⌘ *Author's Choice*

Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC) and Proteome Quantitation of Mouse Embryonic Stem Cells to a Depth of 5,111 Proteins*Ⓢ

Johannes Graumann‡§¶, Nina C. Hubner‡§, Jeong Beom Kim‖, Kinarm Ko‖, Markus Moser‡‡, Chanchal Kumar‡, Jürgen Cox‡, Hans Schöler‖, and Matthias Mann‡§§**

Embryonic stem (ES) cells are pluripotent cells isolated from mammalian preimplantation embryos. They are capable of differentiating into all cell types and therefore hold great promise in regenerative medicine. Here we show that murine ES cells can be fully SILAC (stable isotope labeling by amino acids in cell culture)-labeled when grown feeder-free during the last phase of cell culture. We fractionated the SILAC-labeled ES cell proteome by one-dimensional gel electrophoresis and by isoelectric focusing of peptides. High resolution analysis on a linear ion trap-orbitrap instrument (LTQ-Orbitrap) at sub-ppm mass accuracy resulted in confident identification and quantitation of more than 5,000 distinct proteins. This is the largest quantified proteome reported to date and contains prominent stem cell markers such as OCT4, NANOG, SOX2, and UTF1 along with the embryonic form of RAS (ERAS). We also quantified the proportion of the ES cell proteome present in cytosolic, nucleoplasmic, and membrane/chromatin fractions. We compared two different preparation approaches, cell fractionation followed by one-dimensional gel separation and in-solution digestion of total cell lysate combined with isoelectric focusing, and found comparable proteome coverage with no apparent bias for any functional protein classes for either approach. Bioinformatics analysis of the ES cell proteome revealed a broad distribution of cellular functions with overrepresentation of proteins involved in proliferation. We compared the proteome with a recently published map of chromatin states of promoters in ES cells and found excellent correlation between protein expression and the presence of active and repressive chromatin marks. ***Molecular & Cellular Proteomics 7:672–683, 2008.***

Because of their pluripotency and potentially unlimited capacity of self-renewal as well as developmental inducibility, embryonic stem (ES)[1] cells hold great promise both as model systems in developmental biology and for regenerative medicine (1). ES cells pose a plethora of scientific questions. These range from which factors enable this cell type to retain "stemness" (the undifferentiated and pluripotent state) to the mechanisms of differentiation into various cell and tissue types. Although traditional candidate gene approaches have provided detailed insight into many of these areas, technologies characterizing the cell type as a whole and comparing it with others have the potential to provide an unbiased, "systems-level" view and to uncover unanticipated aspects of ES cell biology.

A rich body of literature describes global stem cell characterization at the level of the transcriptome (2, 3), and more recently several studies on the global chromatin state of ES cells were added to that arsenal (see for example, Ref. 4). However, regulation of chromatin state and transcript abundance represent only two aspects of the realization of any cellular process. Studies centering on them alone implicitly disregard the influences of translational and post-translational regulation of protein levels and activity, such as proteolysis and covalent modifications. For this reason, it is important to complement other large scale approaches with proteomics analysis. The technology of MS-based proteomics has become increasingly powerful in many areas of protein-based research (5), and very recently, proteome-wide quantitation has been demonstrated (6). However, proteomics methods applied to the embryonic stem cell field have not yet used

From the Departments of ‡Proteomics and Signal Transduction and ‡‡Molecular Medicine, Max Planck Institute for Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany and ‖Department of Cell and Developmental Biology, Max Planck Institute for Molecular Biomedicine, Roentgenstr. 20, 48149 Münster, Germany

Received, September 25, 2007, and in revised form, November 19, 2007

Published, MCP Papers in Press, November 28, 2007, DOI 10.1074/mcp.M700460-MCP200

⌘ *Author's Choice*—Final Version Full Access.

[1] The abbreviations used are: ES, embryonic stem; SILAC, stable isotope labeling by amino acids in cell culture; ERAS, embryonic form of RAS; MEF, mouse embryonic fibroblast; BMP4, bone morphogenic protein 4; bis-Tris, 2-[bis(2-hydroxyethyl)amino]-2-(hydroxymethyl)-propane-1,3-diol; IPI, International Protein Index; GO, Gene Ontology; 1D, one-dimensional; GeLCMS, in-gel digest followed by LC-MS/MS; ID, identity; ChIPseq, chromatin immunoprecipitation together with large scale sequencing of the occupied DNA region; H3K4me3, histone 3 lysine 4 trimethylation, H3K27me3, histone 3 lysine 27 trimethylation; P, present.

these recent developments and have had much reduced depth when compared with cDNA-based microarray studies (7). The most extensive studies of the proteome of mouse ES cells feature 1,790 (8) and 1,775 (9) identified proteins, and there is one study identifying 1,532 proteins in murine and human ES cells (9). These experiments were non-quantitative, rendering differential analysis impossible. The only exception (9) used peptide counting, a method suitable for highlighting large scale changes in protein abundance but not appropriate for determining accurate quantitative changes on a protein by protein basis. This is especially true for low abundance-level, regulatory proteins. Methods using stable isotopes provide more accurate quantitation (10). Among these techniques metabolic labeling would be especially attractive because it eliminates error-prone parallel steps in protein purification protocols. However, metabolic labeling methods have so far mainly been used with transformed cell lines, and labeling of ES cells, a cell type that is difficult to culture, has not yet been demonstrated.

We show here that complete metabolic labeling of murine embryonic stem cells using stable isotope labeling by amino acids in cell culture (SILAC (11, 12)) is feasible. Here we used SILAC-labeled ES cells to achieve increased confidence of peptide identification and to construct an initial high quality reference proteome of 5,111 proteins. In addition to other low abundance protein classes such as transcription factors and kinases, this proteome contains well documented stem cell markers, which suggests that the SILAC-labeled cells retain stemness. We also quantified compartmental distribution of the stem cell proteome, and we compared the combination of isoelectric focusing of peptides from in-solution digest with the established in-gel procedure. Bioinformatics analysis of this large and high confidence ES cell proteome revealed overall features of this cell type, including its strong proliferative character.

EXPERIMENTAL PROCEDURES

Culture of Embryonic Stem Cells—The mouse embryonic stem cell lines G-olig2 (13) and R1 were cultured as adherent cells on mouse embryonic fibroblasts (MEFs) mitotically inactivated either by irradiation at 3,000 rads or mitomycin C. Dulbecco's modified Eagle's medium (Invitrogen) devoid of arginine and lysine was supplemented with 15 or 20% fetal bovine serum dialyzed with a cutoff of 10 kDa (Invitrogen, 26400-044); 3.5 mg/ml glucose (to a final concentration of 4.5 mg/ml); 0.1 mM non-essential amino acids without arginine, lysine, and proline; 100 units/ml penicillin/streptomycin (Invitrogen, 115140-122); 2 mM Glutamax (Invitrogen, 35050-038); 100 β-mercaptoethanol (Sigma, M7522); and 1,000 units/ml leukemia inhibitory factor (Chemicon, ESG1107). The medium was replaced every day, and cells were split every 2nd day.

For labeling, arginine and lysine were added in either light (Arg0, Sigma, A5006; Lys0, Sigma, L5501) or heavy (Arg10, Cambridge Isotope Laboratories, CNLM-539; Lys8, Cambridge Isotope Laboratories CNLM-291) form to a concentration of 28 μg/ml for arginine and 49 μg/ml for lysine (Arg0/Lys0: arginine and lysine with normal "light" carbons (^{12}C) and nitrogens (^{14}N); Arg10/Lys8: arginine and lysine derivatives with "heavy" carbons (^{13}C) and nitrogens (^{15}N)). Cells were tested for full incorporation of the label after five passages.

ES cells were either harvested after twice settling for 30 min to separate them from feeder cells or after feeder-free culture on plates coated with 0.1% gelatin for three of the five passages. In the latter case the medium was supplemented with 25 ng/ml recombinant human bone morphogenic protein 4 (BMP4; PeproTech, 120-05).

Cell Lysis and In-solution Digest—To determine the incorporation rate of heavy amino acids, cell pellets were resuspended in cold lysis buffer (1% *N*-octyl glucoside, 0.1% sodium deoxycholate, 150 mM NaCl, 1 mM EDTA, 50 mM Tris-HCl (pH 7.5), EDTA-free Complete protease inhibitor mixture (Roche Applied Science, 11836153001)) and incubated for 10 min on ice. The lysate was then cleared by centrifugation.

Proteins were methanol/chloroform-precipitated (14) and resuspended in 1 pellet volume of 6 M urea, 2 M thiourea in 10 mM Hepes (pH 8.0). After reduction and alkylation with 1 mM DTT and 5.5 mM iodoacetamide, proteins were digested with 5 μg of Lys-C (Wako Chemicals, 129-02541) for 3 h at room temperature. Prior to digestion with 5 μg of trypsin (Promega, V511C) for 12 h at room temperature the urea/thiourea concentration was reduced to 2 M by dilution with 10 mM ammonium bicarbonate. The reaction was stopped by acidifying with trifluoroacetic acid to a pH lower than 2.5. Each sample was loaded on C_{18} StageTips (15).

