Proteomic and phosphoproteomic comparison of human ES and iPS cells

Douglas H Phanstiel[1,2,7], Justin Brumbaugh[2–4,7], Craig D Wenger[1,2], Shulan Tian[4], Mitchell D Probasco[4], Derek J Bailey[1,2], Danielle L Swaney[1,2], Mark A Tervo[1,2], Jennifer M Bolin[4], Victor Ruotti[4], Ron Stewart[4], James A Thomson[4–6] & Joshua J Coon[1–3]

Combining high-mass-accuracy mass spectrometry, isobaric tagging and software for multiplexed, large-scale protein quantification, we report deep proteomic coverage of four human embryonic stem cell and four induced pluripotent stem cell lines in biological triplicate. This 24-sample comparison resulted in a very large set of identified proteins and phosphorylation sites in pluripotent cells. The statistical analysis afforded by our approach revealed subtle but reproducible differences in protein expression and protein phosphorylation between embryonic stem cells and induced pluripotent cells. Merging these results with RNA-seq analysis data, we found functionally related differences across each tier of regulation. We also introduce the Stem Cell–Omics Repository (SCOR), a resource to collate and display quantitative information across multiple planes of measurement, including mRNA, protein and post-translational modifications.

For practical and ethical reasons, induced pluripotent stem cells (iPSCs) hold great potential for therapeutic and research purposes. Based on morphology, capacity to self-renew and developmental potential, iPSCs are nearly indistinguishable from their embryonic stem cell (ESC) counterparts[1–3], but their extent of similarity on the molecular level remains controversial[4–6]. Whereas various studies have stressed the overall similarity of gene expression programs between ESCs and iPSCs[1,2,5,7], a few studies have reported subtle differences in RNA levels, DNA methylation and the efficiency of many iPSC lines to differentiate into neural lineages[6,8–10]. Meanwhile, similarity of human ESCs and iPSCs at the protein level remains completely unexplored to our knowledge. These analyses are critical, as many forms of regulation are enforced post-transcriptionally or through post-translational modifications.

To address the proteomic and phosphoproteomic similarity between ESCs and iPSCs, we used a method that combines isobaric tagging, high-mass-accuracy mass spectrometry and a recently developed software tool. Applying this method to compare two ESC lines, one iPSC line and one fibroblast cell line, we identified 7,952 proteins and 10,499 phosphorylation sites (without any replicates). Leveraging the multiplex nature of our approach, we then examined proteins and their phosphorylation sites in four ESC lines and four iPSC lines in biological triplicate (24 samples total) and identified 6,761 proteins and 19,122 phosphorylation sites in total. Rigorous statistical analysis revealed significant ($P < 0.05$, Student's t-test) and functionally related differences between proteins and phosphorylation sites in human ESCs and iPSCs, which may reflect residual regulation characteristic of iPSCs' somatic origin. Finally, we introduce a searchable online resource, SCOR, for storing large-scale data related to pluripotency.

RESULTS

Peptide identification and quantification

Resonant-excitation, collision-activated dissociation imposes a fundamental low-mass cutoff that hinders the detection of reporter ions generated by isobaric tags. To remove this limitation we used beam-type CAD (HCD) with high-mass-accuracy detection of fragment ions[11–14]. These methods increase peptide identifications over 60% and phosphopeptide identifications over 260% compared to CAD with low-mass-accuracy detection fragment ions (**Fig. 1a**). We attribute these increases to greater specificity in database searches and fewer sequence-directed cleavage events. HCD is compatible with isobaric tagging strategies for multiplexed peptide quantification. Isobaric tags can be used to compare up to eight samples in a single experiment and facilitate analysis of biological replicates and multiple cell lines[15–17]. However, this form of quantification is subject to a unique and widespread quantitative error arising from the 'co-isolation' of multiple peptide precursors before fragmentation[18]. We therefore used recently developed software, TagQuant, which

[1]Department of Chemistry, University of Wisconsin, Madison, Wisconsin, USA. [2]Genome Center of Wisconsin, University of Wisconsin, Madison, Wisconsin, USA. [3]Department of Biomolecular Chemistry, University of Wisconsin, Madison, Wisconsin, USA. [4]Morgridge Institute for Research, Madison, Wisconsin, USA. [5]Department of Cell and Regenerative Biology, University of Wisconsin, Madison, Wisconsin, USA. [6]Department of Molecular, Cellular and Developmental Biology, University of California, Santa Barbara, California, USA. [7]These authors contributed equally to this work. Correspondence should be addressed to J.J.C. (jcoon@chem.wisc.edu).

RECEIVED 22 FEBRUARY; ACCEPTED 8 AUGUST; PUBLISHED ONLINE 11 SEPTEMBER 2011; DOI:10.1038/NMETH.1699

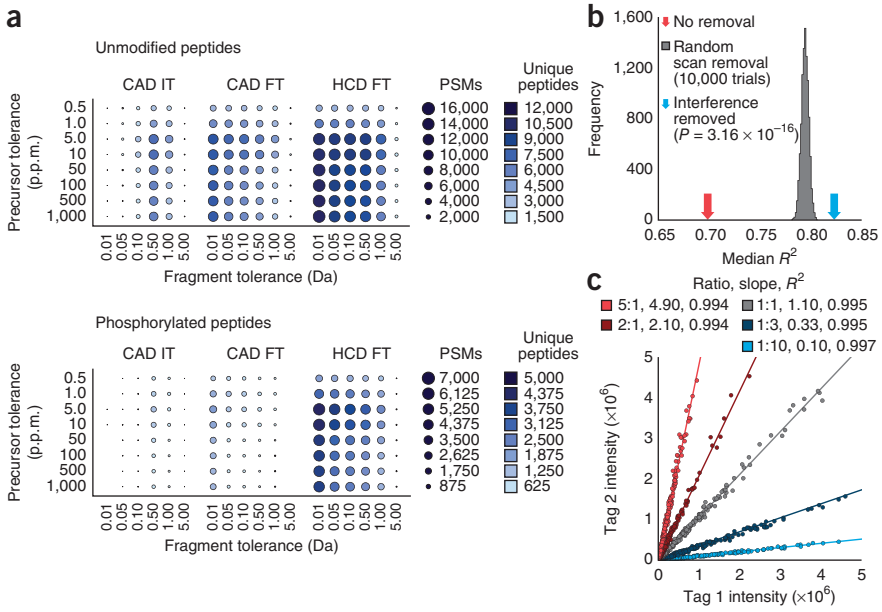

Figure 1 | Figures of merit for peptide identification and quantification. (**a**) Peptide identifications as a function of precursor and product mass tolerance. Using proteins isolated from human ESC whole-cell lysate, we performed liquid chromatography tandem mass spectrometry for each combination of dissociation method and mass analyzer. IT, ion-trap detection; FT, orbitrap detection. We searched data using fragment-ion tolerances of 0.01–5.0 Da, filtered results by precursor mass tolerances of 0.5–1,000 p.p.m. and filtered identifications to achieve 1% FDR. We performed experiments in triplicate and averaged the results. The number of peptide spectrum matches (PSMs) is proportional to circle size; number of unique peptides is represented by circle color as indicated. (**b**) R^2 values for all peptides in each protein (H1 versus NFF comparison; fourplex experiment) were calculated as a metric for quality of quantification. (**c**) Characterization of quantification. Data points represent reporter ion intensities for a single protein mixed in the indicated ratios. Lines represent the theoretical value for the mixtures presented.

