevisiae* and *Schizosaccaromyces pombe*. Analysis of the sequencing data using the *de novo* motif discovery method indicated that ATI could identify the most active TFs in all of these species. Several of the recovered motifs matched known motifs from the respective species[36]. From the study of the yeast *S. cerevisiae*, we detected ABF1, RAP1 and REB1. Notably, of six motifs (five common motifs and one shared motif for RBPJ) that were common to almost all mouse tissues and cell lines, two TFs, RBPJ (CBF11) and the bHLH factor TFE (CBF1), were highly active in the yeast *S. pombe*, and two TFs, TFE (CBF1) and the bZIP factor CREB (CST6), were highly active in *S. cerevisiae*. Although there are only 23 human TF families, the number of distinct motifs is much larger; for example, in humans, there are more than 400 different binding motifs[42], with at least 30 distinctly different motifs for bHLH factors and ten for bZIP factors[9,43]. Thus, the fact that the same motifs were highly active in distant species suggests that the dominant mechanisms of transcriptional regulation may have been conserved during the evolution of eukaryotes. In the different species, we also found many motifs that we could not assign to a known TF on the basis of existing TF specificity databases (**Supplementary Fig. 4**).

Identification of master transcription factors in yeast

One concern with the ATI assay is that it could only identify specific TFs that bound strongly under the *in vitro* conditions used in the assay. To independently validate the assay, we analyzed whether the

most active DNA-binding TFs identified by the method were also the most active *in vivo*. We first compared the ATI results from the yeast *S. cerevisiae* with the four known TF motifs that determine the yeast transcriptome[44]. Of the four motifs that are required to build a model describing yeast transcript start positions, ATI identified three—those for ABF1, RAP1 and REB1—and, in addition, recovered CG-rich motifs that were related to the fourth motif used in the study (that for RSC3; **Fig. 4a**). This result indicates that ATI can identify a complete set of TFs crucial for determining the transcriptional state of yeast.

Transcription factor–binding activities contribute to chromatin landscape

To determine whether ATI could also confer information on the mammalian chromatin landscape, we compared the ATI data with DHSs from the mouse ENCODE project[45]. This analysis revealed that the top 2,000 ten-mers detected by ATI in mES cells, heart, spleen, brain and liver were strongly enriched in the ~5,000 most-significant DHS regions from the respective tissues. As expected, the strongest enrichment was seen in ES cells, as they are more homogenous than tissues containing multiple different cell types (**Fig. 4b** and **Supplementary Fig. 5**). Analysis of the ten-mers enriched in ES cell DHSs and ATI revealed that there were many ten-mers that were enriched in both and that all of these ten-mers were related to the ATI motifs (**Supplementary Fig. 6**). Consistent with the observed enrichment of ATI ten-mers in DHSs, ATI also enriched many DHS sequences from a mouse genomic library (**Supplementary Fig. 7**). However, some ten-mers were enriched only in DHSs (**Supplementary Fig. 6**); these included many repetitive CG-rich sequences that were enriched in gene regulatory elements. Because DHSs represent gene regulatory elements, they are expected to be enriched with both motifs that contribute to the opening of the chromatin and motifs that are involved in downstream activities, such as transactivation or repression of RNA polymerase II. Consistent with this, *de novo* motif discovery analysis of DHSs revealed some motifs that were not enriched by ATI. These included a motif similar to that of ZNF-143 (**Fig. 1b**); this motif has been reported to contribute to interactions between promoters and distal regulatory elements[46].

We further hypothesized that if ATI accurately represented TF-binding activities in cells and revealed subsequences that bound strongly to TFs also *in vivo*, then it would be possible to predict the DHS regions by using the ATI-enriched subsequences. It has previously been shown that DHSs can be predicted on the basis of sequence features from different types of experimental data (for example, DNase-seq data[47] or ChIP-seq data[48]). It is also well established that DHSs and TF-binding clusters are enriched with matches to biochemically obtained TF motifs[49,50] and that they overlap with *in silico*–predicted clusters that are called based on TF motif matches only. However, in our recent study, only ~30% of TF-binding clusters could be predicted on the basis of monomeric-TF-binding models[50], suggesting that additional unknown determinants affected TF binding to DNA inside cells. To determine whether ATI improved on the predictions, we developed a predictor based on the enrichment rank of all ten-mers in ATI. This analysis revealed that >70% of the DHSs could be predicted by the ten-mers derived solely from ATI (**Fig. 4c**; 10% expected by random; $P < 3.2 \times 10^{-226}$ by winflat[51]), indicating that ATI-derived ten-mer enrichment more accurately represented TF activity in cells as compared to that from any other presently available direct biochemical information. In the prediction of the DHSs genome wide, ATI was nearly as effective as using ten-mers from the DHSs themselves (**Fig. 4d**,**e**), indicating that the ATI data contained a substantial fraction of the motif information included in the DHSs, despite the fact that the DHSs were expected to contain additional motif features that related to their functionality in gene regulation and not to their open-chromatin status. Consistently, analysis of DHSs that were hard to predict with ATI ten-mers but that could be predicted with DHS ten-mers again revealed the ZNF-143-like motif (**Supplementary Fig. 8d**). In contrast, DHSs that were hard to predict with both types of ten-mers did not contain enriched sequence motifs (**Supplementary Fig. 8d**), suggesting that their DNase I hypersensitivity could be caused by longer sequence features, such as those affecting intrinsic nucleosome affinity[52]. In summary, both the yeast transcriptome model and the DHS data verified that the TFs found with ATI were active in the cell types and tissues analyzed and contributed to the positioning of transcription start sites (TSSs) in yeast and to accessibility of chromatin in mouse cells.

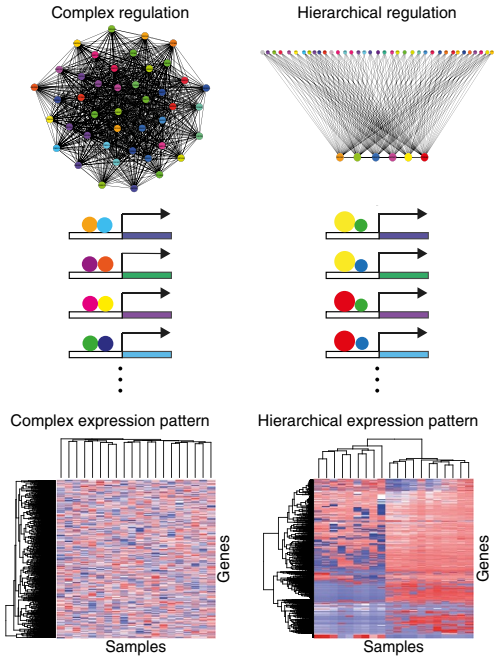

Figure 5 A hierarchical gene regulatory network leads to a hierarchical gene expression profile. As compared to a complex gene regulatory network formed from equally active TFs (left), the hierarchical gene regulatory network formed by using TFs that are either strongly (large circles) or weakly (small circles) active (right) is simpler and yields a gene expression pattern that is similar to the hierarchical gene expression patterns observed in real biological samples. The heat map for complex expression pattern was generated from an artificial random matrix containing 20 'samples' and 2,000 'genes'; the heat map for hierarchical expression pattern was generated from RNA-seq data from 20 different cell and tissues samples in the ENCODE project[45]. In this model, the genes clustered into each group are regulated by the same dominant TFs in collaboration with different weak TFs, and the clustered tissue samples share some dominant TFs.