Subcellular Fractionation and In-gel Digest—Feeder-free cultured ES cells were mixed 1:1 heavy and light to obtain a cell pellet of approximately 60-μl volume. This pellet was subjected to a subcellular fractionation protocol modified according to Dignam *et al.* (16). The pellet was resuspended and incubated for 10 min in ice-cold buffer containing 10 mM Hepes-KOH (pH 7.9), 1.5 mM $MgCl_2$, 10 mM KCl, 0.2% *N*-octyl glucoside, and EDTA-free Complete protease inhibitor mixture (Roche Applied Science, 11836153001). The suspension was homogenized in a 0.1 ml Potter-Elvehjem homogenizer (Neolab, 9-0905). The supernatant containing predominantly cytoplasmic proteins was collected after 15-min centrifugation at 400 × *g* at 4 °C. The remaining pellet was washed in ice-cold PBS, resuspended in cold buffer containing 420 mM NaCl, 20 mM Hepes-KOH (pH 7.9), 20% glycerol, 2 mM $MgCl_2$, 0.2 mM EDTA, 0.1% *N*-octyl glucoside, 0.5 mM DTT, and EDTA-free Complete protease inhibitory mixture and incubated on ice for 1 h. The supernatant containing predominantly nucleoplasmic proteins was collected after 15-min centrifugation at 18,000 × *g* at 4 °C. The chromatin/membrane-containing pellet was resuspended in cold PBS supplemented with 600 mM NaCl, 1% *N*-octyl glucoside, and 125 units of Benzonase (Novagen, 70746); incubated for 30 min in an ultrasonic bath; and centrifuged for 15 min at 18,000 × *g* at 4 °C. Chromatin/membrane proteins were collected with the supernatant.

300 μg of protein of each fraction were separated on a 4–12% NuPage Novex bis-Tris gel (Invitrogen, NP0321) in three lanes each and stained using the Colloidal Blue Staining kit (Invitrogen, LC6025) according to the manufacturer's instructions. The gel was cut into 15 slices containing approximately the same protein amount, and slices from the three identical gel lanes were pooled. The in-gel digest was performed according to Shevchenko *et al.* (17) with minor modifications. Each sample was loaded on C_{18} StageTips (15).

Isoelectric Focusing—ES cells were cultured under feeder-free conditions (during the last three passages) in media containing either the light or heavy version of arginine and lysine, mixed 1:1, and in-solution digested as described above. Peptides obtained from the digestion of 250 μg of protein were focused using the Agilent 3100 OFFGEL Fractionator (Agilent, G3100AA) and the 3100 OFFGEL High Res kit, pH 3–10 (Agilent, 5188-6424) according to the manufacturer's instructions. Peptides were focused for 50 kV-h at a maximum current of 50 μA and maximum power of 200 milliwatts. Peptide fractions were acidified by adding 10% of a solution containing 30% acetonitrile, 10% trifluoroacetic acid, and 5% acetic acid prior to using StageTips and MS analysis.

LC-MS/MS—Peptides were twice eluted from StageTips using 20 μl of 80% acetonitrile, 0.5% acetic acid; the volume was reduced to 5 μl in the SpeedVac, and the peptides were acidified with 5 μl of 2% acetonitrile, 1% trifluoroacetic acid.

All LC-MS/MS experiments were performed essentially as described previously (18). Briefly peptides were separated using an Agilent 1200 nanoflow LC system consisting of a solvent degasser, a nanoflow pump, and a thermostated microautosampler. 5 μl of sample were loaded with constant flow of 500 nl/min onto a 15-cm fused silica emitter with an inner diameter of 75 μm (Proxeon Biosystems) packed in-house with reverse-phase ReproSil-Pur C_{18}-AQ 3-μm resin (Dr. Maisch GmbH). Peptides were eluted with a segmented gradient of 10–60% solvent B over 105 min with a constant flow of 200 nl/min. The HPLC system was coupled to an LTQ-Orbitrap mass spectrometer (ThermoFisher Scientific) via a nanoscale LC interface (Proxeon Biosystems). The spray voltage was set to 2.3 kV, and the temperature of the heated capillary was set to 180 °C. Survey full-scan MS spectra (*m/z* 300–1700) were acquired in the orbitrap with a resolution of 60,000 at *m/z* 400 after accumulation of 1,000,000 ions. The five most intense ions from the preview survey scan delivered by the orbitrap were sequenced by collision-induced dissociation (normalized collision energy, 40%) in the LTQ after accumulation of 5,000 ions concurrently to full-scan acquisition in the orbitrap. Maximal filling times were 1,000 ms for the full scans and 150 ms for the MS/MS scans. Precursor ion charge state screening was enabled, and all unassigned charge states as well as singly charged species were rejected. The dynamic exclusion list was restricted to a maximum of 500 entries with a maximum retention period of 180 s and a relative mass window of 15 ppm. The lock mass option was enabled for survey scans to improve mass accuracy (19). Data were acquired using the Xcalibur software. The raw data will be made available to interested parties upon request.

Bioinformatics Analysis—Mass spectra were analyzed using the in-house developed software MaxQuant (version 1.0.4.11) (20), which performs peak list generation, SILAC- and extracted ion current-based quantitation, false positive rate (21) determination based on search engine results, peptide to protein group assembly, and data filtration and presentation. The data were searched against the mouse International Protein Index protein sequence database (IPI version 3.24 (22)) supplemented with frequently observed contaminants (porcine trypsin, *Achromobacter lyticus* lysyl endopeptidase, and human keratins; a total of 52,355 forward entries) and concatenated with reversed copies of all sequences (23, 24) using Mascot (version 2.1.04, Matrix Science (25)). Enzyme specificity was set to trypsin, allowing for cleavage N-terminal to proline and between aspartic acid and proline (18). Carbamidomethylcysteine was set as a fixed modification, and oxidized methionine, *N*-acetylation, and loss of ammonia from N-terminal glutamine were set as variable modifications. Spectra determined to result from heavy labeled peptides by presearch MaxQuant analysis were searched with the additional fixed modifications Arg10 and Lys8, whereas spectra with a SILAC state not determinable *a priori* were searched with Arg10 and Lys8 as additional variable modifications. Maximum allowed mass deviation (26) was set initially to 5 ppm for monoisotopic precursor ions and 0.5 Da for MS/MS peaks. A maximum of three missed cleavages and three labeled amino acids (arginine and lysine) were allowed. The required false positive rate was set to 5% at the peptide level, the required false discovery rate was set to 1% at the protein level, and the minimum required peptide length was set to 6 amino acids. False positive rates for peptides are calculated by recording Mascot score and peptide sequence length-dependent histograms of forward and reverse hits separately and then, using Bayes' theorem, deriving the probability of a false identification for a given top scoring peptide. The cutoff used on the peptide level ensures that the worst identified peptide has a probability of 0.05 of being false. Proteins are then sorted by the product of the false positive rates of the contained peptides where only peptides with distinct sequences are taken into account. Proteins are successively included starting with the best identified ones until a false discovery rate of 1% is reached, which is estimated based on the fraction of reverse protein hits. If the identified peptide sequence set of one protein was equal to or contained the peptide set of another protein, these two proteins were grouped together by MaxQuant and not counted as independent protein hits. On top of the protein false discovery rate threshold, proteins were considered identified with at least two peptides (thereof one uniquely assignable to the respective sequence) and quantified if at least one MaxQuant-quantifiable SILAC pair was associated with them. No outliers are removed due to the use of robust statistics (median instead of average of the peptides). Significance of protein ratios is determined in two alternative ways. To obtain a robust and asymmetrical estimate of the standard deviation of the main distribution we calculate the 15.87, 50, and 84.13 percentiles r_{-1}, r_0, and r_1 (corresponding to 1 σ in each direction from the mean). We define $r_1 - r_0$ and $r_0 - r_{-1}$ as the right- and left-sided robust standard deviations, respectively. For a normal distribution, these would be equal to each other and to the conventional definition of a standard deviation. A suitable measure for a ratio $r > r_0$ of being significantly far away from the main distribution would be the distance to r_0 measured in terms of the right standard deviation as follows.

$$z = \frac{r - r_0}{r_1 - r_0} \qquad \text{(Eq. 1)}$$

This can be analogously defined for $r < r_0$. To get a more intuitive, probability-like quantity we calculate the value of the complementary error function for the *z* above, which would for normally distributed data correspond to the probability of obtaining a value this large or larger by chance and call it significance A. For instance, a value of 0.0013 for significance A would indicate a distance of 3 standard deviations from the center of the distribution.

Significance B uses the same strategy, but takes into account the dependence of the distribution on the summed protein intensity. The accuracy of a protein ratio is assessed by calculating the coefficient of variability over all redundant quantifiable peptides.

To determine the quality of the subcellular fractionation, a list of all identified proteins was created, containing the average normalized signal intensity of the identified peptides (as calculated by MaxQuant) in any of the three fractions (cytoplasmic, nucleoplasmic, and chromatin/membrane). The resulting 4,041 protein hits were clustered according to their signal intensity (0–100%) in each of the fractions using Genesis (27). The protein clusters were analyzed according to their statistically overrepresented Gene Ontology (GO) categories using BinGO (28), a Cytoscape (29) plug-in. The clusters were compared against a reference set of the complete mouse proteome, a list of all IPI numbers (version 3.24), and their respective GO identifiers. The GO annotations were extracted from the European Bioinformatics Institute Gene Ontology Annotation (GOA) Mouse 36.0 release containing 34,888 proteins. The analysis was done using the hypergeometric test. All GO terms with a *p* value <0.001 were accepted after correcting for multiple terms testing by the Benjamini and Hochberg false discovery rate. The analysis was done for GO cellular compartment and GO biological function categories. The enrichment was calculated according to Adachi *et al.* (30).

