Dynamic proteomic profiling of extra-embryonic endoderm differentiation in mouse embryonic stem cells

Claire M. Mulvey, Christian Schröter, Laurent Gatto, Duygu Dikicioglu, Isik Baris Fidaner, Andy Christoforou, Michael J. Deery, Lily T. Y. Cho, Kathy K. Niakan, Alfonso Martinez-Arias and Kathryn S. Lilley

1Cambridge Centre for Proteomics, Department of Biochemistry, University of Cambridge, CB2 1QR, UK. 2Cambridge System Biology Centre, Wellcome Trust Stem Cell building, University of Cambridge, CB2 1QR, UK. 3Department of Genetics, University of Cambridge, CB2 3EH, UK. 4Computational Proteomics Unit, Department of Biochemistry, Tennis Court Road, University of Cambridge, CB2 1QR, UK. 5Department of Biochemistry, University of Cambridge, CB2 1QR, UK. 6Department of Computer Engineering, Bogazici University, 34342 Istanbul, Turkey. 7Neusentis, Pfizer Worldwide Research & Development, Granta Park Science Park, Great Abington, Cambridge, CB21 6GS, United Kingdom. 8Division of Stem Cell Biology and Developmental Genetics, Medical Research Council National Institute for Medical Research, Mill Hill, London, NW7 1AA, UK.

¥ These authors contributed equally to this work.

*Corresponding authors: k.s.lilley@bioc.cam.ac.uk and ama11@hermes.cam.ac.uk

Running Title: Proteomics of embryonic stem cell differentiation

KEYWORDS: Differentiation / Extra-embryonic Endoderm (XEN) stem cells / Pluripotency / Quantitative Proteomics / Tandem Mass Tags (TMT)
During mammalian pre-implantation development, the cells of the blastocyst’s inner cell mass differentiate into the epiblast and primitive endoderm lineages, which give rise to the fetus and extra-embryonic tissues, respectively. Extra-embryonic endoderm differentiation can be modeled in vitro by induced expression of GATA transcription factors in mouse embryonic stem cells. Here we use this GATA-inducible system to quantitatively monitor the dynamics of global proteomic changes during the early stages of this differentiation event and also investigate the fully differentiated phenotype, as represented by embryo-derived extra-embryonic endoderm (XEN) cells. Using mass spectrometry-based quantitative proteomic profiling with multivariate data analysis tools, we reproducibly quantified 2,336 proteins across three biological replicates and have identified clusters of proteins characterized by distinct, dynamic temporal abundance profiles. We first used this approach to highlight novel marker candidates of the pluripotent state and extra-embryonic endoderm differentiation. Through functional annotation enrichment analysis, we have shown that the downregulation of chromatin-modifying enzymes, the re-organization of membrane trafficking machinery and the breakdown of cell-cell adhesion are successive steps of the extra-embryonic differentiation process. Thus, applying a range of sophisticated clustering approaches to a time-resolved proteomic dataset has allowed the elucidation of complex biological processes which characterize stem cell differentiation and could establish a general paradigm for the investigation of these processes.
INTRODUCTION

The specification of different cell types during development is a complex process that is organized in a hierarchical fashion. Lineage-specific transcription factors activate the expression of batteries of genes which endow cells with particular phenotypes that allow them to change their morphologies and exert defined functions. One of the earliest examples of this process is found in the mammalian pre-implantation embryo. At embryonic day E3.0, the mouse embryo is comprised of an outer epithelial layer, the trophoderm, that will give rise to the placenta and an inner cell mass (ICM) which will be subdivided into the Primitive Endoderm (PrE) and the Epiblast (Epi). Some extra-embryonic tissues are derived from the PrE, while the Epi develops into the embryo proper. Initially, all cells in the ICM co-express transcription factors for the Epi fate, such as OCT4, SOX2 and NANOG, together with the transcriptional regulator GATA6 that promotes PrE specification and is a marker for this fate at later stages. Cells within the ICM that undergo PrE specification progressively downregulate the Epi-specific transcription factors concomitantly with an increase in the expression of GATA6 and other PrE-specific transcriptional regulators such as GATA4 and SOX17 [1-3]. Importantly, in addition to these differences in transcription factor expression, the PrE and the Epi are also morphologically distinct: while the Epi prior to implantation has mesenchymal character, the PrE initially forms an epithelium lining the epiblast. Subsequently, the PrE further differentiates into the epithelial visceral endoderm, a tissue surrounding the post-implantation epiblast, and the parietal endoderm, which deposits the extracellular matrix of Reichert’s membrane [4, 5].