DISCUSSION

It is not possible to determine the binding activities of TFs in different types of cells solely on the basis of RNA expression or protein abundance data, because different TFs have different binding specificities and affinities for DNA. Moreover, the binding activity of TFs is often regulated at the level of posttranslational modifications, protein–protein interactions[53] and nuclear localization. For this reason, we developed the ATI assay, which directly measures

the DNA-binding activities of all TFs in cell extracts and detects the changes in the activities in response to stimuli. In addition, comparison of TF activities between different cell types can yield useful hypotheses about key TFs that determine the identity of the cell types. Independent verification of the results by using MS and prediction of functional features validated ATI's ability to capture the majority of strong DNA-binding activities in cells. However, the sensitivity of ATI in being able to for detect ion of accessory DNA-binding factors and proteins that bind to for DNA weakly, such as the SMAD proteins, may be low. Further studies that miniaturize and standardize the process are expected to further improve the sensitivity and accuracy of this widely applicable method.

Our findings for the tissue specificity of TF activity in mouse tissues, the conservation of TF activities between different species and the dynamics and functional importance of strongly active TFs during differentiation suggest that strongly active TFs have an important and active role in cell differentiation and cell fate determination. We also found that by using the ATI-derived binding information we were able to predict positions of open chromatin using only biochemical parameters far more accurately than what has previously been possible. This result indicates that lack of knowledge of TF DNA-binding activity levels was a major unknown factor that hindered previous computational predictions of regulatory elements. However, our results cannot be interpreted to mean that open chromatin results exclusively from the action of TFs with strong binding activity. It is well known that some TFs can directly or indirectly recruit enzymes that remodel or modify nucleosomes to generate open chromatin and/or derepress closed chromatin states[54,55]. It should also be noted that a large number of binding sites for less-active TFs, or site(s) for combinations of cooperatively bound TFs, can also bind with sufficient energy to displace or compete out nucleosomes. These mechanisms can open a subset of DHSs, but they are unlikely to be the predominant way to open chromatin; if that was the case, then the binding motifs for the cooperative or weak binders would have been detected when conducting the *de novo* motif mining of the DHS regions.

On the one hand, our results indicate that the few TFs that have strong DNA-binding activities in a cell have a major role in setting its overall gene regulatory architecture. On the other hand, ChIP-seq analyses have clearly shown that a large number of TFs can bind open chromatin regions in cells[9,10]. These observations are consistent with a model in which the TFs that are strongly active in DNA binding set up the overall chromatin state of cells, and that the ability of TFs with weaker DNA-binding activity to regulate their target genes is conditional on this chromatin state.

As compared to a complex model in which different TFs are equally active and can collaborate with each other, such a hierarchical gene regulatory model is far simpler and can explain the fact that hierarchical gene expression patterns are commonly observed in analyses of real biological systems (**Fig. 5**). This model also provides a simple combinatorial gene regulation system. If the TF that has strong DNA-binding activity lacks a strong transactivation or repression domain, then it will require a partner that has such a domain to regulate gene expression. It should be noted that different types of activation domains will also further contribute to such combinatorial regulation[56], increasing the number of cooperation partners to three or more.

The biochemical activity–based distinction between TFs that we identified here is related, but not identical, to the concept of 'pioneer' TFs that are able to bind to nucleosomal DNA[57]. The ability to compete against nucleosomes could either be simply due to mass action or due to a specific ability of some TFs to bind to nucleosomal DNA with fast kinetics and recruit nucleosome remodeling enzymes[54,55]. In bioinformatic studies to identify factors occupying most of their consensus motifs[58], both types of factors would be detected equally. Comparison of our data with factors that occupy their motifs[58] revealed that many such factors displayed strong DNA-binding activity in the ATI assay. The ability to predict DHS positions based on ATI data suggests that much of the nucleosome-competing activity in ES cells is due to TFs that are present in such a high abundance relative to their K_d values that they effectively and specifically compete against nucleosome binding. Motif discovery analysis from the DHSs that we could not predict by using ATI data also failed to identify motifs that corresponded to previously identified pioneer factors that can access nucleosomal DNA[59,60] (**Supplementary Fig. 8d**), suggesting that such activities either do not rely on complex sequence motifs[60] or do not occur at so many positions that they would be detectable by motif mining.

In summary, we present a method to determine TF activity in cells and tissues, and our findings suggest that the cellular transcriptional regulatory network may be much simpler than previously thought.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank J. Yan, E. Kaasinen, B. Schmierer and Y. Yin for critical review of the manuscript, and S. Augsten, L. Hu and P. Pandey for technical assistance. This work was supported by the Center for Innovative Medicine at the Karolinska Institutet (2015–2017; J.T.), the Knut and Alice Wallenberg Foundation (KAW 2013.0088; J.T.), the Göran Gustafsson Foundation (2011–2013; J.T.) and the Swedish Research Council (Vetenskapsrådet; Rådsprofessorprogrammet D0815201; J.T.).

AUTHOR CONTRIBUTIONS

I.S. collected mouse tissue samples; B.W. extracted proteins, performed ATI experiments and analyzed the data; F. Zhong and F. Zhu performed the DHS analysis; B.S. performed the iHep reprogramming experiment; L.M.O. and J.L. performed the MS experiments and data analysis; A.J., T.K. and M.T. helped to supervise the project or related experiments; and B.W. and J.T. wrote the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

1. Boyer, L.A. *et al.* Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947–956 (2005).
2. Chen, X. *et al.* Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106–1117 (2008).
3. Wang, J. *et al.* A protein interaction network for pluripotency of embryonic stem cells. *Nature* **444**, 364–368 (2006).
4. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676 (2006).
5. Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861–872 (2007).
6. Feng, B. *et al.* Reprogramming of fibroblasts into induced pluripotent stem cells with orphan nuclear receptor ESRRB. *Nat. Cell Biol.* **11**, 197–203 (2009).
7. Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A. & Luscombe, N.M. A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* **10**, 252–263 (2009).
8. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
9. Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152**, 327–339 (2013).
10. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