We used ProteinCenter (Proxeon Bioinformatics, Odense, Denmark), a proteomics data mining and management software, to compare the results of the two prefractionation methods, subcellular fractionation in combination with SDS gel electrophoresis and isoelectric focusing. Further analysis and plotting were performed using the R statistical computing and graphics environment (31).

Comparison of the complete proteome with a recent microarray analysis of ES cells by Hailesellasse Sene *et al.* (32) was carried out in two steps. We first estimated the basal expression of the ES cell transcriptome, and in a second step we mapped our proteome data set onto the resulting transcriptome. The microarray experiments were carried out with two different array types. We analyzed the triplicates of each array type separately and calculated the MAS5 expression values using the "mas5" function implemented in the "affy" package of the statistical and computational environment R (31). For reporting the MAS5 present (P) *versus* absent calls we used a *p* value cutoff of 0.01, the same as our proteome acceptance stringency, rather than the usual 0.05.

The expression values were then converted to $\log_2$ scale and *z*-transformed to facilitate the comparison of mRNA expression across two array types. Subsequently the data for the MOE430A/B arrays were combined into one set. A probe set was considered expressed if it was present in two of three triplicates, *i.e.* a P call of 66%. Only 7,926 probe sets of a total of 45,265 met this criterion. They in turn mapped to 5,490 unique Entrez gene IDs. For expression comparison with the mRNA data set the protein intensity values were also converted to $\log_2$ scale and *z*-transformed. Finally the overlap between the mRNA (5,490 genes) and our proteome (4,948 genes) data set was identified. This overlapping set was then used to calculate protein-mRNA expression correlation using the *z*-transformed expression values for each entity.

RESULTS

SILAC of Embryonic Stem Cells—For the SILAC technology, cells are grown in the presence of light or heavy forms of amino acids, such as arginine and lysine. Although there is no indication that incorporation of a heavy amino acid has any effect on cells, the SILAC procedure requires the use of dialyzed serum to remove the natural amino acids already present in the serum. In this process, low molecular weight growth factors can also be removed, potentially interfering with growth of susceptible cell types. Secondly ES cells are usually grown on MEFs as "feeder cells" that provide an environment for ES cells allowing them to remain in the undifferentiated state. In proteomics analysis these feeder cells are undesirable because they could contaminate the ES cell proteome.

We first tested whether mouse ES cells would grow in SILAC medium using feeder cells or under feeder-free culturing conditions. We used two common mouse ES cell lines, R1 and G-Olig2 (13), which were derived from the former. Despite the dialyzed serum used, neither of the two cell populations deviated from their normal colony morphology (data not shown).

As mentioned above, ES cells are traditionally cultured on MEF feeder layers inactivated by irradiation or mitomycin C. The feeder layer is renewed when passaging ES cells and may represent a substantial source of unlabeled amino acids. To evaluate this possibility, we grew G-Olig2 ES cells on feeders in medium providing solely heavy arginine and lysine for five passages. ES cells were separated from contaminating feeders via the significantly faster attachment rate of feeders. This led to an ES cell population of 98% purity by visual inspection through light microscopy. We then evaluated the relative enrichment of heavy labeled peptides by LC-MS of in-solution digested whole cell extracts (Fig. 1*A*). The figure clearly shows incomplete labeling with an average ratio between heavy and light SILAC states of about 6 (83% of peptides in the heavy state). The low labeling efficiency of 0.83 and the bimodal

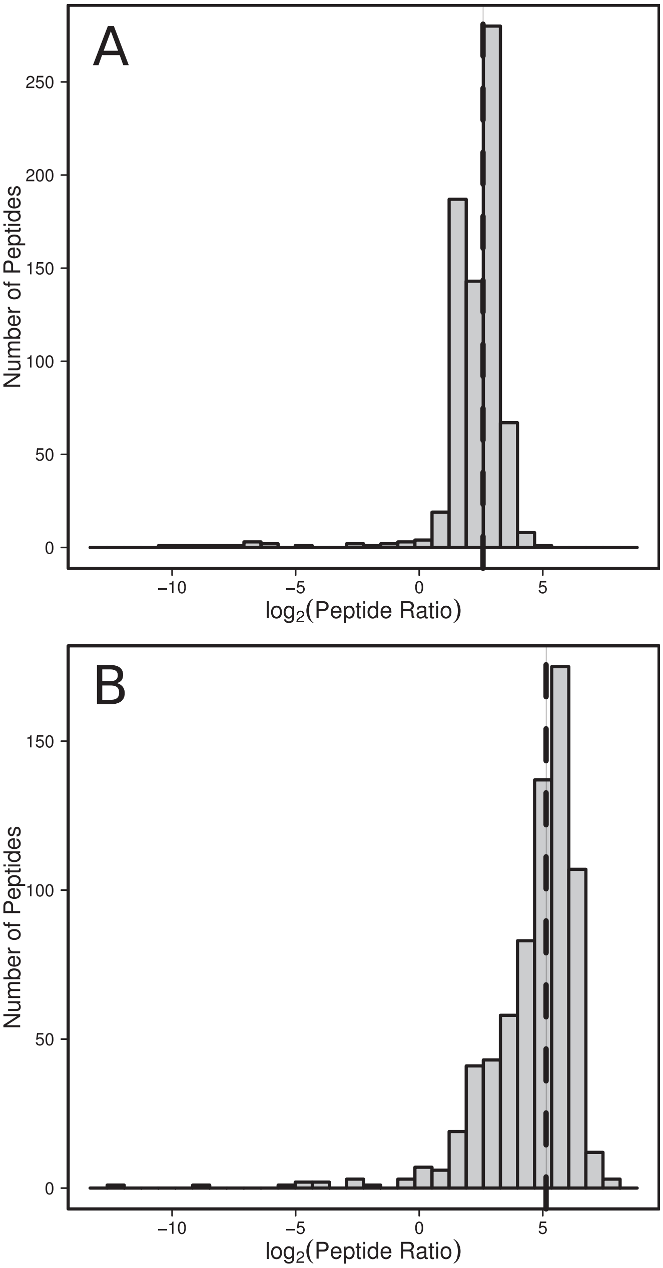

FIG. 1. **Mouse ES cells are readily SILAC-labeled, but feeder cells interfere with labeling efficiency.** G-Olig2 embryonic stem cells were grown for five passages with heavy arginine and lysine as the sole source for the respective amino acids, lysed in modified RIPA buffer, precipitated, digested in solution, and analyzed by LC-MS. *A*, SILAC ES cell culture on feeder cells. The *red line* indicates a median peptide enrichment ratio of 6.02 (83% labeling efficiency). *B*, SILAC ES cell culture under feeder-free conditions. The mean enrichment ratio (*dashed line*) was 36.30 (97% labeling efficiency).

distribution of peptide ratios suggest that the sample is composed of partially labeled feeder cells and of fully labeled ES cells. Likely even low contamination with feeders has a strong contaminating effect because their diameter is approximately twice that of ES cells.

In a second attempt to achieve complete SILAC labeling, we then grew ES cells in BMP4-supported feeder-free culture for three passages prior to harvest (33). As can be seen in Fig. 1*B*, this led to a unimodal distribution of high incorporation ratios of heavy amino acids. The average labeling efficiency after five passages was 97% showing that mouse ES cells can be efficiently and completely SILAC-labeled.

Very recently, van Hoof *et al.* (34) reported high arginine to proline conversion in a human ES cell line, and they proposed a strategy to avoid quantitation errors potentially introduced by this conversion. However, at our arginine concentrations there was no strong arginine to proline conversion in these cell lines.

Subcellular Proteomics of ES Cells—Having established the compatibility of ES cell culture with SILAC, we set out to acquire an initial deep proteome of murine embryonic stem cells. To that end we sought to reduce the complexity of the ES cell lysate by standard subcellular fractionation as described under "Experimental Procedures." The three resulting fractions, cytoplasmic, nucleoplasmic, and chromatin/membrane fraction, were separated on a 1D SDS gel (Fig. 2*A*), and the gel lanes were sliced into 15 gel blocks and subjected to in-gel digest followed by LC-MS/MS ("GeLCMS") analysis. Mass spectrometric measurements were performed on an LTQ-Orbitrap using 140-min gradients per fraction. Mass resolution was set to 60,000 at *m/z* 400, and average absolute mass accuracy was 300 ppb (S.D. 300 ppb) due to the lock mass option and estimation of mass centroids over the elution peak (19, 20). Proteins were accepted for identification using stringent criteria, including the requirement of identification by two fully tryptic peptides (18) with at least one peptide unique to the protein sequence and not shared with any other database entry. Overall protein false discovery rate was required to be less than 1% (see "Experimental Procedures"). The combined analysis of 45 gel slices resulted in the acquisition of 516,649 tandem mass spectra, which yielded 35,963 unique peptide identifications and 4,036 distinct proteins. These proteins mapped to 3,931 locations in the mouse genome (different Ensembl IDs). Identified peptides and proteins are listed in supplemental Tables 2 and 3.