Some of the early fate decisions of the mammalian embryo can be recapitulated in culture and thus provide a good experimental system to further define and investigate both the transcriptional control as well as the execution of cell fate change. Pre-implantation blastocysts can be used to obtain self-renewing cell populations representing the Epi and the
PrE. Embryonic Stem (ES) cells for instance share many functional and molecular characteristics of the Epi lineage, and extra-embryonic endoderm (XEN) cells display a molecular signature that resembles the PrE [6, 7]. ES and XEN cells have very different morphological characteristics: while ES cells grow in tight colonies, XEN cells have a flat mesenchymal appearance and are highly motile. ES cells grown under specific experimental conditions can contribute to extra-embryonic tissues in chimaera experiments [3, 8, 9], and spontaneous differentiation of XEN-like cells in ES cell cultures has been observed [10]. Most significantly, induced expression of GATA factors in ES cells will produce XEN-like cells efficiently [11, 12].

While the role of transcriptional regulators that drive the divergence of pluripotent Epi and PrE specification has been intensively investigated in recent years, the temporal hierarchy of the protein expression programs that they direct and the global changes that result when cells undergo differentiation are only beginning to be elucidated. The most straightforward way of analyzing these global changes is by transcriptomic microarray analysis or RNA-seq (see [10] for an example in the context of differentiation of XEN-like cells). However, changes at the mRNA level do not necessarily reflect changes at the protein expression level [13], emphasizing the need to investigate the execution of cell differentiation with proteomic techniques. Several large-scale proteomics studies have begun to address the processes involved in regulation and maintenance of pluripotency, complementing the wealth of information generated from transcriptomic and epigenetics studies (see [14] for a recent review). However, these studies often use either undirected differentiation protocols, focus on a specific subcellular region, or compare only a few terminally differentiated states [15-17]. We reasoned that the experimentally controlled differentiation of extra-embryonic cells from ES cells through inducible expression of GATA factors would represent an ideal case to
delineate global molecular changes in the differentiating proteome and define the dynamic, temporal sequence of events that underlie the early stages of this process.

Using a TMT-differential labeling mass spectrometry-based approach, we have robustly quantified over 2,000 proteins associated with the transition from an ES- to a XEN-like phenotype following inducible GATA4 expression. A range of multivariate data analysis methods were used to define clusters of proteins whose temporal abundance profiles correlate with morphological changes. This approach has demonstrated the equivalence of GATA4 and GATA6 to implement the ES-to-XEN transition, and identified novel indicators of both the self-renewing, pluripotent ES-cell population and the fully differentiated XEN state. We find that the onset of differentiation induced by GATA4 triggers alterations in chromatin remodeling factors, enhanced membrane trafficking, reorganization of extracellular matrix and components of the adherens junctions, as well as altered metabolic capacity. We also show that some effector molecules of ERK/MAPK signaling may be involved in the transition to a more migratory XEN-like phenotype. We anticipate that the results presented here will provide a valuable resource for the further investigation of the cellular-wide changes induced during lineage commitment in mouse embryonic stem cells and the early cell fate decision \textit{in vivo}.
MATERIALS AND METHODS

Cell culture, GATA4 and GATA6 overexpression

ES and XEN cells were cultured on gelatin-coated plastic in Glasgow Minimal Essential Medium (GMEM) supplemented with 10% ES-cell qualified fetal bovine serum, non-essential amino acids, sodium pyruvate, stable glutamine, β-mercaptoethanol and leukemia inhibitory factor (LIF). Medium was changed daily, and cells passaged as required upon reaching confluence.