identifies mass spectra compromised by interference and excludes these data points from peptide and protein quantification[19]. This filtering method resulted in an increase in quantitative precision. As random removal of spectra also increases R^2 values, we tested significance of the increase in R^2 value resulting from interference filtering by permutation testing. By fitting a Gaussian curve to the distribution, we estimated the significance of the increase in R^2 resulting from interference filtering ($P = 3.16 \times 10^{-16}$; **Fig. 1b**). TagQuant also incorporates mathematical correction of tag impurities, summing of reporter-ion intensities and exclusion of low-intensity reporter ions (Online Methods)[20,21]. We tested our complete workflow using a whole-cell lysate from *Saccharomyces cerevisiae*. Separate pools of protein were labeled with isobaric tags, combined in known ratios and analyzed via mass spectrometry. The observed results matched closely to the expected ratios for the range of mixtures tested ($R^2 > 0.99$; **Fig. 1c**).

Comparison of ESC and iPSC proteomes

We first compared transcripts, proteins and phosphorylation sites across two human ESC lines (H1 and H9), one iPSC line (DF19.7) and one fibroblast (newborn foreskin fibroblast (NFF)) cell line (**Supplementary Fig. 1**) using isobaric tags. With less than two weeks of analysis with an instrument, we identified 7,952 total proteins (1% false discovery rate (FDR); **Supplementary Table 1**) and 10,499 total sites of phosphorylation (localized with 95% confidence; **Supplementary Table 2** and **Fig. 2a**,**b**). We validated measurements for selected, representative proteins by western blots (**Supplementary Fig. 2**). Identified proteins include key regulators of pluripotency, such as OCT4 or POU5F1, NANOG and SOX2 (**Fig. 2c**), and nearly every major component of the developmentally related epigenetic regulators, polycomb group and trithorax proteins (**Supplementary Fig. 3**).

Comparing ESC and NFF lines revealed that 35% of proteins and 59% of phosphorylation sites differed by at least twofold in abundance. The genes corresponding to these differentially regulated proteins and phosphorylation sites were functionally related and representative of the two cell states. For example, proteins found in twofold higher amounts in ESCs were enriched for cell cycle–related processes (for example, DNA replication, cell division and others) (**Supplementary Table 3**), reflecting the rapid proliferation and shorter doubling times characteristic of pluripotent cells[22]. Conversely, proteins observed at higher levels in NFFs were enriched for processes pertinent to differentiated cell types. Differential regulation of phosphorylation sites was likewise apparent. Phosphorylation sites that were at least twofold higher in either ESCs or NFFs were enriched for several different amino acid motifs (**Supplementary Table 4**). To test whether this reflected differences in kinase activity between the two cell types, we mapped potential kinases to each phosphorylated site using group-based prediction system software[23]. We then used Fisher's exact test followed by Benjamini-Hochberg adjustment to determine whether substrates for particular kinases were enriched in sets of phosphorylation sites that were at least twofold different between ESCs and NFFs and mapped them to the human kinome tree (**Fig. 3**; adapted from ref. 24). Entire kinase families appear highly active in distinct cell types. For example, targets of CMGC kinases were more highly phosphorylated in the ESCs relative to NFFs, and substrates of CAMK and AGC kinases were more heavily occupied in the NFFs ($P < 0.05$, Fisher's exact test with Benjamini-Hochberg correction)[25]. The large number of differences and their functional enrichment confirm that two sample comparisons, without replicate analysis, are sufficient to characterize major differences between highly dissimilar cell types.

A complete map of the similarities and differences between ESCs and iPSCs will be key for both fundamental science and clinical applications. Single replicate comparison of one ESC line and one iPSC line, however, revealed twofold or greater differences in less than 1% of proteins and phosphorylation sites. This small set of proteins and phosphorylation sites showed no functional commonality (Gene Ontology terms[26], Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways[27] or phosphorylation motifs). Moreover, comparing ESCs and iPSCs yielded roughly the same number of absolute protein differences as a comparison between

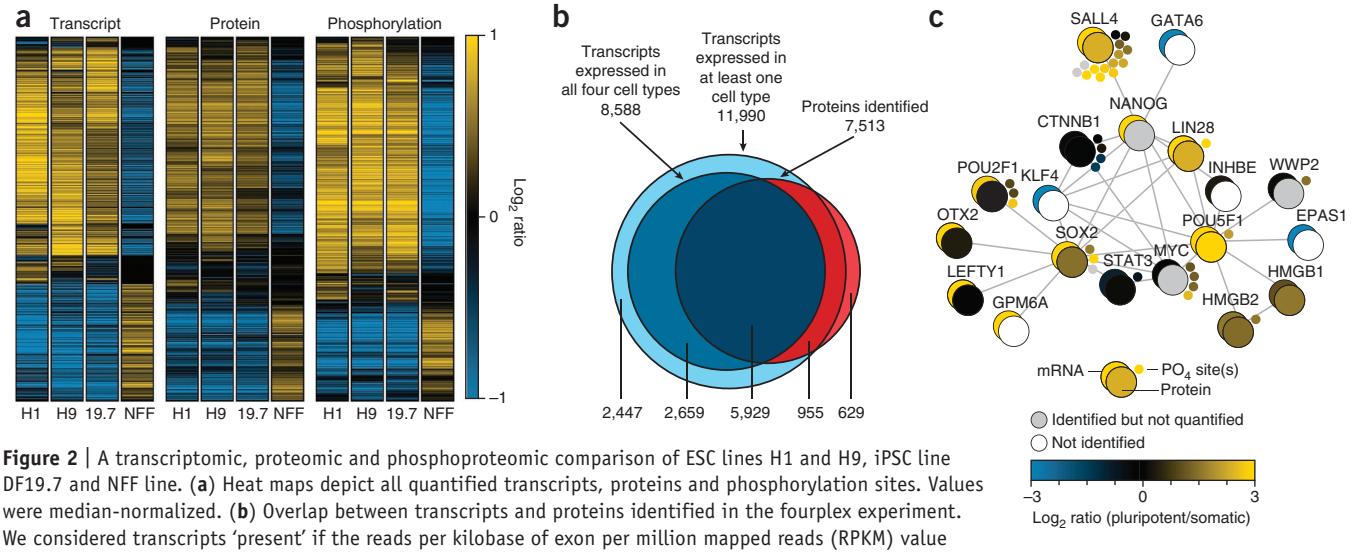

Figure 2 | A transcriptomic, proteomic and phosphoproteomic comparison of ESC lines H1 and H9, iPSC line DF19.7 and NFF line. (**a**) Heat maps depict all quantified transcripts, proteins and phosphorylation sites. Values were median-normalized. (**b**) Overlap between transcripts and proteins identified in the fourplex experiment. We considered transcripts 'present' if the reads per kilobase of exon per million mapped reads (RPKM) value was greater than 1 for all four cell types, and we determined protein identification via *P*-value filtering (1% FDR). (**c**) Cytoscape schematic of mRNA, protein and phosphorylation quantification from the fourplex experiment for genes known to have an interaction with NANOG, SOX2 or POU5F1 (search tool for the retrieval of interacting genes-proteins (STRING) database, confidence score > 0.90). Data are identified by protein name.

two ESC lines, regardless of fold difference (**Supplementary Fig. 4**). Together, these data suggested an overall inability to differentiate between ESCs and iPSCs at the protein level.