The overlap of protein identifications between the subcellular compartments was surprisingly high (Fig. 2*B*). More than half of all proteins were identified in all three compartments, and only 20% were found solely in one compartment. Visual inspection of the subcellular fractionation, however, indicated good separation. The histone bands, for example, appear to be unique to the chromatin/membrane fraction (Fig. 2*A*). To resolve this apparent discrepancy and to gain insight into the subcellular distribution of the mouse ES cell proteome, we then quantified all peptide signals across the three fractions whether they were sequenced or not. This was aided by the very high peptide mass accuracy, which facilitated matching of peptides between runs (20). In this way, we obtained the percentage of protein present in each fraction, which we then used for hierarchically clustering (Fig. 2*C*). Three major clusters emerged (labeled *A*, *B*, and *C* in the figure). GO enrichment analysis of cluster B revealed significant overrepresentation for membrane-bound organelle, mitochondria, nucleus, nucleolus, and related terms ($p < 10^{-21}$ for each category). As can be seen in Fig. 2*B*, cluster B encompassed proteins quantified as most abundant in the chromatin/membrane fraction, unambiguously supporting the success of the cellular fractionation. Likewise proteins from cluster C were by far most abundant in the nucleoplasmic fraction, and this cluster was overrepresented in nucleus, chromosome, nucleoplasm, spliceosome, etc. ($p < 10^{-15}$ for each category). Finally cluster A (most abundant in the cytosolic fraction) was overrepresented in cytoplasm and cytosol ($p < 10^{-48}$). The complete list of overrepresented GO terms for all clusters is shown in supplemental Table 4, and the percent distribution of each protein between subcellular fractions is shown in supplemental Table 3.

The above analysis shows that the subcellular fractionation indeed performed as expected with cytosolic, nucleoplasmic, and chromatin proteins most abundant in the appropriate fractions. Nevertheless a small fraction of these proteins was also found in the other compartments. Due to the high sensitivity of LC-MS/MS, for most proteins this is sufficient for identification.

Analysis of the ES Proteome by Isoelectric Focusing of Peptides—In two-dimensional gel electrophoresis, proteins are first separated according to their isoelectric point using IPG strips (35). In principle, peptides can also be separated on these strips. In a recently introduced commercial instrument, the OFFGEL Fractionator (Agilent), the IPG strip connects 24 solvent-filled reservoirs. During isoelectric focusing peptides migrate to the appropriate reservoir and can easily be retrieved from solution (36, 37). Here we wanted to evaluate this relatively new technology for large scale proteome analysis and to complement our 1D gel-based method with a completely different separation approach.

We applied in-solution digested whole ES cell extract to the instrument and separated peptides for 50 kV-h. Each of the 24 resulting peptide fractions was cleaned up on StageTips (15) and analyzed by standard on-line HPLC-MS/MS (see "Experimental Procedures"). From the 264,372 tandem mass spectra acquired, we identified a total of 27,362 unique peptides with an average absolute mass accuracy of 559 ppb (S.D. 476 ppb) using the same stringency as described above for the GeLCMS analysis (supplemental Table 6). This yielded 3,972 proteins, which mapped to 3,892 different Ensembl entries (supplemental Table 7).

OFFGEL analysis identified almost the same number of

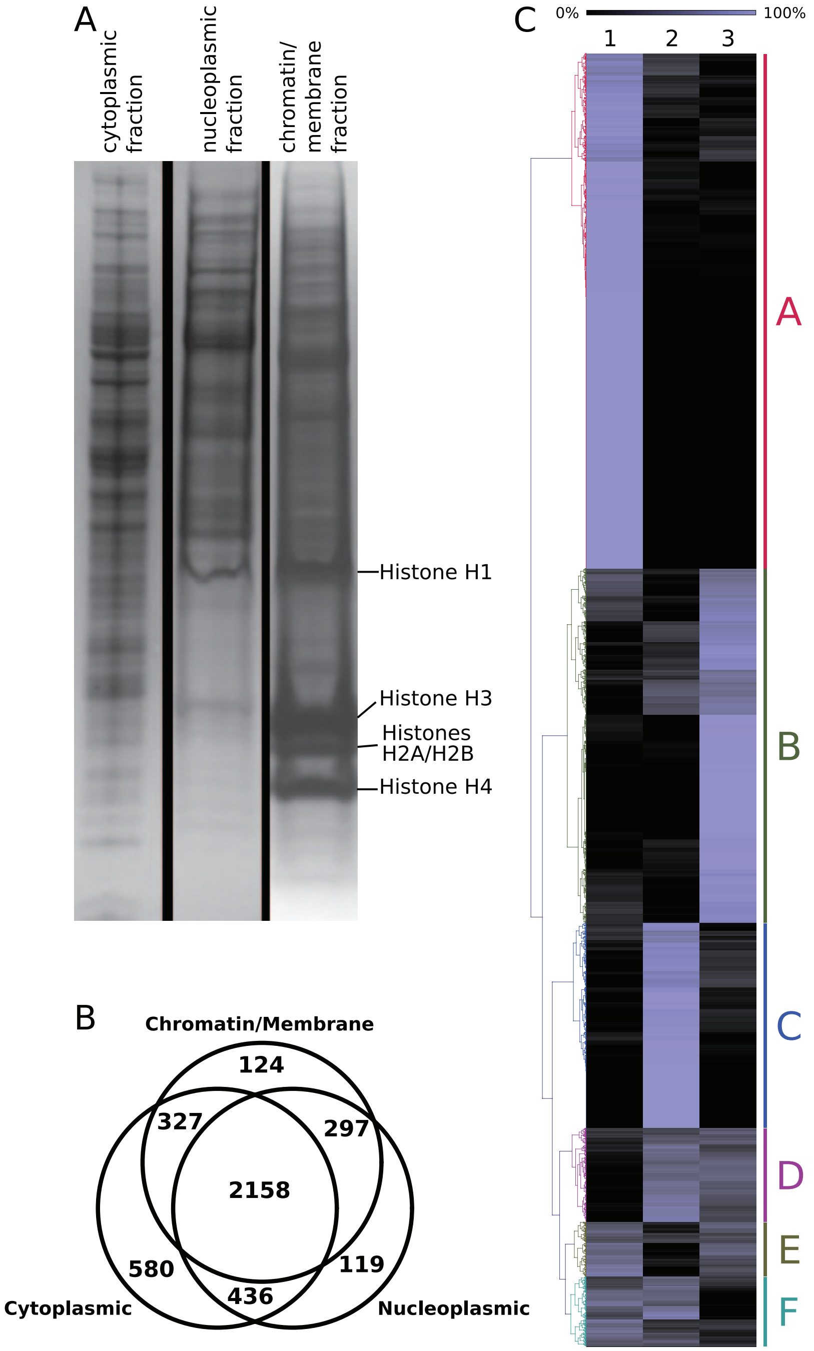

FIG. 2. **Subcellular fractionation of G-Olig2 ES cells.** *A*, Coomassie-stained gel of subcellular fractions; note the separation of histones. *B*, Venn diagram representing how subcellular fractions contribute to total protein identifications. *C*, clustering of protein groups retrieved according to their total peptide signal (normalized extracted ion current). The clustered groups are labeled by *letters* (*A–F*) according to visual inspection: *1*, cytoplasmic fraction; *2*, nucleoplasmic fraction; *3*, chromatin/membrane fraction. See text for proteins overrepresented in clusters A, B, and C.

proteins as the GeLCMS analysis combined with subcellular fractionation (3,972 *versus* 4,036). This is intriguing because the OFFGEL approach involved less sample preparation and only about half the mass spectrometric analysis time (24 compared with 45 LC-MS/MS runs). Furthermore GO analysis showed that essentially all categories are covered equally well by both approaches.

The Mouse ES Cell Proteome at a Depth of More than 5000 Proteins—We combined the two large scale experiments described above to arrive at a high confidence proteome of mouse ES cells. All raw MS files were imported into the MaxQuant software together and analyzed as a whole using uniform statistical criteria, in particular the requirement for two fully tryptic peptides in the correct SILAC states with very low mass deviation and a 99% certainty of identification at the protein level as assessed by reverse database searching. In this way, we arrived at 781,021 tandem mass spectra, resulting in 49,445 unique peptide sequences with an average absolute mass error of 400 ppb (S.D. 400 ppb; supplemental Table 9). This yielded a mouse ES cell proteome of 5,111 proteins (supplemental Table 10; comprising all identified proteins but excluding common contaminants such as human

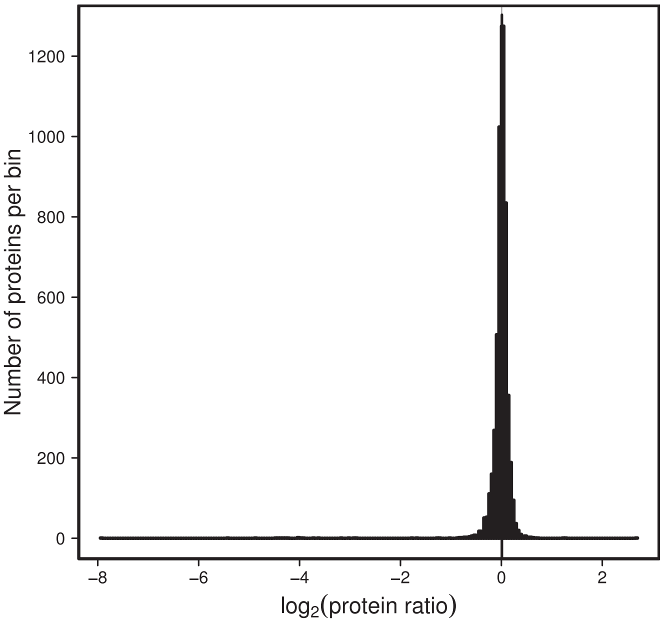

FIG. 3. **Quantitation of the ES cell proteome.** The figure shows the $\log_2$-transformed protein ratios. Protein ratios have a median of 0 on the log scale (*dashed line*) as expected for a 1:1 mixture and cluster tightly around the median.

keratins, BSA, and trypsin). These proteins map to 4,972 distinct locations in the mouse genome. Thus ES cells express at least about a quarter of the genes in the genome. Fig. 3 demonstrates quantitation of more than 5,000 proteins in an equal mixture of the heavy and light mouse ES cell proteome. As can be seen in the figure, protein ratios are distributed closely around the expected 1:1 value.