11. Garber, M. *et al.* A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. *Mol. Cell* **47**, 810–822 (2012).
12. Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9**, 72–74 (2011).
13. Nitta, K.R. *et al.* Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *eLife* **4** (2015).
14. Lim, L.S. *et al.* ZIC3 is required for maintenance of pluripotency in embryonic stem cells. *Mol. Biol. Cell* **18**, 1348–1358 (2007).
15. Loh, Y.H. *et al.* The OCT4 and NANOG transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.* **38**, 431–440 (2006).
16. Ivanova, N. *et al.* Dissecting self-renewal in stem cells with RNA interference. *Nature* **442**, 533–538 (2006).
17. Jiang, J. *et al.* A core KLF circuitry regulates self-renewal of embryonic stem cells. *Nat. Cell Biol.* **10**, 353–360 (2008).
18. Nichols, J. *et al.* Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor OCT4. *Cell* **95**, 379–391 (1998).
19. Jolma, A. *et al.* DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527**, 384–388 (2015).
20. Tun, T. *et al.* Recognition sequence of a highly conserved DNA-binding protein RBP-Jκ. *Nucleic Acids Res.* **22**, 965–971 (1994).
21. Wei, G.H. *et al.* Genome-wide analysis of ETS-family DNA-binding *in vitro* and *in vivo*. *EMBO J.* **29**, 2147–2160 (2010).
22. Wikström, L. *et al.* Abnormal heart rate and body temperature in mice lacking thyroid hormone receptor–α1. *EMBO J.* **17**, 455–461 (1998).
23. Adams, B. *et al. Pax5* encodes the transcription factor BSAP and is expressed in B lymphocytes, the developing CNS and adult testis. *Genes Dev.* **6**, 1589–1607 (1992).
24. Urbánek, P., Wang, Z.Q., Fetka, I., Wagner, E.F. & Busslinger, M. Complete block of early B cell differentiation and altered patterning of the posterior midbrain in mice lacking PAX5 (BSAP). *Cell* **79**, 901–912 (1994).
25. Mach, B., Steimle, V., Martinez-Soria, E. & Reith, W. Regulation of MHC class II genes: lessons from a disease. *Annu. Rev. Immunol.* **14**, 301–331 (1996).
26. Poirier, R. *et al.* Distinct functions of *Egr* gene family members in cognitive processes. *Front. Neurosci.* **2**, 47–55 (2008).
27. Nakakura, E.K. *et al.* Mammalian Scratch: a neural-specific Snail family transcriptional repressor. *Proc. Natl. Acad. Sci. USA* **98**, 4010–4015 (2001).
28. Sugitani, Y. *et al.* BRN-1 and BRN-2 share crucial roles in the production and positioning of mouse neocortical neurons. *Genes Dev.* **16**, 1760–1765 (2002).
29. Wang, N.D. *et al.* Impaired energy homeostasis in C/EBP-α-knockout mice. *Science* **269**, 1108–1112 (1995).
30. Frain, M. *et al.* The liver-specific transcription factor LF-B1 contains a highly diverged homeobox DNA-binding domain. *Cell* **59**, 145–157 (1989).
31. Pontoglio, M. *et al.* Hepatocyte nuclear factor 1 inactivation results in hepatic dysfunction, phenylketonuria and renal Fanconi syndrome. *Cell* **84**, 575–585 (1996).
32. Fonjallaz, P., Ossipow, V., Wanner, G. & Schibler, U. The two PAR leucine zipper proteins TEF and DBP display similar circadian and tissue-specific expression but have different target promoter preferences. *EMBO J.* **15**, 351–362 (1996).
33. Du, Y. *et al.* Human hepatocytes with drug metabolic function induced from fibroblasts by lineage reprogramming. *Cell Stem Cell* **14**, 394–403 (2014).
34. Huang, P. *et al.* Direct reprogramming of human fibroblasts to functional and expandable hepatocytes. *Cell Stem Cell* **14**, 370–384 (2014).
35. Weirauch, M.T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).
36. Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L. & Noble, W.S. Quantifying similarity between motifs. *Genome Biol.* **8**, R24 (2007).
37. Branca, R.M. *et al.* HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nat. Methods* **11**, 59–62 (2014).
38. Gabig, T.G., Mantel, P.L., Rosli, R. & Crean, C.D. Requiem: a novel zinc-finger gene essential for apoptosis in myeloid cells. *J. Biol. Chem.* **269**, 29515–29519 (1994).
39. Gaarenstroom, T. & Hill, C.S. TGF-β signaling to chromatin: how SMADs regulate transcription during self-renewal and differentiation. *Semin. Cell Dev. Biol.* **32**, 107–118 (2014).
40. Zhang, J. *et al.* Retinoic acid induces embryonic stem cell differentiation by altering both encoding RNA and microRNA expression. *PLoS One* **10**, e0132566 (2015).
41. Kokkinopoulos, I. *et al.* Cardiomyocyte differentiation from mouse embryonic stem cells using a simple and defined protocol. *Dev. Dyn.* **245**, 157–165 (2016).
42. Lambert, S.A. *et al.* The human transcription factors. *Cell* **172**, 650–665 (2018).
43. Yin, Y. *et al.* Impact of cytosine methylation on DNA-binding specificities of human transcription factors. *Science* **356**, eaaj2239 (2017).
44. de Boer, C.G. *et al.* A unified model for yeast transcript definition. *Genome Res.* **24**, 154–166 (2014).
45. Yue, F. *et al.* A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355–364 (2014).
46. Bailey, S.D. *et al.* ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. *Nat. Commun.* **2**, 6186 (2015).
47. Noble, W.S., Kuehn, S., Thurman, R., Yu, M. & Stamatoyannopoulos, J. Predicting the *in vivo* signature of human gene regulatory sequences. *Bioinformatics* **21** (Suppl. 1), i338–i343 (2005).
48. Lee, D., Karchin, R. & Beer, M.A. Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res.* **21**, 2167–2180 (2011).
49. Thurman, R.E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
50. Yan, J. *et al.* Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell* **154**, 801–813 (2013).
51. Audic, S. & Claverie, J.M. The significance of digital gene expression profiles. *Genome Res.* **7**, 986–995 (1997).
52. Kaplan, N. *et al.* The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**, 362–366 (2009).
53. Chronis, C. *et al.* Cooperative binding of transcription factors orchestrates reprogramming. *Cell* **168**, 442–459.e20 (2017).
54. Fryer, C.J. & Archer, T.K. Chromatin remodeling by the glucocorticoid receptor requires the BRG1 complex. *Nature* **393**, 88–91 (1998).
55. Li, Z. *et al.* FOXA2 and H2A.Z mediate nucleosome depletion during embryonic stem cell differentiation. *Cell* **151**, 1608–1616 (2012).
56. Stampfel, G. *et al.* Transcriptional regulators form diverse groups with context-dependent regulatory functions. *Nature* **528**, 147–151, 10.1038/nature15545 (2015).
57. Zaret, K.S. & Mango, S.E. Pioneer transcription factors, chromatin dynamics and cell fate control. *Curr. Opin. Genet. Dev.* **37**, 76–81 (2016).
58. Sherwood, R.I. *et al.* Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat. Biotechnol.* **32**, 171–178 (2014).
59. Cirillo, L.A. *et al.* Opening of compacted chromatin by early developmental transcription factors HNF3 (FOXA) and GATA-4. *Mol. Cell* **9**, 279–289 (2002).
60. Soufi, A. *et al.* Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell* **161**, 555–568 (2015).