Next we examined RNA-seq data, which we acquired concomitantly with proteome data (**Supplementary Table 5**). Consistent with our proteomic results, we detected large differences between ESCs and NFFs: 48% of transcripts differed by twofold or more. Whereas 9% of transcripts differed by greater than twofold when comparing ESC (H9) and iPSC transcripts for a single replicate, the two ESC lines showed even greater variation (12% of transcripts). This suggested that it was not possible to distinguish differences between cell types from line-to-line variability. Unlike in the initial proteomic experiments, we carried out the RNA measurements in biological triplicate. Statistical analysis afforded by replicates enabled us to move beyond arbitrary fold cutoffs and establish statistical significance. Using Student's *t*-test with Benjamini-Hochberg correction ($P < 0.05$), we observed 623 differentially regulated transcripts between ESCs (H9) and iPSCs. From these data, we reasoned that proteomic differences likely existed between these similar cell types but were subtle and therefore masked by our inability to perform statistical analyses without replicates.

Replicate analyses

To test this hypothesis we leveraged the multiplexing capabilities of eight-plex isobaric tags to compare proteins and phosphorylation sites across four ESC lines (H1, H7, H9 and H14) and four iPSC lines (DF4.7, DF6.9, DF19.11 and DF19.7) in biological triplicate (**Supplementary Fig. 1**). To facilitate comparison between all 24 samples, we median-normalized reporter-ion intensities. Proteomic and phosphoproteomic analyses took less than six weeks to acquire and resulted in the identification of 6,761 total proteins (<1% FDR; **Supplementary Table 1**) and 19,122 total sites of phosphorylation (localized with at least 95% confidence; **Supplementary Table 2**). We quantified 4,742, 3,396 and 2,234 proteins in at least one, two or three replicates, respectively, and 14,162, 8,217 and 4,564 localized phosphorylation sites in at least one, two or three replicates, respectively. Again we performed an mRNA-seq analysis for each of the samples using an Illumina Genome Analyzer IIx.

Analysis of a single biological replicate (8 cell lines) revealed only 1 transcript, 5 proteins and 4 phosphorylation sites that were statistically different ($P < 0.05$, Student's *t*-test with Benjamini-Hochberg correction; **Fig. 4a**). However, inclusion of two more biological replicates permitted detection of many differentially regulated elements: 1,560 transcripts, 293 proteins and 292 phosphoisoforms differed significantly between ESCs and iPSCs ($P < 0.05$, Student's *t*-test with Benjamini-Hochberg correction; **Fig. 4a–c** and **Supplementary Table 6**). Greater than 90% of the differentially regulated transcripts, proteins and phosphorylation sites differed by less than twofold. These minor deviations were only detectable through biological replicate analysis, which increased sample size, and with it, statistical power.

Though biological replicates provide the statistical power to detect differences, they may not always distinguish pervasive differences between cell types from variance between cell lines. This is best illustrated by considering just H1 ESC lines and DF4.7 iPSC lines. Biological triplicate analysis of transcripts from these lines indicates 990 differentially regulated transcripts ($P < 0.05$, Student's *t*-test with Benjamini-Hochberg correction; **Fig. 4d**). However, most (63%) of these differences did not overlap with differentially regulated transcripts as determined by the full 24-sample comparison. Moreover, we did not detect 72% of the differences identified by analysis of all eight cell lines in biological triplicate by comparison between H1 and DF4.7 cells alone (**Fig. 4d**). We conclude that analyzing multiple cell lines is an essential addition to biological replicates for comparing ESCs and iPSCs.

Despite the subtlety of the differences observed here, their functional enrichment suggests a consistent distinction in regulation between ESCs and iPSCs. Transcripts, proteins and phosphorylation sites found at higher levels in iPSCs were enriched for many biological processes required for somatic cell function, including system process, organ development, blood

circulation and muscle-system process (**Supplementary Table 7**). However, motif analysis of differentially regulated phosphorylation sites did not implicate any specific kinases or phosphatases in these differences. Despite the functional relationship of the differentially regulated elements, the differences at each level of regulation (transcript, protein and phosphorylation) often did not correspond to the same genes (**Fig. 4e**).

To determine whether differences between ESCs and iPSCs represented incomplete reprogramming of iPSCs, we contrasted ESC and iPSC comparisons with ESC and NFF comparisons.