We first checked the quantified proteome for the presence of known stem cell markers. We found OCT4 (38) with seven peptides, SOX2 (39) with nine peptides, and NANOG (40, 41) with two peptides (Fig. 4). These three “master regulators” are intimately involved in the maintenance of stemness, and loss of their expression is concomitant with exit from the pluripotent state. The presence of these factors in our proteome suggests that SILAC-labeled mouse ES cells retain stemness. We did not detect SALL4 (42) and the very recently discovered DPPA2 and DPPA4 (43), known stem cell markers that are presumably expressed in the mouse ES cells investigated here. This is most likely due to their low abundance. Table I lists these factors as well as others that have been identified here and designated “stem cell-specific” in the literature. However, several proteomics studies use this term for proteins that are clearly not exclusive to stem cells, such as proteasome subunits and alkaline phosphatases (8), and these are not listed in the table.

To further evaluate the completeness of coverage we determined the number of protein kinases and transcription factors in our data set. We found 156 protein kinases (GO Term 0004672 protein kinase activity) and 131 transcription factors (GO Term 0003700 transcription factor activity). These

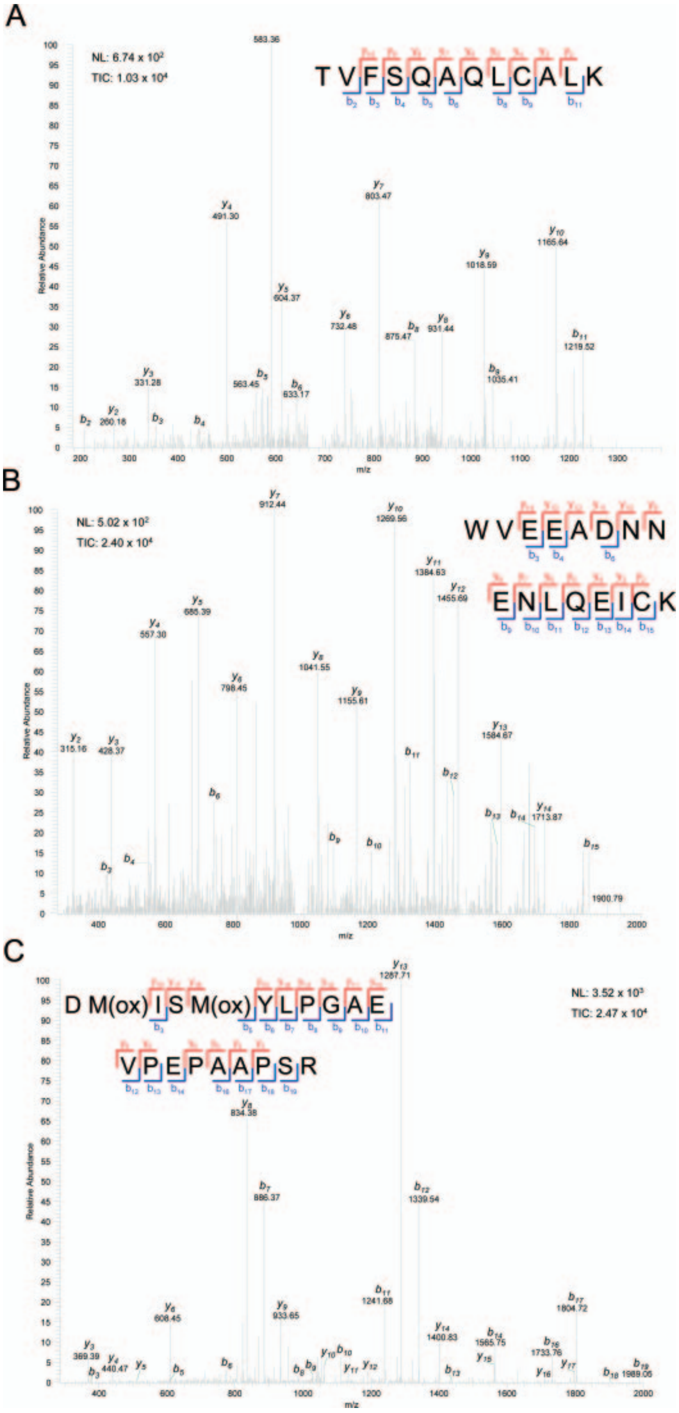

FIG. 4. **Fragmentation spectra of master stem cell regulators.** *A*, one of the tandem mass spectra identifying NANOG. The spectrum is labeled with b and y ions from the identified sequence shown in the *inset*. For explanation of fragmentation scheme see Ref. 59. *B*, one of the tandem mass spectra identifying OCT4. *C*, one of the spectra identifying SOX2. *M(ox)* signifies oxidized methionine. *TIC*, total ion current; *NL*, neutral loss.

are 4.1 and 3.5% of all proteins identified. For kinases this is the same proportion as annotated (4.2%), whereas for transcription factors it is slightly less than the 5% annotated in the

TABLE I
ES cell-specific markers

Subcell. fract., subcellular fractionation; Norm., normalized; H, heavy; L, light.

Stem cell marker	UniProt ID (IPI)	Ref.	Experiment					
			Subcell. fract., GeLCMS		OFFGEL, LC/MS		Combined analysis	
			Peptides: all (unique)	Norm. ratio H/L	Peptides: all (unique)	Norm. ratio H/L	Peptides: all (unique)	Norm. ratio H/L
				%		%		%
Catenin α-1	P26231	60	31 (26)	1.03 ± 7.03	29 (29)	1.05 ± 5.51	37 (30)	1.04 ± 6.02
ERAS	Q7TN89	61	2 (2)	0.9 ± 25.7	5 (5)	1.07 ± 6.38	5 (5)	1.04 ± 9.35
ESG1	Q9CQS7	62	4 (4)	1.22 ± 6.3	6 (6)	1.04 ± 16.26	8 (8)	1.16 ± 11.37
ESRRB	Q61539	63	13 (12)	0.83 ± 5.58	5 (5)	0.99 ± 4.25	14 (13)	0.85 ± 6.42
FGF4	P11403	64	1 (1)[a,b]	0.95 ± 14.1	—[c]	—	1 (1)	1.1
NANOG	Q80Z64	40, 41	1 (1)[a]	0.95 ± 3.01	1 (1)[a]	NA[d]	2 (2)	0.97 ± 1.94
OCT4	P20263	38	4 (4)	1.11 ± 5.71	7 (7)	1.12 ± 6.38	7 (7)	1.1 ± 5.22
REX1	P22227	65	4 (3)	1.05 ± 8.37	—	—	4 (3)	0.98 ± 10.64
RIF1	Q62521	9, 66	2 (2)	1.05 ± 6.26	—	—	74 (74)	1 ± 5.12
RNF2	Q9CQJ4	9, 66	6 (6)	1.01 ± 5.23	5 (5)	1.01 ± 4.63	9 (9)	1.01 ± 4.52
SOX2	P48432	39	6 (6)	0.99 ± 4.59	4 (4)	0.88 ± 14.43	9 (9)	0.98 ± 6.08
STELLA	Q8QZY3	67	1 (1)[a]	1.46	1 (1)[a]	1.49	2 (2)	1.47 ± 1.26
TCL1	P56280	63	2 (2)	1.07 ± 3.57	—	—	2 (2)	1.13 ± 5.7
UTF1	O70530	68	10 (10)	1.09 ± 6.85	2 (2)	1.03 ± 2.68	10 (10)	1.1 ± 4.31
ZIC3	Q62521	69	2 (2)	1.05 ± 6.26	—	—	2 (2)	1.03 ± 7

[a] Supporting material to the single peptide identifications is in supplemental Material 2.
[b] This single peptide identification is not part of the final ES proteome protein count.
[c] —, not detected.
[d] Not applicable.

complete mouse genome. Taken together, these observations suggest that we covered the mouse ES cell proteome in considerable but not yet complete depth.

We analyzed the obtained ES cell proteome for over- and underrepresented categories by GO using GOSlim (see "Experimental Procedures"). Overall there were few categories significantly differently populated in the proteome compared with the entire mouse genome. Some underrepresented terms include receptor activity, signal transducer activity, cell communication, signal transduction, and extracellular region (supplemental Table 11). Unfortunately at this point it is difficult to determine whether this underrepresentation was due to experimental design because our fractionation did not include a specific plasma membrane preparation or whether ES cells really express fewer of the proteins that somatic cells need to communicate with each other. Several categories were significantly overrepresented (supplemental Table 11). These include cell cycle, DNA metabolism, biosynthesis, and other categories related to cell growth and division. This shows that ES cells are very actively engaged in proliferation, which correlates well with their short doubling times.