ONLINE METHODS

Cell culture and protein extraction. mES cells (C57BL/6J; from the KCTT Center at Karolinska Institutet) were thawed, plated in ES1+LIF (composition below) medium and cultured in 2i+LIF medium (composition below) without mouse embryonic fibroblast (MEF) feeder layers (first set, corresponding to **Figs. 1** and **3**) or on low-density MEF feeder layers (277,000 irradiated MEFs per 60-mm dish; corresponding to **Fig. 2**) until 70–80% confluence was achieved. Cells were passaged every 2 or 3 d and collected by trypsinization, and the MEF feeder cells were removed by means of the differential adhesion method. The ES1+LIF (250 ml) medium included: 204 ml knockout Dulbecco's modified Eagle's medium (DMEM; Gibco, cat. no. 10829-018), 37.5 ml FBS (ES qualified, Sigma, cat. no. F7524), 2.5 ml of 200 mM L-glutamine (Gibco, cat. no. 25030-024), 2.5 ml of 1 M HEPES (Gibco, cat. no. 15630-056), 2.5 ml of 100× non-essential amino acids (Gibco, cat. no. 11140-035), 0.5 ml of 50 mM β-mercaptoethanol (Thermo Scientific, cat. no. 31350-010), 0.25 ml of 10 mg/ml gentamicin (Thermo Scientific, cat. no. 15710-049) and 0.25 ml leukemia inhibitory factor (LIF; 1×10^6 U/ml stock, Millipore, cat. no. ESG1107). The 2i+LIF medium (50 ml) included: 38.785 ml knockout DMEM, 10 ml KnockOut Serum Replacement (Gibco, cat. no. 10828-028), 0.5 ml of 200 mM L-glutamine (Gibco, cat. no. 25030-024), 0.5 ml of 100× non-essential amino acids, 0.1 ml of 50 mM β-mercaptoethanol, 50 µl of 10 mg/ml gentamicin, 50 µl LIF (1×10^6 U/ml stock), 1 µM of the mitogen-activated protein kinase (MEK) inhibitor PD0325901 (Miltenyi Biotec, cat. no. 130-103-923) and 2 µM of the glycogen synthase kinase (GSK)-3α–GSK-3β inhibitor BIO (Sigma, cat. no. B1686). *Drosophila* S2 cells were cultured in Schneider's *Drosophila* medium (Thermo Scientific, cat. no. 21720024) at 27 °C without CO_2 and were collected by trypsinization. Collected cells were washed once with ice-cold PBS.

For the differentiation of mES cells to different lineages, the ES cells were thawed and plated in ES1+LIF medium and cultured in 2i+LIF medium without feeder layers, and then the cells were split to several plates for different treatments. The control ES cells were cultured in 2i+LIF medium without feeder layers; the ES cells for neural differentiation were cultured with 2i medium supplemented with 2 µM retinoic acid for 2 d; the ES cells for mesodermal differentiation were first cultured in 2i+LIF medium for 16 h, and then changed to the mesodermal medium (composition below) for 30 h. The mesodermal medium (206 ml) included: 100 ml Iscove's modified Eagle's medium (IMDM) supplemented with GlutaMAX (Thermo Scientific, cat. no. 31980030), 100 ml Ham's F-12 nutrient mix (Thermo Scientific, cat. no. 21765029), 2 ml of 100× N2 supplement (Thermo Scientific, cat. no. 17502048), 4 ml of 50× B27 supplement (Thermo Scientific, cat. no. 17504044), 0.5 mM ascorbic acid (Sigma, cat. no. A92902), 4.5×10^{-4} M monothioglycerol (Sigma, cat. no. M1753), 5 ng/ml vascular endothelial growth factor (VEGF; Thermo Scientific, cat. no. PHC9391), 8 ng/ml activin A (Thermo Scientific, cat. no. PHG9014) and 0.5 ng/ml BMP4 (Thermo Scientific, cat. no. PHC9534).

All of the tissues were from four 1-year-old C57BL/6J male mice, one for the ATI assay and the three others for the MS. The heart and spleen samples were collected and then cut into small pieces (~4 mm) before protein extraction; the liver and brain samples were lysed directly without cutting. The soluble nuclear proteins in cells or tissues were extracted by using the Subcellular Protein Fractionation Kit for Tissues (Life Technologies, cat. no. 87790). Ice-cold CEB buffer (1 ml per 100 mg tissue sample or 1 ml per 1 million cells) complemented with protease inhibitors (Roche, cat. no. 05892791001) and phosphatase inhibitors (Roche, cat. no. 04906845001) was added, followed by Dounce homogenization. Homogenized samples were transferred through a strainer into a clean tube, and the tube was then centrifuged at 500*g* for 5 min. Subsequently the supernatant was discarded, and ice-cold MEB buffer with protease and phosphatase inhibitors was added to extract the membranes of cells, followed by centrifugation at 3,000*g* for 5 min. The supernatant was removed, and detergent-free NEB buffer with inhibitors was added to the remaining nuclear pellet; the sample was vortexed for 15 s and incubated at 4 °C for 45 min with gentle mixing. The supernatant after lysing with NEB buffer was collected, supplemented with glycerol (5% vol/vol) and stored at −80 °C in aliquots for future use. All cells were tested regularly for Mycoplasma infection. All mouse work was conducted after approval by the ethics committee of the Board of Agriculture, Experimental Animal Authority, Stockholm South, Sweden (Dnr S50/13, S11/15 and S16/15).

Lentivirus production and generation of iHeps. Sequences encoding the full-length ORFs were cloned into the pLenti6/V5 lentiviral expression vector using the Gateway recombination system. Viruses were generated by co-transfection of the expression vectors with the packaging vectors psPAX2 and pMD2.G (Addgene) into 293FT (Thermo Scientific, cat. no. R70007) cells with Lipofectamine 2000 (Thermo Fisher Scientific). The following day, the cells were replenished with fresh culture medium, and virus-containing medium was collected after 48 h. The virus was concentrated using a Lenti-X concentrator (Clontech).

The human fibroblast cell line CCD-1112Sk was obtained from the ATCC (#CRL 2429) and cultured in fibroblast medium containing DMEM plus 10% FBS with antibiotics. Early-passage fibroblasts were seeded on day 0 and transduced on day 1 with constructs encoding cocktails of TFs as previously reported in studies from Morris *et al.*[61] (FOXA1, HNF4A, KLF5), Du *et al.*[33] (HNF4A, HNF1A, HNF6, ATF5, PROX1, CEBPA) and Huang *et al.*[34] (FOXA3, HNF4A, HNF1A), and with the nine specific TFs identified by ATI from mouse liver (HNF1A, HNF1B, DBP, MAFG, CEBPA, CEBPB, HNF4A, HNF6 (ONECUT1) and ESRRA). The transduction was performed as two independent experiments overnight in the presence of 8 µg/ml polybrene. The virus-containing medium was replaced the following morning with fresh fibroblast medium containing β-mercaptoethanol. On day 3, the cells were changed to a defined hepatocyte growth medium (HCM, Lonza). On day 7, the cells were re-plated on type I collagen–coated plates in HCM in several technical replicates, and thereafter, the HCM was changed every second day. On day 29, the cells were passaged to new type I collagen–coated plates and cultured until 6 weeks after transduction.

The cells were harvested for each condition at several time points for total RNA isolation followed by cDNA synthesis using the Transcriptor High-fidelity cDNA synthesis kit (Roche) and real-time PCR using SYBR green (Roche) for primers specific for the transcripts of the housekeeping gene *GAPDH* and albumin. The albumin C_t values were normalized to those for *GAPDH*, and the mean values of sample replicates were shown for different conditions at the indicated time points.