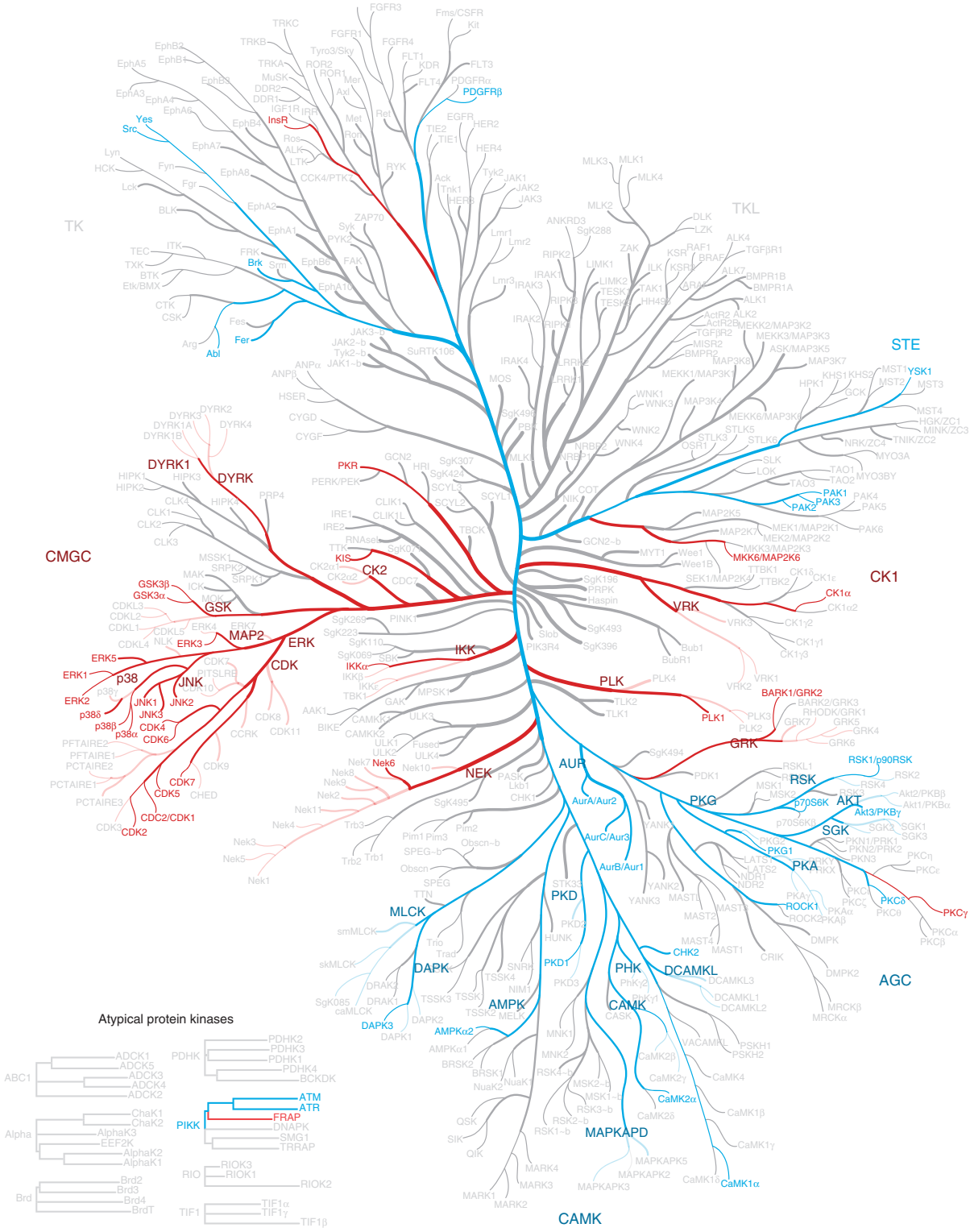

Figure 3 | Kinase substrate analysis between ESCs and NFFs (adapted from ref. 24 with permission from the American Association for the Advancement of Science). Highlighted are kinase substrates for sets of phosphorylation sites that were enriched (changed by more than twofold) in ESCs (red; $P < 0.05$) and in NFFs (blue; $P < 0.05$).

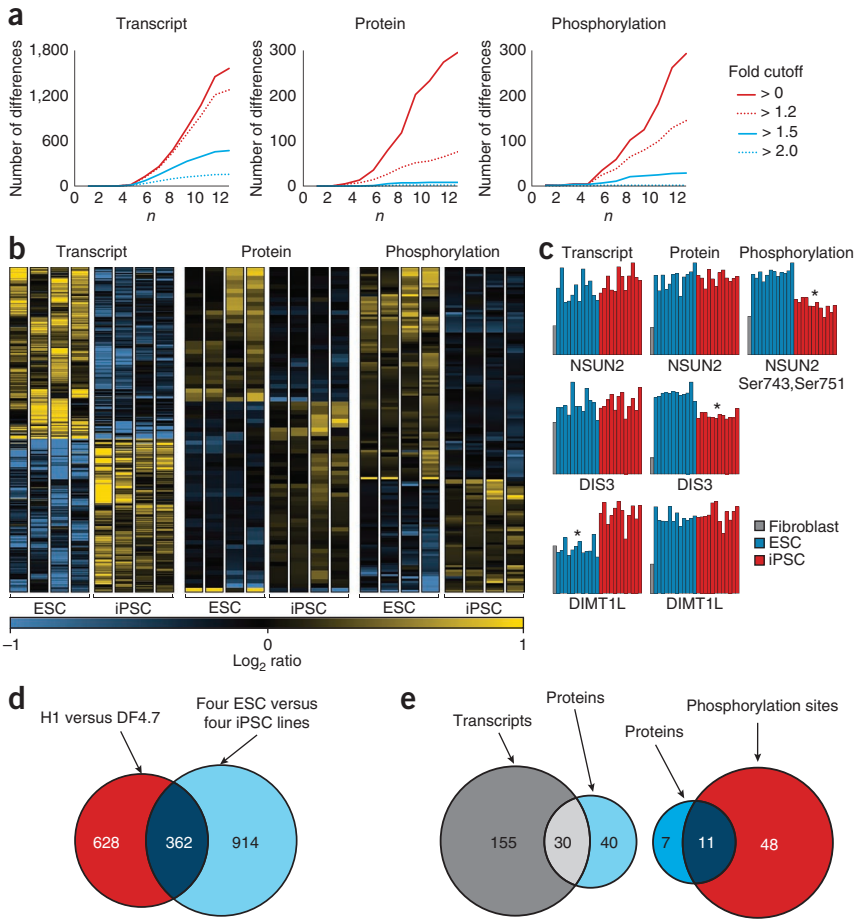

Figure 4 | Comparison of four ESC and four iPSC lines. (**a**) Differentially regulated transcripts, proteins and phosphorylation sites are shown as a function of the number of comparisons (*n*). We performed differential expression analysis using subsets of data. For example, the $n = 2$ value reflects the number of differences detected from comparing just two ESC lines and two iPSC lines without biological replicate, whereas $n = 12$ represents the differences detected from comparing all four ESC lines and all four iPSC lines in biological triplicate. The number of differentially regulated elements for a given fold difference is indicated by different colors. The lines connect data point for ease of interpretation. (**b**) Heatmaps depicting differentially regulated transcripts, proteins and phosphorylation sites ($P < 0.05$, Student's *t*-test, with Benjamini-Hochberg correction). Only transcripts exhibiting at least a 1.5-fold difference and protein and phosphorylation sites exhibiting at least a 1.2-fold difference are shown. (**c**) Randomly selected examples of differentially regulated transcripts, proteins and phosphorylation sites. Bar heights represent relative reporter ion intensity (arbitrary units). $^*P < 0.05$ (Student's *t*-test), (ESCs compared to iPSCs). (**d**) Differentially regulated transcripts detected based on either a comparison between biological triplicates of H1 and DF4.7 cell lines or a comparison of biological triplicates of all four ESC and all four iPSC lines. (**e**) Overlap between differentially regulated proteins and transcripts (left; only genes with both a quantified protein and transcript were included) and differentially regulated proteins and phosphorylation sites (right; only genes with both a quantified protein and phosphorylation site were included).

Based on gene-enrichment analysis, three biological processes showed enrichment at transcript, protein and phosphorylation level in iPSCs compared to ESCs: muscle system process, muscle contraction and wound healing. These terms reflect cellular function characteristic of mesodermal lineages and may represent the NFF origin of the iPSCs. Supporting this hypothesis, we found that all three terms were enriched in the transcripts, proteins and phosphorylation sites that were at least twofold higher in NFFs than ESCs (**Supplementary Table 8**). In fact, more than half of the Gene Ontology terms enriched among transcripts, proteins and phosphorylation sites that were significantly higher in iPSCs compared to ESCs were also enriched in NFF compared to ESCs. Among this dataset were multiple phosphorylation events on NSUN2, encoded by a proto-oncogene implicated in cell proliferation[28] (**Fig. 4c**). *NSUN2* transcript and total protein levels for NSUN2 were not different between ESCs and iPSCs, suggesting the changes are not simply a matter of protein abundance. Additionally, phosphorylation of these sites in iPSCs was similar to the levels observed in fibroblast cells, which may reflect residual regulation from kinases and phosphatases more characteristic of the differentiated NFFs. NSUN2 acts downstream of c-MYC[28], one of a handful of factors commonly used to improve reprogramming efficiency. At the transcript level, the set of mRNAs more abundant in iPSCs, which included *TBX15* and *PITX2*, were enriched for developmental function and exposed a connection to mesoderm differentiation[29,30]. All of these results suggest that somatic cell programs are not completely silenced during reprogramming. Although this has been observed before in gene expression studies[31], to our knowledge this is the first evidence that incomplete silencing is also reflected in regulation of proteins and post-translational modifications[10,32].