Microarray studies provide an estimate of the transcript (mRNA) levels in a particular biological state at any given time and have so far been the predominant technology to study various aspects of murine ES cell biology (32, 44–46). As proteomics measures protein expression including translational and post-translational regulations, we explored the quantitative and qualitative overlap between a recent mRNA microarray study by Hailesellasse Sene *et al.* (32) and our proteome data set. We chose that particular study because the cell line and experimental conditions used matched closely with our proteome analysis protocol. The data are of high quality as assessed from the expression correlation and box plots of the triplicates for each chip (provided as supplemental Fig. 1). The 7,926 probe sets deemed "present" (see "Experimental Procedures") correspond to 5,490 unique Entrez identifiers of which we were able to map 3,322 to our proteome data set. Fig. 5*A* depicts the overlap between the proteome and mRNA data sets and shows that proteomic coverage compares favorably with gene expression given criteria of similar stringency. We recently reported a very similar finding in a study of the HeLa cell proteome (6). mRNA expression correlates moderately with protein expression (Pearson correlation coefficient of 0.43; Fig. 5*B*). This suggests that in general steady state protein expression is not in direct stoichiometric relationship with the gene expression and rather results from the complex interplay of regulation on the transcriptional, translational, and post-translational levels. Unraveling contributions of the different regulatory processes is beginning to be feasible by proteomics methods (47) but is beyond the scope of this study.

The epigenetic state of ES cells is of central interest with regard to their pluripotent state and loss thereof during differentiation (48). In particular, the N-terminal tails of histones carry post-translational modifications that are known to correlate with transcriptional activity of the locus that is modified (49, 50). Very recently, a number of studies have described the genome-wide detection of active, repressive, and bivalent histone marks in mouse ES cells. These marks are histone 3 lysine 4 trimethylation (H3K4me3), histone 3 lysine 27 trim-

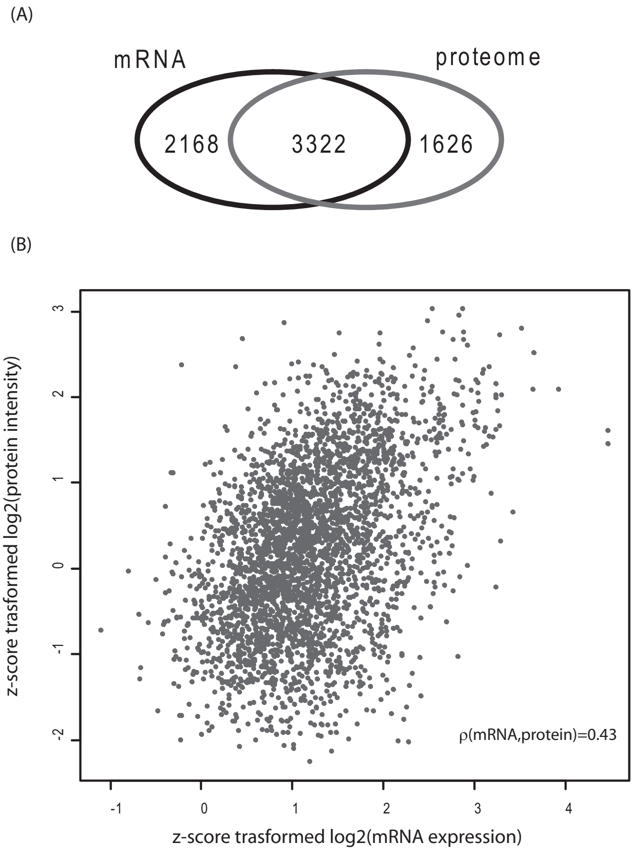

FIG. 5. **Correlation of mRNA expression with the proteome of ES cells.** *A*, Venn diagram representing the overlap between Entrez ID mapped mRNA probe sets deemed present ($p \geq 0.01$, P call ≥66%, see "Experimental Procedures") from a recent ES cell study (32) and the Entrez ID mapped combined proteome (false positive rate ≤0.01). *B*, correlation of *z*-score-transformed summed protein intensity (extracted ion current) with *z*-score-transformed mRNA expression.

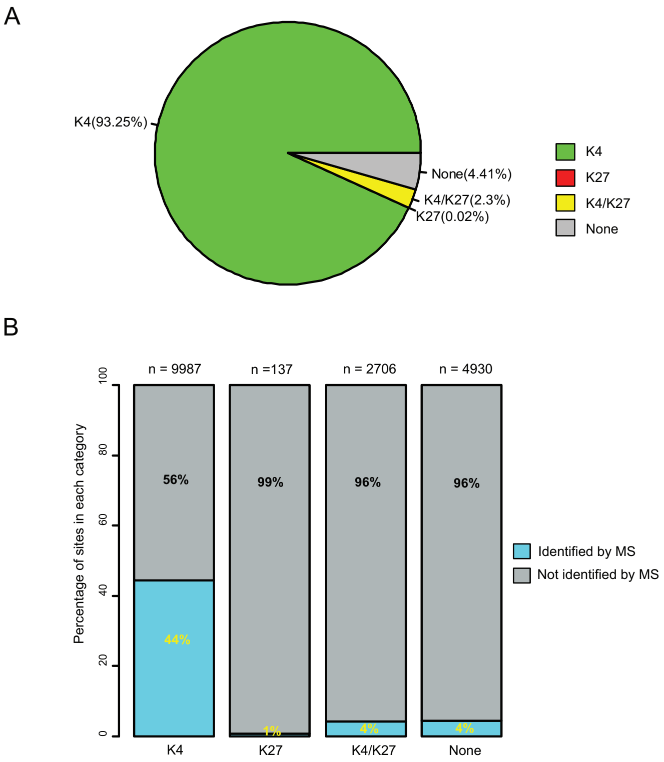

FIG. 6. **Correlation of chromatin state with the proteome of ES cells.** *A*, distribution of activating (*K4*), repressing (*K27*), and bivalent markers (*K4/K27*) in the ES proteome data set (comparison with Mikkelsen *et al.* (4)). The vast majority of detected ES cell proteins have an activating histone mark in the promoter region of the corresponding gene, and only one has only the repressive K27 mark. *B*, proportion of detected proteins for genes with activating (K4), repressive (K27), and bivalent (K4/K27) chromatin marks. The column labeled "none" refers to genes in which none of the two marks was found. The number of genes in each category is indicated on *top* of the *bar*.

ethylation (H3K27me3), and H3K4me3 together with H3K27me3, respectively. The presence of these marks on stem cell promoters should correlate with our observed proteome. Genes whose protein product is detected should have active histone marks, whereas proteins that are not expressed should carry repressor marks. We compared our data set against the data set of Mikkelsen *et al.* (4), who used chromatin immunoprecipitation together with large scale sequencing of the occupied DNA region (ChIPseq). For the vast majority of proteins detected in our study (93%), the activating H3K4me3 mark was indeed present on the corresponding gene (Fig. 6). Another 2% (108 proteins) had the bivalent mark thought to be present on genes needed for differentiation and poised for transcription (48). Interestingly GO enrichment analysis using GOSlim on these 108 proteins revealed significant overrepresentation of categories potentially involved in these processes, namely morphogenesis and cell development ($p < 0.001$). Strikingly only one of the proteins detected in our ES cell proteome had a repressive mark. If the ChIPseq or the proteomics data had been random 60 proteins containing a repressive mark should have been detected. Furthermore the one detected protein whose promoter had a repressive mark encodes for Calponin-1, a protein reported to be highly expressed in mesenchymal stem cells upon mechanical strain (51). Finally we identified 207 proteins for which no data had been obtained in the genome-wide chromatin ChIPseq experiment. Conversely the ChIPseq study found 5,616 genes with activating marks for which we did not identify the corresponding protein product. Many of the genes in this set may not actually be expressed as proteins, and the data set may contain false positives for the ChIPseq study and false negatives for our proteome study (for example, proteins with extremely low expression level).

DISCUSSION

In this study, we evaluated several ways to SILAC label mouse ES cells. We found that growing the cells for two passages on feeder cells followed by three passages in BMP4-supplemented, feeder-free conditions led to essentially complete incorporation (median value of 97%). We then used this SILAC condition to analyze the mouse ES cell proteome in depth with two different approaches. Although we did not use SILAC to quantify two different states against each other, the one-to-one mixtures analyzed here greatly aided in establishing a high quality proteome. SILAC distinguished peptides from non-peptide peaks and noise and yielded the number of arginines and lysines for each peptide, which substantially decreased the search space in database matching and thereby increases the number of statistically significant peptide identifications (20). Furthermore we demonstrated here that more than 5,000 proteins can not only be identified but also quantified in a single cell type, making this the largest study of its kind to date.