Active transcription factor identification assay. Protein extract (4 µg) was incubated with 5 µl barcoded double-stranded DNA oligos containing 40 bp random sequences (10 pmol, 900 ng), together with poly-deoxy-inosinic-deoxy-cytidylic acid (poly-dIdC) as a competitor (80 ng) in 1× binding buffer (140 mM KCl, 5 mM NaCl, 1 mM K_2HPO_4, 2 mM $MgSO_4$, 100 µM EGTA and 3 µM $ZnSO_4$, in 20 mM HEPES, pH 7.5) at room temperature for 30 min. After incubation, an EMSA was performed on ice for 70 min by using a 6% DNA retardation gel (Invitrogen, cat. no. EC63652BOX) in 0.5× TBE buffer (1 mM EDTA in 45 mM Tris-borate, pH 8.0) with 106 V constant voltage. The gel was then dyed with SYBR gold fluorescence dye for 10–20 min and washed with milliQ water. Fragments that migrated above the 300-bp marker were collected and eluted in TE buffer (1 mM EDTA in 10 mM Tris-Cl, pH 8.0), followed by incubation at 65 °C for 3 h. PCR was performed using Phusion polymerase (Thermo Scientific cat. no. F530L) to amplify the eluted DNA oligos for 20 cycles with 4 pmol of each primer, using a Bio-Rad S1000 Thermal Cycler with the following settings: initial denaturation 97 °C for 60 s, denaturation 97 °C for 15 s, annealing 65 °C for 15 s, elongation 72 °C for 40 s, final elongation 72 °C for 180 s. An additional 4 pmol of primers were added before the last cycle of PCR with a 20-min elongation time to convert the remaining single-stranded DNA into double-stranded DNA. The PCR product was then incubated again with the same extract, and the cycle was repeated. After three or four cycles of enrichment, PCR products bearing different barcodes were pooled and purified with the QIAquick PCR Purification Kit (Qiagen, cat. no. 28106) for next-generation sequencing (NGS) library preparation. NGS was performed with HiSeq 2000 or 4000 instruments (Illumina). The sequencing data from different cycles were compared with each other to determine the enrichment of specific motifs that related to the overall DNA-binding activities of specific TFs. Note also that because multiple cycles of enrichment were used, and the individual cycles of ATI were independent of each other, the cycles could be separately analyzed to determine reproducibility of enrichment of specific motifs (for example, compare **Fig. 2b** and **Supplementary Fig. 2b**). Analysis of the amount of DNA recovered from the gel, and the absolute number of motifs recovered during

doi:10.1038/nbt.4138

sequencing revealed that TFs from one nucleus bound approximately 5×10^6 DNA ligands, of which 3×10^5 represented specific binding events with clearly identifiable motif. This estimate was broadly consistent with an earlier estimate of TF abundance[62].

In principle, ATI analysis using 1 µg of the 40-bp random oligonucleotides (consisting of more than 6×10^{12} DNA ligands) can identify exact sequences that are approximately 20 bp long, or redundant sequences that consist of ~40 bits of information content. Most known TF motifs are well below this limit, with the exception of long arrays of zinc fingers found in repressor proteins that suppress mobile genetic elements[63,64]. Motifs for some of these proteins cannot be identified by ATI due to lack or extreme rarity of the potentially bound sequences in the initial library pool. To address this, we also ran ATI using fragmented mouse genomic sequences.

ATI assay using genomic fragments. The genomic DNA extracted from mES cells was sheared to make fragments of ~150 bp in size. Then, 270 ng of fragmented genomic DNA was incubated with 5 µg mES cell nuclear extract, together with 270 ng poly-dIdC, in 20 µl of 1× binding buffer. After incubation, an EMSA was performed on ice for 70 min using a 6% DNA retardation gel (Invitrogen, cat. no. EC63652BOX) in 0.5× TBE buffer (1 mM EDTA in 45 mM Tris-borate, pH 8.0) with 106 V constant voltage. Fragments that were >450 bp ('bound') and between 150 bp and 300 bp ('unbound') were collected and prepared for Illumina sequencing using the NEBNext Ultra DNA Library Prep Kit for Illumina (New England Biolabs, cat. no. E7370S). Peak calling for bound and unbound DNA fragments (setting original sheared DNA as background) was carried out using the MACS[65] program. The peaks from both samples were separately compared with DHSs from mES cells by using the BEDOPS program[66].

Synthetic DNA library design. The design of the synthetic DNA library was based on our previous SELEX results[9,19] that included monomeric and dimeric motifs for TFs. First, a dominating set of motifs, which consisted of 921 position-weight matrices (PWMs)[9,19], was extracted from the motifs. Subsequently, the seeds of these motifs were reformatted to include only five different IUPAC nucleotide ambiguity codes, A, T, C, G and N. A set of sequences containing the consensus seed, and seeds with each individual defined base replaced by an N, were then generated, which resulted in a total of 13,847 sequences. These sequences were then flanked with eight different sets of background sequences in such a way that the total length of all sequences was 35 nt. The background sequences were derived from the human genome that did not contain TF-binding sites (based on ChIP-seq data), exons or high-affinity matches to any of the 921 PWMs. Finally, standard Illumina adapters and three undefined bases were added to each sequence to generate unique molecular identifiers (UMIs). The DNA library, which consisted of single-stranded DNA, was then amplified by PCR to make double-stranded DNA for the experiment.

ATI data analysis. Two methods were used to identify the most important TFs in different cell types. The *de novo* motif discovery method was based on the 'Autoseed' program[13]. In Autoseed, seed sequences representing subsequences whose counts are higher than any other closely related subsequence (using the Huddinge distance metric[13]) are used as seeds. This method is based on direct counts of subsequences (enrichment relative to random sequence), and not on direct comparison between selected and unselected ten-mer sequences, as the latter approach would increase noise due to the low counts of all ten-mer sequences in the unselected library. The method can identify seeds that are separated by a Hamming distance of two or more. Up to 200 highest-count local-maxima 10-bp sequences (with or without a gap at the center) were used as seeds to generate the initial PWM motifs, which were then investigated manually to remove low-complexity motifs and motifs that were highly similar. Background correction was performed by using the subtractive method described in Jolma *et al.*[67]. To facilitate comparison of similar motifs, logos were generated in such a way that the frequency of each base was directly proportional to the height of the corresponding letter; because the absolute molecular counts for most motifs ranged from 10,000 to 100,000, even relatively small differences between motifs were statistically significant; therefore, the motifs were not rescaled based on information content. Counts of motifs were assigned based on the number of reads that matched the seeds in cycle 4 minus those in cycle 0. Because >93% of the reads were unique in each case, all of the reads were used to estimate the absolute number of molecular events. ATI could measure all TF activities, but its sensitivity was limited by the number of observed binding events as compared to the background occurrence of the motif in the random sequence population. In practice, activities that were >500-fold lower than the maximal activity identified in a particular cell type were commonly detected.

Known-motif enrichment analysis was used to study the enrichment of known motifs. First the number of reads for known motifs were counted with the MOODS program[68,69] before and after enrichment based on a particular cutoff value ($P \leq 0.0001$; score > 11). Subsequently, the enrichment and P value (Winflat[51]) were calculated for each motif; the sensitivity to detect differences by using this method was very high, and it could detect highly statistically significant differences whose fold changes were probably too low to be biologically meaningful. Due to this reason, we also reported the fold changes in each case.

To analyze the combinatorial binding of TFs in mES cells, seven 'strong' motifs (more than 10% of the highest activity, as mentioned in **Fig. 2a**) were taken into account. Each read in the sequencing data was analyzed for the presence of perfect matches to each of the seven strong motif seeds, and the total number of all the seed matches in each read were counted. For symmetric motifs, only one strand was taken into account; for asymmetric motifs, both strands were analyzed. This analysis revealed that after four cycles, ~5% of reads contained a seed match, and only 0.05% contained matches to more than one seed (**Supplementary Fig. 1b**), indicating that in the early rounds of ATI, the motifs could not effectively compete against each other. Thus, the presence of only a few strongly active motifs in the ATI data could not be due to over-enrichment of one motif that competed out the other motifs during the enrichment cycles.