Data resource and sharing

To facilitate integration of these results with other datasets we created the Stem Cell–Omics Repository (SCOR; http://scor.chem.wisc.edu/) a web-based resource that collates quantitative biological analyses of ESCs and iPSCs. A key feature of SCOR is the ability to visualize quantitative information for transcripts, proteins and post-translational modifications from many sources (**Supplementary Fig. 5**). Included in the database are several large-scale analyses from other laboratories, all of which are queried during standard searches. To ensure that SCOR remains relevant, we added an option for users to submit published data for inclusion on the website. Our intention is that the resource will expand as the field grows. A separate tab in the tools section provides open-access, downloadable programs used for postacquisition data processing, including the interference filtering program, TagQuant. All datasets are downloadable from the SCOR database.

To demonstrate the value of this resource, we applied SCOR to evaluate results from this and several other microarray and

RNA-seq experiments[1,4]. This analysis, encompassing iPSCs derived using integrating viral vectors and non-integrating episomal vectors, identified several transcripts that were consistently different in ESCs versus iPSCs (**Supplementary Table 9**). To include data from outside laboratories we intersected our results with a similar dataset[4] (**Supplementary Table 9**). This dataset contained two transcripts (*TCERG1L* and *FAM19A5*) that were consistently higher in ESCs relative to iPSCs. Recent work reported that both of these genes exhibit promoter hypermethylation and ultimately lower expression in several iPSC lines[10]. These and other genes that show consistent differential regulation are of great interest for further studies. As more proteomic studies of ESCs and iPSCs become available, we anticipate that SCOR will facilitate similar interlaboratory comparisons to determine the most pervasive transcriptomic, proteomic and phosphoproteomic discrepancies.

DISCUSSION

This comparison offers important insights into the nature of reprogrammed cells. One subtle but critical conclusion is the remarkable similarity between ESCs and iPSCs, which is highlighted by the technical rigor required in our study to detect even minor differences. Although the exact biological relevance of these differences remains unknown, functional similarity of the genes that contribute to them suggest that iPSCs retain residual regulation characteristic of the cells from which they were derived. These differences do not appear to appreciably alter cellular function in the pluripotent state but instead may surface during differentiation as cells invoke gene expression programs needed for development. Although iPSCs can produce mesoderm, endoderm and ectoderm, the process of reprogramming selects for cells predisposed to the pluripotent state, not necessarily for cells that differentiate with equal efficiency to all lineages. For example, recent studies have reported that ESC lines differentiate into neural lineages with higher efficiency than most iPSC lines[33]. From our data, many transcripts with lower expression in iPSCs relative to ESCs, such as *NNAT* (neuronatin) and *SOX11*, were also functionally related through their role in neural development[34,35]. At the post-translational modification level, the extent of phosphorylation was consistently lower in iPSCs for several microtubule-related proteins that are directly (DPYSL2; ref. 36) or indirectly (FAM29A[37]) implicated in neural differentiation and development. Understanding how these genes contribute to neural differentiation in both ESCs and iPSCs will be the subject of additional study.

A major advantage of combining multiple planes of measurement is the ability to dissect regulatory mechanisms not apparent in a single dimension. For instance, many of the protein kinases whose substrates exhibited significant differences ($P < 0.05$, Fisher's exact test) in phosphorylation exhibited little to no change at the transcript or protein level. For example, while *CDK2* mRNA and CDK2 protein amounts were largely unchanged (less than twofold) in pluripotent cells relative to NFFs, CDK2 substrates were more highly phosphorylated in pluripotent cells. A possible explanation for this observation was apparent in our global post-translational modification data. Phosphorylation of CDK2 at Thr160, a mark required for kinase activity[38], was upregulated by nearly sixfold in all three pluripotent cell lines. Likewise, CDK4, CDK5 and CDK6 all have similar amounts of transcript and protein in the pluripotent cells, but the motifs they target show a significant ($P < 0.05$, Fisher's exact test) increase in phosphorylation. In contrast, the higher transcript and corresponding protein expression of cAMP-dependent protein kinase and protein kinase C in NFFs may explain the corresponding high levels of substrate phosphorylation. Taken together, these data suggest multiple mechanisms for the regulation of kinases. For instance, proteins involved in transitory functions, such as the aforementioned cell cycle–related kinases, may be regulated via rapid and dynamic signals (phosphorylation and dephosphorylation) rather than by slower and longer-lasting transcriptional and translational changes.

The results presented here highlight the importance of including multiple biological replicates to overcome biological and technical variability and to establish statistical significance. Moreover, evaluating multiple cell lines or subjects ensures that observed differences are persistent and not merely single sample aberrations. In this study we grew and collected cells simultaneously under identical conditions to minimize variation introduced by sample handling. In interpreting these results, we recognize the importance of expanding the comparison of ESCs and iPSCs to cover as many lines, reprogramming methods and growth conditions as possible. To date, 75 ESC lines are listed on the US National Institutes of Health–approved registry and innumerable iPSC lines are available from diverse sources. Comparing all of these cell lines is a daunting task for a single research group. We therefore created SCOR, an open-access resource to collate, visualize and analyze large-scale datasets related to pluripotency. As research expands, we hope that the SCOR website will bring datasets together and facilitate cross-laboratory comparisons at every tier of regulation.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We acknowledge A.J. Bureta for helping with and providing illustrations, A. Williams and K. Eastman for editing the text, G.McAlister for assistance with instrumentation, J. Yu for providing transcriptomic data before publication and Thomson lab members for critical reading and discussion of this manuscript. This work was supported by the University of Wisconsin, the Beckman Foundation and US National Institutes of Health (NIH) grants R01GM080148 (to J.J.C.) and P01GM081629 (to J.A.T. and J.J.C.). D.H.P. acknowledges support from a NIH predoctoral traineeship, the Genomic Sciences Training Program, NIH grant 5T32HG002760.