We used two methods for large scale proteome analysis. First we combined a standard cell fractionation protocol with 1D gel electrophoresis and analysis of 45 gel slices by LC-MS/MS. Qualitative analysis showed that most proteins were identified in all three subcellular compartments, and only a small proportion were identified in a single fraction. We then performed a quantitative analysis by summing the peptide signals for each protein in the three cell fractions. In this way, we obtained an intensity profile of each protein in each of the fractions. The quantitative analysis clearly showed that proteins are distributed as expected from their intracellular location. However, the benefit of subcellular fractionation for additional protein identification is not as great as might be expected because the high sensitivity of modern MS methods means that a low percentage of proteins from a different compartment will still be identified. Additionally our analysis showed that purely qualitative interpretation of the results of subcellular fractionation is likely to be misleading. However, the subcellular fractionation did increase dynamic range in each fraction as well as peptide sequence coverage. The main use of subcellular fractionation in proteomics will be in learning about protein localization, which can be achieved by methods such as protein correlation profiling (52, 53). Here we have, for the first time, comprehensively determined the percentage distribution of more than 4,000 proteins between three cellular fractions.

In a second approach to the characterization of the mouse ES cell proteome, we digested the proteome in-solution, separated the resulting tryptic peptides by isoelectric focusing in the OFFGEL apparatus followed by 24 LC-MS/MS runs. This analysis yielded almost as many proteins as the cell fractionation and GeLCMS approach at a considerable time saving in sample preparation and analysis time. This is mainly due to less redundancy in the OFFGEL fractions compared with the subcellular fractionation-GeLCMS experiment as also evident from the substantially lower number of required MS/MS events. Although more detailed evaluation still needs to be performed, we conclude that the OFFGEL approach is very promising for complex proteome characterization.

The mouse ES cell proteome reported here is as least as complex as any other cell type that we have investigated in this laboratory. Although it was already known that the transcriptome of ES cells is very complex, it was possible that ES cells store many messages that would only be translated upon differentiation. Because we measured a very diverse ES cell proteome, our results now make this hypothesis unlikely.

Our ES cell proteome contains most of the well known stem cell markers, arguing that the SILAC technology is well suited to the quantitative analysis of markers during differentiation. The number of regulatory proteins quantified is similar to the number expected from the theoretical proteome as a whole. Together these observations argue that we covered the stem cell proteome in considerable depth and without obvious bias. Nevertheless several stem cell markers were still missing, and protein identification on our data set using less stringent criteria showed evidence for the presence of at least another 1,000 proteins. Thus further technology development is still needed for more comprehensive coverage of the ES cell proteome. This will especially be true for the quantitation of ES cell-specific protein isoforms, some of which, such as ERAS, we already detected here, and for the quantitation of regulatory modifications in the ES cell proteome. Compared with other “omics” approaches, such as microarray analysis of ES cells (54), however, we believe that quantitative proteomics is already similarly comprehensive and potentially much more quantitative. This is also the conclusion we previously reached when comparing the HeLa cell proteome and the transcriptome detected in microarray experiments (6).

The SILAC-labeled cells described here can be used in two ways in proteomics studies. In the first approach, one ES cell population can be differentially modified with respect to the other, and differences in the proteome can be directly quantified. For example, obligate stem cell factors can be knocked down by small interfering RNA, and the differentiation response can be followed. In a second approach one would produce a large quantity of fully labeled ES cells and then use them as internal standards for proteomics studies of ES cells. In this format, an equal amount of SILAC-labeled ES cells would be added to experiment and control or to the samples in a time course experiment. This would have the advantage that standard protocols could be used and no special care would have to be taken for SILAC conditions.

The question of what constitutes an ES cell has recently become even more interesting in light of reports on the “reprogramming” of terminally differentiated fibroblasts into pluripotent ES-like cells (55–57). We hope that quantitative proteomics can shed light on such events in the future just as has already been demonstrated for the differentiation of adult stem cells (58).

Acknowledgments—We thank our colleagues for fruitful discussions and help, especially Michael Sixt for advice on stem cell culture, Reinhard Faessler for use of facilities, Michiel Vermeulen for critical reading of the manuscript, and Peter Bandilla for exceptional technical support.

* This work was supported in part by the European Union Grant High-throughput Epigenetic Regulatory Organisation In Chromatin (HEROIC). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[S] The on-line version of this article (available at http://www.mcponline.org) contains supplemental material.

§ Both authors contributed equally to this work.

¶ Supported by the European Network Grant RUBICON.

** Supported by the Interdepartmental Graduate Program for Experimental Life Sciences (IGEL) program (University of Münster).

§§ To whom correspondence should be addressed. Tel.: 49-89-8578-2557; Fax: 49-89-8578-3209; E-mail: mmann@biochem.mpg.de.

REFERENCES

1. Wobus, A. M., and Boheler, K. R. (2005) Embryonic stem cells: prospects for developmental biology and cell therapy. *Physiol. Rev.* **85,** 635–678
2. Robson, P. (2004) The maturing of the human embryonic stem cell transcriptome profile. *Trends Biotechnol.* **22,** 609–612
3. Araki, R., Fukumura, R., Sasaki, N., Kasama, Y., Suzuki, N., Takahashi, H., Tabata, Y., Saito, T., and Abe, M. (2006) More than 40,000 transcripts, including novel and noncoding transcripts, in mouse embryonic stem cells. *Stem Cells* (*Dayton*) **24,** 2522–2528
4. Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T. K., Koche, R. P., Lee, W., Mendenhall, E., O'Donovan, A., Presser, A., Russ, C., Xie, X., Meissner, A., Wernig, M., Jaenisch, R., Nusbaum, C., Lander, E. S., and Bernstein, B. E. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448,** 553–560
5. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422,** 198–207
6. Cox, J., and Mann, M. (2007) Is proteomics the new genomics? *Cell* **130,** 395–398
7. Baharvand, H., Fathi, A., van Hoof, D., and Salekdeh, G. H. (2007) Concise review: trends in stem cell proteomics. *Stem Cells* (*Dayton*) **25,** 1888–1903
8. Nagano, K., Taoka, M., Yamauchi, Y., Itagaki, C., Shinkawa, T., Nunomura, K., Okamura, N., Takahashi, N., Izumi, T., and Isobe, T. (2005) Large-scale identification of proteins expressed in mouse embryonic stem cells. *Proteomics* **5,** 1346–1361
9. van Hoof, D., Passier, R., Ward-Van Oostwaard, D., Pinkse, M. W., Heck, A. J., Mummery, C. L., and Krijgsveld, J. (2006) A quest for human and mouse embryonic stem cell-specific proteins. *Mol. Cell. Proteomics* **5,** 1261–1273
10. Ong, S. E., and Mann, M. (2005) Mass spectrometry-based proteomics turns quantitative. *Nat. Chem. Biol.* **1,** 252–262
11. Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., and Mann, M. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1,** 376–386
12. Mann, M. (2006) Functional and quantitative proteomics using SILAC. *Nat. Rev.* **7,** 952–958
13. Xian, H. Q., McNichols, E., St Clair, A., and Gottlieb, D. I. (2003) A subset of ES-cell-derived neural cells marked by gene targeting. *Stem Cells* (*Dayton*) **21,** 41–49
14. Wessel, D., and Flugge, U. I. (1984) A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Anal. Biochem.* **138,** 141–143
15. Rappsilber, J., Ishihama, Y., and Mann, M. (2003) Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* **75,** 663–670
16. Dignam, J. D., Lebovitz, R. M., and Roeder, R. G. (1983) Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. *Nucleic Acids Res.* **11,** 1475–1489
17. Shevchenko, A., Tomas, H., Havlis, J., Olsen, J. V., and Mann, M. (2006) In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protoc.* **1,** 2856–2860
18. Olsen, J. V., Ong, S. E., and Mann, M. (2004) Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol. Cell. Proteomics* **3,** 608–614
19. Olsen, J. V., de Godoy, L. M., Li, G., Macek, B., Mortensen, P., Pesch, R., Makarov, A., Lange, O., Horning, S., and Mann, M. (2005) Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol. Cell. Proteomics* **4,** 2010–2021
20. Cox, J., de Godoy, L., de Souza, G., Olsen, J. V., Ren, S., and Mann, M. (2007) Bioinformatics algorithms and software enabling whole proteome quantitation applied to diploid vs. haploid yeast cells, in *55th ASMS Conference on Mass Spectrometry, Indianapolis, June 3–7, 2007*, ThOB pm-04:10, American Society for Mass Spectrometry, Santa Fe, NM
21. Gingras, A. C., Gstaiger, M., Raught, B., and Aebersold, R. (2007) Analysis of protein complexes using mass spectrometry. *Nat. Rev.* **8,** 645–654
22. Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E., and Apweiler, R. (2004) The International Protein Index: an integrated database for proteomics experiments. *Proteomics* **4,** 1985–1988
23. Moore, R. E., Young, M. K., and Lee, T. D. (2002) Qscore: an algorithm for evaluating SEQUEST database search results. *J. Am. Soc. Mass Spectrom.* **13,** 378–386
24. Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J., and Gygi, S. P. (2003) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* **2,** 43–50
25. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567
26. Zubarev, R., and Mann, M. (2007) On the proper use of mass accuracy in proteomics. *Mol. Cell. Proteomics* **6,** 377–381
27. Sturn, A., Quackenbush, J., and Trajanoski, Z. (2002) Genesis: cluster analysis of microarray data. *Bioinformatics* (*Oxf.*) **18,** 207–208
28. Maere, S., Heymans, K., and Kuiper, M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* (*Oxf.*) **21,** 3448–3449
29. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13,** 2498–2504
30. Adachi, J., Kumar, C., Zhang, Y., Olsen, J. V., and Mann, M. (2006) The human urinary proteome contains more than 1500 proteins, including a large proportion of membrane proteins. *Genome Biol.* **7,** R80
31. R Development Core Team (2004) *R: a Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria
32. Hailesellasse Sene, K., Porter, C. J., Palidwor, G., Perez-Iratxeta, C., Muro, E. M., Campbell, P. A., Rudnicki, M. A., and Andrade-Navarro, M. A. (2007) Gene function in early mouse embryonic stem cell differentiation. *BMC Genomics* **8,** 85
33. Ying, Q. L., Nichols, J., Chambers, I., and Smith, A. (2003) BMP induction of Id proteins suppresses differentiation and sustains embryonic stem cell self-renewal in collaboration with STAT3. *Cell* **115,** 281–292
34. van Hoof, D., Pinkse, M. W., Oostwaard, D. W., Mummery, C. L., Heck, A. J., and Krijgsveld, J. (2007) An experimental correction for arginine-to-proline conversion artifacts in SILAC-based quantitative proteomics. *Nat. Methods* **4,** 677–678
35. Gorg, A., Obermaier, C., Boguth, G., Harder, A., Scheibe, B., Wildgruber, R., and Weiss, W. (2000) The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis* **21,** 1037–1053
36. Heller, M., Ye, M., Michel, P. E., Morier, P., Stalder, D., Junger, M. A., Aebersold, R., Reymond, F., and Rossier, J. S. (2005) Added value for tandem mass spectrometry shotgun proteomics data validation through isoelectric focusing of peptides. *J. Proteome Res.* **4,** 2273–2282
37. Horth, P., Miller, C. A., Preckel, T., and Wenz, C. (2006) Efficient fractionation and improved protein identification by peptide OFFGEL electrophoresis. *Mol. Cell. Proteomics* **5,** 1968–1974
38. Scholer, H. R., Dressler, G. R., Balling, R., Rohdewohld, H., and Gruss, P. (1990) Oct-4: a germline-specific transcription factor mapping to the mouse t-complex. *EMBO J.* **9,** 2185–2195
39. Yuan, H., Corbi, N., Basilico, C., and Dailey, L. (1995) Developmental-specific activity of the FGF-4 enhancer requires the synergistic action of Sox2 and Oct-3. *Genes Dev.* **9,** 2635–2645
40. Chambers, I., Colby, D., Robertson, M., Nichols, J., Lee, S., Tweedie, S.,