DNase I hypersensitive site (DHS) analysis. The DHS data for different mouse tissues and ES cells were obtained from the ENCODE project[45], which included 14 replicates for mouse liver, seven replicates for mouse brain and two replicates each for mouse heart, spleen and ES cells. First the BroadPeak data were downloaded, and the top 5,000 regions for each replicate were selected based on signal values. For tissues with two replicates, the intersected regions were used for downstream analysis, which resulted in ~4,000 DHSs for each tissue; for liver, DHSs that overlapped in more than eight replicates were selected to reach a similar size of data as compared with that of the other tissues; for brain, DHSs that overlapped in more than four replicates were selected. For each tissue, the frequencies of all ten-mers were counted in the initial ATI library and in each ATI enrichment cycle; the fold change for each ten-mer was calculated by comparing the frequencies of it in the last cycle (cycle 4) and first cycle (cycle 1). After that, the DHSs and the ten-mer results for the same tissues were analyzed. First, each DHS region was flanked with adjacent genomic sequences to make a 10-kb region, which resulted in ~4,000 regions with the length of 10 kb for each tissue and cell type; all of these 10-kb regions were then aligned by using the middle of the DHS as the center position. A score for each position of the 10-kb sequences was then calculated based on the $\log_2$(fold change of the ATI ten-mers). The histogram in **Figure 4b** was calculated from the average of the scores for all DHSs. For visualization (**Fig. 4b**), the position containing a ten-mer that was ranked at 2,000 or higher in enrichment based on ATI was indicated by a blue dot.

The extended 10-kb regions from ES cells, as well as those from different mouse tissues, were used in the ATI-based prediction of DHSs. All ten-mers enriched in ATI were given a score of 1, whereas all of the others were given a score of 0. The scoring of the ten-mers was optimized by trying different cut-off values using a separate training set (setting separately the top 0.1%, top 0.5%, top 1%, top 5%, top 10%, top 20%, top 40%, top 60% and top 100% (all) of the ten-mers as score 1 and the remaining ten-mers as score 0; 100% of ten-mers was considered as a negative control). The 10-kb regions were divided to 50-bp bins, and each bin was then assigned with the mean score of all ten-mers inside the bin. The DHS position was then called based on identification of the highest score in a sliding window of 17 bins. The optimal width of the smoothing window was determined by using half of the 10-kb regions as a training set; only the test set data are shown in **Figure 4b** for ES cells and in **Supplementary Figure 5b** for different mouse tissues.

Precision-recall analysis. For prediction of DHS regions genome wide, DHS regions representing the intersection of the two top 30,000 mES cell DHS

sequences were selected, and for each DHS region, 10-kb control sequences (non-DHS regions) were also taken from both ends of 50-kb windows centered by it. A score was assigned to each ten-mer as the $\log_2$(fold change) by comparing the counts of the ten-mer in DHS regions and the control regions derived from chromosome 12 to chromosome 18. For the ATI data, the ten-mer scores were calculated as mentioned in the subsection "DNase I hypersensitive site (DHS) analysis". The scores were kept for a fraction of the most-enriched ten-mers, with the remaining ten-mers being assigned a score of 0. To plot the precision-recall curves, a score was assigned to each non-overlapping 1-kb window by adding up the scores of all ten-mers inside the window. Each window was labeled as "DHS" if >500 bp of it was covered by a DHS, otherwise it was labeled as "non-DHS". Then the precision-recall curve was plotted by predicting the labels of all of the windows with their scores using varied thresholds. For the final prediction plots, the fraction of nonzero ten-mer scores was identified by optimizing the area under the curve (AUC) of a precision recall curve by using chromosome 11 and 19 DHS data, and the DHS data from the remaining chromosomes were used as ground truth for the prediction.

For the analysis of DHSs that were easy or difficult to predict by ATI or the DHS ten-mers (**Supplementary Fig. 8**), the 1-kb windows labeled as 'DHS' in the remaining chromosomes were evenly divided into three categories based on scores: tritile_1 with the highest scores, tritile_2 with the intermediate scores, and tritile_3 with the lowest scores. Those 'DHS' windows were further divided into (i) easy to be predicted with both ATI and DHS ten-mer data (intersection of tritile_1 of DHS-based prediction and tritile_1 of ATI-based prediction, 'Easy to predict'), (ii) easy to be predicted with ATI ten-mer data but hard to be predicted with DHS data (intersection of tritile_3 of DHS-based prediction and tritile_1 of ATI-based prediction, 'ATI_predicted'), (iii) easy to be predicted with DHS ten-mer data but hard to be predicted with ATI data (intersection of tritile_1 of DHS-based prediction and tritile_3 of ATI-based prediction, 'DHS_predicted') and (iv) hard to be predicted with both types of data (intersection of tritile_3 of DHS-based prediction and tritile_3 of ATI-based prediction, 'Hard to predict').

Enrichment of DNA-binding proteins using biotinylated ATI ligands. The proteins in the nuclear extract were pulled down by biotinylated DNA, as previously reported[70]. First, DNA oligonucleotides were amplified with biotinylated primers (modified with biotin incorporated with triethylene glycol spacer) and purified to remove extra primers. Subsequently, 2 µg of biotinylated double-stranded DNA was incubated with 4 µl of high-performance streptavidin–sepharose suspension (GE Healthcare, cat. no. 17511301) in DNA-binding buffer (10 mM HEPES, pH 8.0, 1 M NaCl, 10 mM EDTA and 0.05% NP40) for 1 h at room temperature by shaking. Beads were then washed twice with DNA-binding buffer and twice with protein-binding buffer (140 mM KCl, 5 mM NaCl, 1 mM K_2HPO_4, 2 mM $MgSO_4$, 100 µM EGTA and 3 µM $ZnSO_4$, in 20 mM HEPES, pH 7.5). Nuclear extracts from feeder-free mES cells (200 µg in 200 µl), supplemented with 2 µg poly-dIdC competitor DNA and EDTA-free complete protease inhibitors (Sigma, cat. no. 000000004693159001), was added to the beads and incubated for 1.5 h with shaking at room temperature. The beads were then washed with ice-cold low-stringency buffer (10 mM Tris-Cl, pH 7.5, 4% glycerol, 500 µM EDTA and 50 mM NaCl) ten times, followed by on-bead digestion for MS analysis[70].

Mass spectrometry (MS) sample preparation. For on-bead digestion of captured DNA-binding proteins from the nuclear extract, washed beads were incubated in 50 µl of 25 mM ammonium bicarbonate and 1 mM DTT for 1 h at 37 °C. Iodoacetic acid (IAA) was then added to the samples to a final concentration of 5 mM, and the samples were incubated at room temperature in the dark for 10 min. The IAA was then quenched by addition of DTT to a final concentration of 5 mM. Protein samples were then digested, first by using Lys-C protease (0.2 µg/sample; Thermo Scientific, cat. no. 90051) overnight at 37 °C. In the second digestion step, trypsin protease (0.1 µg/sample; Thermo Scientific, cat. no. 90057) was added, and the samples were incubated overnight at 37 °C and then lyophilized with a vacuum microcentrifuge (SpeedVac).