AUTHOR CONTRIBUTIONS

D.H.P. designed research, prepared samples, performed mass spectrometry, wrote software, analyzed data and wrote the manuscript. J.B. designed research, grew cells, prepared samples, analyzed data and wrote the manuscript. C.D.W. wrote software. S.T. and V.R. analyzed data. M.D.P. grew cells. D.J.B. designed websites. D.L.S. helped with phosphorylation analysis. M.A.T. optimized the labeling procedure. J.M.B. performed RNA sequencing. R.S. designed research and analyzed data. J.A.T. and J.J.C. designed research and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturemethods/.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Yu, J. *et al.* Human induced pluripotent stem cells free of vector and transgene sequences. *Science* **324**, 797–801 (2009).
2. Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861–872 (2007).
3. Yu, J. *et al.* Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318**, 1917–1920 (2007).
4. Chin, M.H. *et al.* Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures. *Cell Stem Cell* **5**, 111–123 (2009).
5. Guenther, M.G. *et al.* Chromatin structure and gene expression programs of human embryonic and induced pluripotent stem cells. *Cell Stem Cell* **7**, 249–257 (2010).
6. Chin, M.H., Pellegrini, M., Plath, K. & Lowry, W.E. Molecular analyses of human induced pluripotent stem cells and embryonic stem cells. *Cell Stem Cell* **7**, 263–269 (2010).
7. Bock, C. *et al.* Reference maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines. *Cell* **144**, 439–452 (2011).
8. Stadtfeld, M. *et al.* Aberrant silencing of imprinted genes on chromosome 12qF1 in mouse induced pluripotent stem cells. *Nature* **465**, 175–181 (2010).
9. Doi, A. *et al.* Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nat. Genet.* **41**, 1350–1353 (2009).
10. Lister, R. *et al.* Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* **471**, 68–73 (2011).
11. Olsen, J.V. *et al.* Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* **4**, 709–712 (2007).
12. McAlister, G.C., Phanstiel, D., Wenger, C.D., Lee, M.V. & Coon, J.J. Analysis of tandem mass spectra by FTMS for improved large-scale proteomics with superior protein quantification. *Anal. Chem.* **82**, 316–322 (2010).
13. Olsen, J.V. *et al.* A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. *Mol. Cell. Proteomics* **8**, 2759–2769 (2009).
14. Nagaraj, N., D'Souza, R.C.J., Cox, J., Olsen, J.V. & Mann, M. Feasibility of large-scale phosphoproteomics with higher energy collisional dissociation fragmentation. *J. Proteome Res.* **9**, 6786–6794 (2010).
15. Thompson, A. *et al.* Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* **75**, 1895–1904 (2003).
16. Ross, P.L. *et al.* Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* **3**, 1154–1169 (2004).
17. Choe, L. *et al.* 8-plex quantitation of changes in cerebrospinal fluid protein expression in subjects undergoing intravenous immunoglobulin treatment for Alzheimer's disease. *Proteomics* **7**, 3651–3660 (2007).
18. Ow, S.Y. *et al.* iTRAQ underestimation in simple and complex mixtures: the good, the bad and the ugly. *J. Proteome Res.* **8**, 5347–5355 (2009).
19. Wenger, C.D., Phanstiel, D.H., Lee, M.V., Bailey, D.J. & Coon, J.J. COMPASS: a suite of pre- and post-search proteomics software tools for OMSSA. *Proteomics* **6**, 1064–1074 (2011).
20. Shadforth, I.P., Dunkley, T.P.J., Lilley, K.S. & Bessant, C. i-Tracker: for quantitative proteomics using iTRAQ (TM). *BMC Genomics* **6**, 145 (2005).
21. Griffin, T.J. *et al.* iTRAQ reagent-based quantitative proteomic analysis on a linear ion trap mass spectrometer. *J. Proteome Res.* **6**, 4200–4209 (2007).
22. Becker, K.A., Stein, J.L., Lian, J.B., van Wijnen, A.J. & Stein, G.S. Establishment of histone gene regulation and cell cycle checkpoint control in human embryonic stem cells. *J. Cell. Physiol.* **210**, 517–526 (2007).
23. Xue, Y. *et al.* GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. *Mol. Cell. Proteomics* **7**, 1598–1608 (2008).
24. Manning, G., Whyte, D.B., Martinez, R., Hunter, T. & Sudarsanam, S. The protein kinase complement of the human genome. *Science* **298**, 1912–1934 (2002).
25. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J. R. Stat. Soc., B* **57**, 289–300 (1995).
26. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
27. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
28. Frye, M. & Watt, F.M. The RNA methyltransferase Misu (NSun2) mediates Myc-induced proliferation and is upregulated in tumors. *Curr. Biol.* **16**, 971–981 (2006).
29. Singh, M.K. *et al.* The T-box transcription factor Tbx15 is required for skeletal development. *Mech. Dev.* **122**, 131–144 (2005).
30. Dong, F. *et al.* Pitx2 promotes development of splanchnic mesoderm-derived branchiomeric muscle. *Development* **133**, 4891–4899 (2006).
31. Kim, K. *et al.* Epigenetic memory in induced pluripotent stem cells. *Nature* **467**, 285–290 (2010).
32. Polo, J.M. *et al.* Cell type of origin influences the molecular and functional properties of mouse induced pluripotent stem cells. *Nat. Biotechnol.* **28**, 848–855 (2010).
33. Hu, B.Y. *et al.* Neural differentiation of human induced pluripotent stem cells follows developmental principles but with variable potency. *Proc. Natl. Acad. Sci. USA* **107**, 4335–4340 (2010).
34. Siu, I.M. *et al.* Coexpression of neuronatin splice forms promotes medulloblastoma growth. *Neuro-oncol.* **10**, 716–724 (2008).
35. Hargrave, M. *et al.* Expression of the Sox11 gene in mouse embryos suggests roles in neuronal maturation and epithelio-mesenchymal induction. *Dev. Dyn.* **210**, 79–86 (1997).
36. Kawano, Y. *et al.* CRMP-2 is involved in kinesin-1-dependent transport of the Sra-1/WAVE1 complex and axon formation. *Mol. Cell. Biol.* **25**, 9920–9935 (2005).
37. Zhu, H., Coppinger, J.A., Jang, C.Y., Yates, J.R. III & Fang, G. FAM29A promotes microtubule amplification via recruitment of the NEDD1-gamma-tubulin complex to the mitotic spindle. *J. Cell Biol.* **183**, 835–848 (2008).
38. Bourke, E., Brown, J.A.L., Takeda, S., Hochegger, H. & Morrison, C.G. DNA damage induces Chk1-dependent threonine-160 phosphorylation and activation of Cdk2. *Oncogene* **29**, 616–624 (2010).

ONLINE METHODS

Cell growth and lysis. We maintained human embryonic stem cells (lines H1, H7, H9 and H14) and induced pluripotent cells (lines DF4.7, DF6.9, DF19.7 and DF19.11) in a feeder-independent system, as previously described[39]. We karyotyped all ESC and iPSC lines before experiments using standard G-banding chromosome analysis (WiCell Research Institute). When cells reached 70% confluency, we passaged cells enzymatically using dispase (Invitrogen) at a 1:4 splitting ratio. We cultured human NFFs (CRL-2097; American Type Culture Collection (ATCC)) essentially according to ATCC recommendations. We maintained cells in 10% (v/v) FBS (Hyclone Laboratories), 1 mM L-glutamine (Invitrogen), 0.1 mM β-mercaptoethanol (Sigma-Aldrich) and 0.1 mM nonessential amino acids in DMEM (both from Invitrogen). We passaged cells at roughly 70% confluency at a 1:3 splitting ratio, using Tryp-LE (Invitrogen).

For proteomics experiments, we collected all cells by incubation for 10 min with an adequate volume of prewarmed (37 °C), 0.05% Tryp-LE to cover the culture surface. After cell detachment, we added an equivalent volume of either ice-cold growth medium for NFFs, or ice-cold DPBS (Invitrogen) for ESCs, before collecting the cells. We subsequently washed cell pellets twice in ice-cold DPBS and stored them at −80 °C. We collected ~10^8 cells for each analysis. We lysed samples via sonication in lysis buffer containing 8 M urea, 40 mM NaCl, 50 mM Tris (pH 8), 2 mM $MgCl_2$, 50 mM NaF, 50 mM b-glyceraldehyde phosphate, 1 mM sodium orthovanadate, 10 mM sodium pyrophosphate, 1× mini EDTA-free protease inhibitor (Roche Diagnostics) and 1× phosSTOP phosphatase inhibitor (Roche Diagnostics).