and Smith, A. (2003) Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. *Cell* **113,** 643–655

41. Mitsui, K., Tokuzawa, Y., Itoh, H., Segawa, K., Murakami, M., Takahashi, K., Maruyama, M., Maeda, M., and Yamanaka, S. (2003) The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell* **113,** 631–642
42. Zhang, J., Tam, W. L., Tong, G. Q., Wu, Q., Chan, H. Y., Soh, B. S., Lou, Y., Yang, J., Ma, Y., Chai, L., Ng, H. H., Lufkin, T., Robson, P., and Lim, B. (2006) Sall4 modulates embryonic stem cell pluripotency and early embryonic development by the transcriptional regulation of Pou5f1. *Nat. Cell Biol.* **8,** 1114–1123
43. Maldonado-Saldivia, J., van den Bergen, J., Krouskos, M., Gilchrist, M., Lee, C., Li, R., Sinclair, A. H., Surani, M. A., and Western, P. S. (2007) Dppa2 and Dppa4 are closely linked SAP motif genes restricted to pluripotent cells and the germ line. *Stem Cells* (*Dayton*) **25,** 19–28
44. Ivanova, N. B., Dimos, J. T., Schaniel, C., Hackney, J. A., Moore, K. A., and Lemischka, I. R. (2002) A stem cell molecular signature. *Science* **298,** 601–604
45. Ramalho-Santos, M., Yoon, S., Matsuzaki, Y., Mulligan, R. C., and Melton, D. A. (2002) “Stemness”: transcriptional profiling of embryonic and adult stem cells. *Science* **298,** 597–600
46. Tesar, P. J., Chenoweth, J. G., Brook, F. A., Davies, T. J., Evans, E. P., Mack, D. L., Gardner, R. L., and McKay, R. D. (2007) New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature* **448,** 196–199
47. Lu, P., Vogel, C., Wang, R., Yao, X., and Marcotte, E. M. (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.* **25,** 117–124
48. Spivakov, M., and Fisher, A. G. (2007) Epigenetic signatures of stem-cell identity. *Nat. Rev. Genet.* **8,** 263–271
49. Jenuwein, T., and Allis, C. D. (2001) Translating the histone code. *Science* **293,** 1074–1080
50. Kouzarides, T. (2007) Chromatin modifications and their function. *Cell* **128,** 693–705
51. Kurpinski, K., Chu, J., Hashi, C., and Li, S. (2006) Anisotropic mechanosensing by mesenchymal stem cells. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 16095–16100
52. Andersen, J. S., Wilkinson, C. J., Mayor, T., Mortensen, P., Nigg, E. A., and Mann, M. (2003) Proteomic characterization of the human centrosome by protein correlation profiling. *Nature* **426,** 570–574
53. Foster, L. J., de Hoog, C. L., Zhang, Y., Xie, X., Mootha, V. K., and Mann, M. (2006) A mammalian organelle map by protein correlation profiling. *Cell* **125,** 187–199
54. Evsikov, A. V., and Solter, D.(2003) Comment on “‘Stemness’: transcriptional profiling of embryonic and adult stem cells” and “a stem cell molecular signature”. *Science* **302,** 393, author reply 393
55. Meissner, A., Wernig, M., and Jaenisch, R. (2007) Direct reprogramming of genetically unmodified fibroblasts into pluripotent stem cells. *Nat. Biotechnol.* **25,** 1177–1181
56. Okita, K., Ichisaka, T., and Yamanaka, S. (2007) Generation of germline-competent induced pluripotent stem cells. *Nature* **448,** 313–317
57. Wernig, M., Meissner, A., Foreman, R., Brambrink, T., Ku, M., Hochedlinger, K., Bernstein, B. E., and Jaenisch, R. (2007) In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. *Nature* **448,** 318–324
58. Kratchmarova, I., Blagoev, B., Haack-Sorensen, M., Kassem, M., and Mann, M. (2005) Mechanism of divergent growth factor effects in mesenchymal stem cell differentiation. *Science* **308,** 1472–1477
59. Steen, H., and Mann, M. (2004) The ABC’s (and XYZ’s) of peptide sequencing. *Nat. Rev.* **5,** 699–711
60. Torres, M., Stoykova, A., Huber, O., Chowdhury, K., Bonaldo, P., Mansouri, A., Butz, S., Kemler, R., and Gruss, P. (1997) An alpha-E-catenin gene trap mutation defines its function in preimplantation development. *Proc. Natl. Acad. Sci. U. S. A.* **94,** 901–906
61. Takahashi, K., Mitsui, K., and Yamanaka, S. (2003) Role of ERas in promoting tumour-like properties in mouse embryonic stem cells. *Nature* **423,** 541–545
62. Western, P., Maldonado-Saldivia, J., van den Bergen, J., Hajkova, P., Saitou, M., Barton, S., and Surani, M. A. (2005) Analysis of Esg1 expression in pluripotent cells and the germline reveals similarities with Oct4 and Sox2 and differences between human pluripotent cell lines. *Stem Cells* (*Dayton*) **23,** 1436–1442
63. Ivanova, N., Dobrin, R., Lu, R., Kotenko, I., Levorse, J., DeCoste, C., Schafer, X., Lun, Y., and Lemischka, I. R. (2006) Dissecting self-renewal in stem cells with RNA interference. *Nature* **442,** 533–538
64. Niswander, L., and Martin, G. R. (1992) Fgf-4 expression during gastrulation, myogenesis, limb and tooth development in the mouse. *Development* (*Camb.*) **114,** 755–768
65. Rogers, M. B., Hosler, B. A., and Gudas, L. J. (1991) Specific expression of a retinoic acid-regulated, zinc-finger gene, Rex-1, in preimplantation embryos, trophoblast and spermatocytes. *Development* (*Camb.*) **113,** 815–824
66. Wang, J., Rao, S., Chu, J., Shen, X., Levasseur, D. N., Theunissen, T. W., and Orkin, S. H. (2006) A protein interaction network for pluripotency of embryonic stem cells. *Nature* **444,** 364–368
67. Payer, B., Saitou, M., Barton, S. C., Thresher, R., Dixon, J. P., Zahn, D., Colledge, W. H., Carlton, M. B., Nakano, T., and Surani, M. A. (2003) Stella is a maternal effect gene required for normal early development in mice. *Curr. Biol.* **13,** 2110–2117
68. van den Boom, V., Kooistra, S. M., Boesjes, M., Geverts, B., Houtsmuller, A. B., Monzen, K., Komuro, I., Essers, J., Drenth-Diephuis, L. J., and Eggen, B. J. (2007) UTF1 is a chromatin-associated protein involved in ES cell differentiation. *J. Cell Biol.* **178,** 913–924
69. Lim, L. S., Loh, Y. H., Zhang, W., Li, Y., Chen, X., Wang, Y., Bakre, M., Ng, H. H., and Stanton, L. W. (2007) Zic3 is required for maintenance of pluripotency in embryonic stem cells. *Mol. Biol. Cell* **18,** 1348–1358