For analysis of nuclear proteins, nuclear extracts were prepared as described above, and protein concentration was determined (Bio-Rad DC assay). For digestion using filter-aided sample prep (FASP), 250 µg of protein sample was mixed with 1 mM DTT, 8 M urea and 25 mM HEPES pH 7.6 in a centrifugation filtering unit with a 10-kDa cut-off (Nanosep Centrifugal Devices with Omega Membrane, 10k). The samples were then centrifuged for 15 min at 14,000*g*, followed by addition of the 8 M urea buffer and centrifugation. Protein samples were digested on the filter, first by using Lys-C (Thermo Scientific) for 3 h in 37 °C, at an enzyme:protein ratio of 1:50. In the second digestion step, trypsin (Thermo Scientific), at an enzyme:protein ratio of 1:50 in 50 mM HEPES, was added and incubated overnight at 37 °C. After digestion, the filter units were centrifuged for 15 min at 14,000*g*, followed by another centrifugation after the addition of 50 µl MilliQ water. Peptides were collected, and the peptide concentration was determined (Bio-Rad DC assay). For the label-free experiment, peptide samples were cleaned up individually by solid-phase extraction (SPE strata-X-C, Phenomenex) and dried in a vacuum microcentrifuge (SpeedVac).

For the relative quantification (TMT) experiment, peptide samples were pH-adjusted using TEAB buffer with pH 8.5 (30 mM final concentration). The resulting peptide mixtures were labeled with isobaric TMT tags (Thermo Scientific). High labeling efficiency was verified by liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) before pooling of samples. Sample clean-up was performed by strong cation-exchange solid-phase extraction (SPE strata-X-C, Phenomenex). Purified samples were dried in a vacuum microcentrifuge.

For peptide prefractionation by high-resolution isoelectric focusing (IEF)[37], 500 µg of the labeled peptide pool was dissolved in 250 µl of rehydration solution (8 M urea, 1% Pharmalyte for pH range 3–10 from GE Healthcare), which was then used to re-swell an immobilized pH gradient (IPG) gel strip (GE Healthcare) pH 3–10. IEF was then run on an Ettan IPGphor isoelectric focusing system (GE Healthcare) until at least 150 kVh (~1 d running time). After focusing was complete, a well-former with 72 wells was applied onto the strip, and liquid-handling robotics (GE Healthcare prototype) was used to add MilliQ water for a 30-min incubation/extraction of peptides. Extracted peptides from the 72 fractions were then transferred into a microtiter plate (96 wells, V-bottom, Corning cat. no. 3894). The extraction was repeated three times after which the combined samples on the microtiter plate were dried using a vacuum microcentrifuge.

Mass spectrometry. Label-free MS of peptides from captured DNA-binding proteins was performed using a hybrid Q-Exactive mass spectrometer (Thermo Scientific). Each sample was resuspended in 10 µl of solvent A (95% water, 5% DMSO and 0.1% formic acid (FA)) of which 3 µl was injected. Peptides were trapped on an Acclaim PepMap nanotrap column (C18, 3 µm, 100 Å, 75 µm × 20 mm) and separated on an Acclaim PepMap RSLC column (C18, 2 µm, 100 Å, 75 µm × 50 cm, Thermo Scientific). Peptides were separated using a gradient of A (5% DMSO and 0.1% FA) and B (90% acetonitrile (ACN), 5% DMSO and 0.1% FA), which ranged from 6% to 37% B in 240 min with a flow of 0.25 µl/min. Q-Exactive (QE) was operated in a data-dependent manner, performing FTMS (Fourier Transform Mass Spectrometry) survey scans at 70,000 resolution (and mass range 300–1,700 *m/z*) followed by MS/MS (35,000 resolution) of the top five ions using higher-energy collision dissociation (HCD) at 30% normalized collision energy. Precursors were isolated with a 2-*m/z* window. Automatic gain control (AGC) targets were 1×106 for MS1 and 1×10^5 for MS2. Maximum injection times were 100 ms for MS1 and 150 ms for MS2. The entire duty cycle lasted ~1 s. Dynamic exclusion was used with a 60-s duration. Precursors with unassigned charge state or a charge state of 1 were excluded. An underfill ratio of 1% was used.

LC-MS of TMT-labeled peptides from nuclear extracts was also performed using a hybrid Q-Exactive mass spectrometer (Thermo Scientific). For each LC-MS/MS run, the auto sampler (Dionex UltiMate 3000 RSLCnano System) dispensed 15 µl of solvent A (95% water, 5% DMSO and 0.1% formic acid) to the well in the 96-well plate, mixed, and 7 µl proceeded to injection. Peptides were trapped on an Acclaim PepMap nanotrap column (C18, 3 µm, 100 Å, 75 µm × 20 mm), and separated on an Acclaim PepMap RSLC column (C18, 2 µm, 100Å, 75 µm × 50 cm, Thermo Scientific). Peptides were separated using a gradient of A (5% DMSO and 0.1% FA) and B (90% ACN, 5% DMSO and 0.1% FA), ranging from 6% to 37% B in 50 min with a flow of 0.25 µl/min. QE was operated as described above.

Label-free MS of proteins in the nuclear extract was performed using Orbitrap Fusion Tribrid mass spectrometer (Thermo Scientific). Before the analysis, peptides were separated with an Ultimate 3000 RSLCnano system.

doi:10.1038/nbt.4138

Samples were trapped on an Acclaim PepMap nanotrap column (C18, 3 µm, 100Å, 75 µm × 20 mm) and separated on an Acclaim PepMap RSLC column (C18, 2 µm, 100Å, 75 µm × 50 cm), (Thermo Scientific). Peptides were separated using a gradient of A (5% DMSO and 0.1% FA) and B (90% ACN, 5% DMSO and 0.1% FA) that ranged from 6% to 37% B in 240 min with a flow of 0.25 µl/min. The Orbitrap Fusion was operated in a data-dependent manner, selecting the top ten precursors for sequential fragmentation by higher-energy collision dissociation (HCD) and collision-induced dissociation (CID). The survey scan was performed in the Orbitrap at 120,000 resolution from 350–1,550 *m/z*, with a maximum injection time of 50 ms and a target of 2×10^5 ions. Precursors were isolated by the quadrupole with a 1.4-*m/z* window and a 0.5-*m/z* offset, and they were put on the exclusion list for 30 s. Charge states between 2 and 7 were considered for precursor selection. For generation of HCD fragmentation spectra, a maximum ion injection time of 100 ms and an AGC target of 1×10^5 were used before fragmentation at 37% normalized collision energy and analysis in the Orbitrap at 30,000 resolution. For generation of CID fragmentation spectra, a maximum ion injection time of 100 ms and an AGC target of 1×10^4 were used before fragmentation at 35% activation energy, activation Q of 0.25 and analysis in the ion trap, using normal scan range and rapid scan rate.