For RNA-seq analysis, we washed celled twice in prewarmed (37 °C) DPBS and lysed them on the culture dish using Trizol reagent (Invitrogen). We added chloroform (Sigma) to a final concentration of 16.7% (v/v) and centrifuged the sample for 15 min at 12,000*g* at 4 °C. We combined the resulting supernatant with an equal volume of 70% ethanol and processed it using the Qiagen RNeasy kit with on-column DNAse digestion. We linearly amplified poly(A)$^+$ RNAs using a modified T7 amplification method[40] that retains directionality of the transcripts. This protocol generates Illumina RNA-seq libraries with uniform coverage of the entire length of mRNAs. Samples were run on an Illumina Genome Analyzer IIx. We then aligned each lane to the genome and the exon splice sites database using bowtie[41], allowing up to ten multiple matches and three mismatches. For data processing, we filtered 42-base-pair reads to remove adapters in each lane. We used Enhanced Read Analysis of Gene Expression (ERANGE)[42] to obtain expression values in reads per kilobase of exon model per million mapped reads (RPKM).

mRNA analysis. We performed microarray raw data processing and normalization as previously described[1,3]. We assessed ESC and iPSC specificity of transcripts as follows. First, we fit a linear model to estimate all the fold changes across the iPSC and ESC lines, and then applied Bayesian smoothing to the standard errors among the same type of cell lines. Finally, we calculated a *P* value based on the moderated *t*-statistics for the differentially expressed genes and then adjusted them based on Benjamini and Hochberg's method to control the FDR[25]. Second, we required the fold change to be at least threefold different between the two cell types, with an adjusted $P \leq 0.05$. The data in **Supplementary Table 9** were generated from 15 microarrays for ESCs, 25 microarrays for iPSCs, and three microarrays for differentiated cell types (NFF and IMR90) pooled from the work reported in refs. 1,3.

Western blot analysis. To confirm quantification determined by mass spectrometry, we analyzed several proteins by western blot analysis (**Supplementary Fig. 2**). After cell lysis, we loaded equal amounts of total protein from H1 cells, H9 cells, iPSC and NFFs onto a 4–15% acrylamide gel (Biorad). We used the following primary antibodies to detect the indicated protein: mouse monoclonal antibody to human OCT4 (1:2,000, sc-5279, Santa Cruz Biotechnology), goat antibody to human DNMT3B (1:1,000, sc-10235, Santa Cruz Biotechnology), mouse antibody to human GAPDH (1:2,000, MAB374, Chemicon), and mouse antibody to human CD44 (1:10, 550989, Pharmingen-BD). We used the following horseradish peroxidase–linked secondary antibodies: goat antibody to mouse IgG (1:2,000, sc-2005, Santa Cruz Biotechnology), donkey antibody to goat IgG (1:2,000, sc-2056, Santa Cruz Biotechnology). We loaded a biotin-labeled ladder according to the manufacturer's specification (Cell Signaling). We used a Super Signal West Pico Chemiluminescent Substrate (Thermo Scientific Pierce) according to the manufacturer's protocol to image blots on a LAS-3000 Imaging System (Fujifilm Life Science). We determined quantification according to manufacturer's instructions with MultiGauge software, ver 2.0 (Fujifilm Life Science). Between detections, we stripped the membrane using Restore Western Blot Stripping buffer (Thermo Scientific Pierce).

Digestion and labeling. We reduced cysteine residues in the proteomics-experiment samples with 5 mM dithiothreitol, alkylated them using 10 mM iodoacetamide and digested proteins in a two-step process. We added proteinase Lys-C (Wako Chemicals) (enzyme:protein ratio = 1:100) and incubated the samples for ~2 h at 37 °C in lysis buffer. We then diluted samples with 50 mM Tris (pH 8) until the urea concentration was 1.5 M and digested them with trypsin (Promega) (enzyme:protein ratio = 1:50) at 37 °C overnight. We quenched reactions using trifluoroacetic acid (TFA). We dried samples to completion after purification using C18 solid-phase extraction columns (SepPak, Waters). We performed 'isobaric tags for relative and absolute quantitation' (iTRAQ) labeling according to manufacturer supplied protocols (Applied Biosystems)[16,17]. To ensure that each of the samples contained the same amount of protein we prepared a small 1:1:1:1 (1:1:1:1:1:1:1:1 for eightplex experiment) aliquot and analyzed it by mass spectrometry. We used summed reporter ion ratios from this experiment to inform mixing ratios of the remaining labeled digests. Once mixed, we dried samples to completion and purified by them by solid-phase extraction.

Fractionation. We resuspended the labeled peptides in strong cation exchange buffer A (5 mM KH_2PO_4 and 30% acetonitrile; pH 2.65) and injected them onto a polysulfoethylaspartamide column (9.4 mm × 200 mm; PolyLC). Buffer B comprised 5 mM KH_2PO_4, 30% acetonitrile and 350 mM KCl (pH 2.65), and buffer C comprised 50 mM KH_2PO_4 and 500 mM KCl (pH 7.5). We performed separations using a Surveyor liquid chromatography quaternary pump (Thermo Scientific) at a flow rate of 3.0 ml min^{-1}. We used the following gradient for separation: 0–2 min, 100% buffer A, 2–5 min, 0–15% buffer B, 5–35 min, 15–100% buffer

doi:10.1038/nmeth.1699

B. Buffer B was held at 100% for 10 min. Finally, the column was washed extensively with buffer C and water before recalibration. We collected the samples by hand and desalted them by solid-phase extraction.

Phosphopeptide enrichment. After strong cation exchange fractionation, we enriched phosphopeptides using magnetic beads (Qiagen). We washed the beads three times with water, three times with 40 mM EDTA (pH 8.0) for 30 min with shaking and three times with water again. We then incubated beads with 100 mM $FeCl_3$ for 30 min with shaking. Finally, we resuspended beads in 1 ml 1:1:1 acetonitrile:methanol:0.01% acetic acid and washed them three times with 80% acetonitrile and 0.1% TFA in water. We resuspended samples in 80% acetonitrile and 0.1% TFA, and incubated them with beads for 30 min with shaking. We washed the beads six times with 200 µl 80% acetonitrile and 0.1% TFA, and eluted the peptides using 1:1 acetonitrile:5% NH_4OH in water. We acidified eluted phosphopeptides immediately with 4% formic acid, lyophilized them to ~10 µl, and diluted them with 50 mM phosphate buffer before analysis.

Mass spectrometry. We performed tandem mass spectrometry using a NanoAcquity ultra high-pressure liquid chromatography system (Waters) coupled to a dcQLT-orbitrap (Thermo Fisher Scientific). Samples were loaded onto a precolumn (75 µm inner diameter, packed with 5-cm C18 particles, Alltech) for 10 min at a flow rate of 1 µm min^{-1}. Samples were then eluted over an analytical column (50 µm inner diameter, packed with 15 cm C18 particles, Alltech) using a 120-min linear gradient from 1% to 35% acetonitrile with 0.2% formic acid and a flow rate of 300 nl min^{-1}. An additional 30 min were used for column washing and equilibration. We constructed columns as previously described[12].

All mass spectrometer instrument methods consisted of one scan (survey (MS1) scan) (resolving power = 30,000–60,000) followed by data dependent tandem mass spectrometry MS2 scans (resolving power = 7,500) of the ten most intense precursors. Protein identification experiments used exclusively beam-type CAD (HCD) with orbitrap mass analysis. Some phosphopeptide identification experiments included alternating HCD and electron transfer dissociation (ETD) MS2 scans. We quantified any peptides identified by ETD using the corresponding HCD scan. We used an exclusion list for 60 s using a window of −0.55 Th to 2.55 Th. We excluded precursors with unassigned charges states or charge states of one (and two for ETD scans). We used automatic gain control target values of 1,000,000 for MS1 analysis and 50,000 for orbitrap MS2 analysis. To maximize quantified identifications we used QuantMode for some analyses.

Database search and FDR filtering. We used DTA generator to extract peak information from .Raw files and print it into a searchable text file[43]. This software removed fragment ions related to the iTRAQ reagents and as well as charged reduced precursors. We searched spectra against the International Protein Index (IPI) human database version 3.75 with full enzyme specificity using The Open Mass Spectrometry Search Algorithm (OMSSA; version 2.1.4)[44,45]. We used a mass tolerance of ± 4.5 Da precursors and a monoisotopic mass tolerance of ± 0.01 Da for fragments ions. We set carbamidomethylation of cysteines, iTRAQ fourplex on the N terminus, and iTRAQ (fourplex or eightplex) on lysines as fixed modifications, and oxidation of methionines and iTRAQ (fourplex or eightplex) on tyrosines as variable modifications. For phosphopeptide searches we included variable phosphorylation of serine, threonine, and tyrosine as variable modifications. We used the 'Coon OMSSA Proteomic Analysis Software Suite (COMPASS) software suite to filter peptides to a 1% FDR. COMPASS groups peptides into proteins following the rules previously established[46]. COMPASS multiplies probability values for unique peptides of each protein to obtain protein probability values and then filters proteins by this score to achieve a 1% FDR at the protein level.

Peptide and protein quantification. We used custom software, TagQuant[19], to perform iTRAQ quantification. TagQuant is written in C# programming language and distributed along with Compass software suite. TagQuant extracts reporter ion intensities and multiplies them by injection times to determine counts. TagQuant performs purity correction as previously described[20]. TagQuant normalizes intensities such that the total signal from each channel is equal. We summed reporter ion intensities for each channel for all peptides in a given protein with three exceptions: (i) scans corresponding to peptides found in multiple protein groups were not used for quantification, (ii) peptides found to be phosphorylated were not used for protein quantification and (iii) if peaks not related to the precursor were present in the MS1 scan within ± 1.8 Th of the selected precursor at an intensity greater than 25% of the selected precursor the resulting MS2 scan was not used for quantification. We median normalized protein and phosphorylation site quantification in order to compare across all three replicate experiments.

Phosphorylation analysis. We filtered phosphopeptides to a 1% FDR based on unique peptides as described above. To avoid over-reporting of phosphorylation sites, we combined phosphorylated peptides and nonphosphorylated peptides and grouped them into proteins together, following previously established rules[46].

We used the phosphinator software program to localize phosphorylation sites[47]. The algorithm calculates theoretical fragment ion m/z ratios for all possible permutations of phosphopeptide isoforms given the sequence and number of phosphorylations. The algorithm then compares the experimental spectrum against the theoretical product ions for each candidate phosphopeptide isoform, using a product mass tolerance of ± 0.02 Th. Two criteria are required for localization. First, the candidate with the highest number of matching product ions must have at least one more matching product ion than the second highest. Second, the algorithm performs a statistical test to determine the significance of the observed product ions supporting phosphorylation at a specific residue. We take the null hypothesis to be that there is no evidence that a given phosphorylation is localized and that any site-determining fragments observed are merely spurious matches. We calculated a probability value (p value) that represents the likelihood of obtaining the observed number of site-determining fragments or more based on random chance, using the following equation, the cumulative distribution function for a binomial distribution:

$$P(n) = \sum_{k=n}^{N} \binom{N}{k} p^k (1-p)^{N-k}$$

where P is the p-value, N is the number of possible site-determining fragment ions, n is the number of observed site-determining

fragment ions, and p is the probability of a single spurious fragment ion match. The algorithm calculates p as the product of the number of observed tandem mass spectrometry (MS/MS) peaks and the twice the product mass tolerance (±), divided by the MS/MS m/z range.

The algorithm performs this significance test twice for every phosphorylation site in the top isoform: once on each side of the phosphorylated residue. The site-determining fragment ions are those between the phosphorylation site and the closest amino acid residue that could be phosphorylated but are not in the top isoform. The algorithm considers doubly charged products for +3 m/z and higher precursors when the product is comprised of a sufficient portion of the peptide. Phosphinator converts the p-value to a human-readable score by taking $-10 \log_{10}(P)$. We only consider sites where this score is above 13 (that is, $P < 0.05$) on both the left and right side of the residue to be localized, and we only use peptides with all phosphorylations localized for quantitative analysis.

Next, we counted phosphorylation sites. We summed quantitative information from all phosphopeptides that contained the same sites to get the most accurate quantification for each site or combination of sites. We grouped peptides containing multiple sites with other peptides containing the exact same combination of sites. Therefore, we presented a list of phosphorylation isoforms rather than a list of phosphorylated sites. Phosphorylation isoforms can have information regarding one site or a combination of multiple sites. We only counted redundant sites that were found in more than one isoform once in the final count of phosphorylation sites.

Enrichment analysis. We performed two-tailed Student's t-test assuming equal variance in Microsoft Excel. To correct for multiple-hypothesis testing, we applied Benjamini-Hochberg adjustment using the R statistics package. We used a local gene ontology MySQL database installation for analysis of function and cellular location and another local MySQL database populated with information from the Kyoto Encyclopedia of Genes and Genomes Application Programming Interface. We determined putative kinase targets using the group-based prediction system software. To perform Fisher's exact test and subsequent Benjamini-Hochberg correction, we wrote custom software in the C++ programming language and interfaced to the R statistics package through the R Component Object Model library.

39. Ludwig, T.E. *et al.* Derivation of human embryonic stem cells in defined conditions. *Nat. Biotechnol.* **24**, 185–187 (2006).
40. Sengupta, S. *et al.* Highly consistent, fully representative mRNA-Seq libraries from ten nanograms of total RNA. *Biotechniques* **49**, 898–904 (2010).
41. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10** R25 (2009).
42. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
43. Good, D.M. *et al.* Post-acquisition ETD spectral processing for increased peptide identifications. *J. Am. Soc. Mass Spectrom.* **20**, 1435–1440 (2009).
44. Geer, L.Y. *et al.* Open mass spectrometry search algorithm. *J. Proteome Res.* **3**, 958–964 (2004).
45. Kersey, P.J. *et al.* The International Protein Index: An integrated database for proteomics experiments. *Proteomics* **4**, 1985–1988 (2004).
46. Nesvizhskii, A.I. & Aebersold, R. Interpretation of shotgun proteomic data - The protein inference problem. *Mol. Cell. Proteomics* **4**, 1419–1440 (2005).
47. Swaney, D.L., Wenger, C.D., Thomson, J.A. & Coon, J.J. Human embryonic stem cell phosphoproteome revealed by electron transfer dissociation tandem mass spectrometry. *Proc. Natl. Acad. Sci. USA* **106**, 995–1000 (2009).

doi:10.1038/nmeth.1699