Peptide and protein identification. For the label-free capture experiment, MS raw files were searched using Sequest-percolator under the software platform Proteome Discoverer 1.4 (Thermo Scientific) against the Uniprot mouse database (version 2016_10, canonical and isoforms, 85,832 protein entries) and filtered to a 1% false discovery rate (FDR) cut-off (peptide-spectrum-match level). A maximum of two missed cleavages was used together with: carbamidomethylation (C) set as fixed modification and oxidation (M) set as a variable modification. We used a precursor ion mass tolerance of 10 p.p.m. and a product ion mass tolerance of 0.02 Da for HCD spectra. For calculation of the precursor ion area, a mass precision of 2 p.p.m. between scans was used, and the average area of the top three PSMs for each protein group was used to calculate protein area. Only unique peptides in the data set were used for quantification. In total the database search resulted in the identification of 3,889 proteins (**Supplementary Table 3**). TFs were assigned to related motifs detected in ATI based on the current database (HT-SELEX motifs, CIS-BP, TOMTOM[9,13,35,36]). If only one TF was assigned to a motif, then this particular TF was regarded as the candidate TF; if more than one TF was assigned to a motif, then TFs with >20% of the highest abundance (based on values in "Aver_area(c4)") were selected as the candidate TFs that were dominant in the cells (**Supplementary Table 1**). These parameter values should be considered only as an example, as the optimal cut-offs depended on the TFs and the purpose of the individual projects.

For the nuclear-extract analysis, MS raw files were searched using Sequest-percolator under the software platform Proteome Discoverer 1.4 (Thermo Scientific) against the Uniprot mouse reference database (version 2014_03, canonical only, 43,386 protein entries) and filtered to a 1% FDR cut off (peptide-spectrum-match level). For TMT experiments, a maximum of two missed cleavages was used together with: carbamidomethylation (C), TMT-labels (on lysine and N-terminal residues) set as fixed modifications, and oxidation (M) set as a variable modification. We used a precursor ion mass tolerance of 10 p.p.m. and a product ion mass tolerance of 0.02 Da for HCD spectra. Quantification of reporter ions was done by Proteome Discoverer on HCD–FTMS tandem mass spectra using an integration window tolerance of 10 p.p.m. Only unique peptides in the data set were used for quantification. In total the database search resulted in the identification and quantification of 8,578 proteins in the TMT experiment (**Supplementary Table 9**).

For label-free analysis, a maximum of one missed cleavage was used together with: carbamidomethylation (C) set as fixed modifications and oxidation (M) set as a variable. We used a precursor ion mass tolerance of 12 p.p.m., and a product ion mass tolerance of 0.02 Da for HCD spectra and 0.36 for CID spectra. For calculation of precursor ion area, a mass precision of 3 p.p.m. between scans was used, and the average area of the top three PSMs for each protein group were used to calculate protein area. In total, the database search resulted in the identification of 6,239 proteins in the label-free experiment (**Supplementary Table 10**).

Statistical analysis. No statistical methods were used to predetermine sample size. For experiments without independent duplicates (the ATI assay), the results were expressed as individual values; for experiments with duplicates (the iHep reprogramming assay for example), the results were shown as means of the two duplicates; for experiments with triplicate samples, the group results were expressed as mean ± s.d., unless stated otherwise. For comparisons between groups with triplicates, the *P* values were calculated using two-tailed Student's *t*-test in **Figure 2c** and one-sided Student's *t*-test in **Supplementary Table 4**. The statistical analysis was performed using Winflat program for sequencing-data-derived results.

Life Sciences Reporting Summary. Further information about experimental design is available in the **Life Sciences Reporting Summary**.

Data availability. All next-generation sequencing data have been deposited in the European Nucleotide Archive (ENA) under accession PRJEB15639. All of the computer programs and scripts used are either published or available upon request. All data is available upon request.

61. Morris, S.A. *et al.* Dissecting engineered cell types and enhancing cell fate conversion via CellNet. *Cell* **158**, 889–902 (2014).
62. Simicevic, J. *et al.* Absolute quantification of transcription factors during cellular differentiation using multiplexed targeted proteomics. *Nat. Methods* **10**, 570–576 (2013).
63. Schmitges, F.W. *et al.* Multiparameter functional diversity of human C2H2 zinc finger proteins. *Genome Res.* **26**, 1742–1752 (2016).
64. Imbeault, M., Helleboid, P.Y. & Trono, D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**, 550–554 (2017).
65. Zhang, Y. *et al.* Model-based analysis of ChIP-seq (MACS). *Genome Biol.* **9**, R137 (2008).
66. Neph, S. *et al.* BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28**, 1919–1920 (2012).
67. Jolma, A. *et al.* Multiplexed massively parallel SELEX for characterization of human transcription factor–binding specificities. *Genome Res.* **20**, 861–873 (2010).
68. Korhonen, J., Martinmäki, P., Pizzi, C., Rastas, P. & Ukkonen, E. MOODS: fast search for position-weight-matrix matches in DNA sequences. *Bioinformatics* **25**, 3181–3182 (2009).
69. Pizzi, C., Rastas, P. & Ukkonen, E. Finding significant matches of position-weight matrices in linear time. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **8**, 69–79 (2011).
70. Hubner, N.C., Nguyen, L.N., Hornig, N.C. & Stunnenberg, H.G. A quantitative proteomics tool to identify DNA–protein interactions in primary cells or blood. *J. Proteome Res.* **14**, 1315–1329 (2015).

nature research

Corresponding author(s): Jussi Taipale

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.	ATI sequencing depth was set in such a way that on average each experiment would result in millions of independent sequence reads. Statistical power to detect differences between counts of short subsequences in the samples is thus extremely high. Effect size was shown in each case to clarify the magnitude of the difference between the samples.

2. Data exclusions

Describe any data exclusions.	No data were excluded

3. Replication

Describe whether the experimental findings were reliably reproduced.	ATI experiments consist of multiple cycles that consistently enrich the motifs. ATI data from ES cells also shows high reproducibility across the samples. MS data was performed in triplicate samples with similar result. The iHep reprogramming experiment was repeated twice with similar result. The ES cell differentiation assay was done in technical duplicates, and further proceeded to ATI assay.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.	Samples were analyzed directly and individually, and not randomized to experimental groups

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.	Most analyses were performed using computational algorithms. Investigators were not blinded.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

Nature Biotechnology: doi:10.1038/nbt.4138

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	The commercial and public softwares include MEME (version 4.9.0), MOODS (version 1.9.1), MACS (version 1.4), BEDOPS (version 2.4.26), Autoseed and Proteome Discoverer 1.4 (Thermo Scientific).

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	No unique materials were required.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	No antibodies were used.

Nature Biotechnology: doi:10.1038/nbt.4138

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

mouse embryonic stem cells and MEF feeder cells were from KI core facility (Karolinska Center for Transgene Technologies); human fibroblast cell line CCD-1112Sk was obtained from ATCC (#CRL 2429); Drosophila S2 cells were purchased from Thermo (#CRL R69007) and directly used.

b. Describe the method of cell line authentication used.

The human fibroblast cell line CCD-1112Sk and Drosophila S2 cells were directly obtained from trusted vendors and not from other laboratories and used within short time. The mouse embryonic stem cells were authenticated by production of germline chimeric mice, AP staining and cell morphology. The MEF feeder cells were not authenticated.

c. Report whether the cell lines were tested for mycoplasma contamination.

The mouse embryonic stem cells and MEF feeder cells were negative for mycoplasma test; the human fibroblast cell line CCD-1112Sk and drosophila S2 cells were bought from ATCC and Thermo Scientific, therefore are negative for mycoplasma.

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

No commonly misidentified cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

Tissues used in the study were from one-year old C57BL/6J male mouse.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

The study did not involve human research participants.

Nature Biotechnology: doi:10.1038/nbt.4138