The full-length mouse cDNA for Gata4 or Gata6 was amplified from XEN cells, FLAG-tagged and cloned into a plasmid carrying a bi-directional tetracycline-inducible minimal CMV promoter that simultaneously drives the expression of a gene coding for a red fluorescent protein (dsRed). Co-electroporation in mouse ES cells of this plasmid together with a plasmid expressing the FLPe recombinase allowed for successful targeting to a modified ColA1 locus, as described previously [18]. The ES cells also contain a modified ROSA26-locus that drives doxycycline-inducible expression of an M2rtTA transactivator. Successfully targeted cells were selected using puromycine (1 µg/ml) and hygromycin (50-100 µg/ml), and will be described in further detail elsewhere (Wamaitha S.E. et al., submitted). GATA4-FLAG and GATA6-FLAG expression was induced by addition of 500 ng/ml doxycycline to the culture medium.

Western blotting and immunohistochemistry

Cells were harvested in 8 M urea, 0.1% SDS, 25 mM Triethyl ammonium bicarbonate (TEAB), pH 8.5, with protease inhibitors (Roche, Burgess Hill, West Sussex, UK) and protein concentration was evaluated with the BCA protein assay (Thermo Fisher Scientific, Loughborough, Leics., UK, www.lifetechnologies.com). To further ensure comparable
protein levels between samples, for each of the biological replicates a control blot was probed with an antibody directed against alpha-tubulin. Between 8 and 20 µg of each sample was loaded onto 4 - 15% pre-cast SDS PAGE gels (Bio-Rad, Hercules, CA, www.bio-rad.com) and transferred onto nitrocellulose membranes using the Trans Blot Turbo transfer system (Bio-Rad, Hercules, CA, www.bio-rad.com). Primary antibody information can be found in Supporting Information Table S1. Secondary antibodies were fluorescently labeled (IRDye, LI-COR) or HRP-conjugated (Goat anti-Rabbit IgG-HRP, BioRad; Donkey anti-Goat IgG-HRP, Santa Cruz, Dallas, TX, www.scbt.com; Goat anti-Mouse IgG-HRP, Bio-Rad), and visualized by infrared fluorescence detection (LI-COR, Lincoln, NE, www.licor.com) or enhanced chemiluminescence (GE Healthcare, Amersham, Bucks, UK, www.gehealthcare.co.uk).

For immunocytochemistry, cells were plated on ibidi µ-slides, and staining was performed according to standard procedures. Briefly, cells were washed, fixed for 15 minutes in buffered 4% formaldehyde and permeabilized with 0.1% Triton X-100 in the presence of 1% bovine serum albumin. Primary antibodies were diluted in permeabilization/blocking buffer as described in Supporting Information Table S1 and incubated overnight. Secondary antibodies were Alexa-Fluor conjugated IgG from Molecular Probes (Loughborough, Leics., UK, www.lifetechnologies.com) and diluted 1:500. Hoechst 33342 was added at 10µg/ml to the secondary antibody solution to visualize nuclei. Images were captured on a Zeiss LSM700 confocal microscope using a 40X/1.4 NA oil immersion lens.

**TMT labeling, peptide fractionation, mass spectrometry and data analysis**

100 µg each of the 0hr, 16hr, 24hr, 48hr, 72hr doxycycline-inducible GATA4 lysates as well as the XEN lysates were prepared for labeling with six-plex TMT reagents according to the manufacturer’s instructions (Thermo Fisher Scientific, Loughborough, Leics., UK, www.lifetechnologies.com). TMT labeling was performed separately for each replicate
experiment, and the order of the tags was reversed for one of the replicates. The same time-points for the GATA6-induced time-course were also separately labeled and processed for mass spectrometry.


Multivariate analysis

The multivariate analysis was conducted using the Bioconductor [19] pRoloc package [20, 21] implemented in the R statistical programming language. Support vector machines (SVM) classification and \( k \)-means clustering analysis were performed on the GATA4-induced dataset. The ClueGO functional annotation plugin for Cytoscape was used to identify functionally-related groups of enriched GO terms and pathway networks associated with the identified SVM and \( k \)-means clusters [22]. The CLUSTERnGO (CnG) software and source code to carry out model based clustering are available online.
Biological replicates were treated as replicates in CnG when prompted. See Supplementary Methods for further details on the multivariate analysis.
RESULTS

A doxycycline-inducible GATA expression system to study proteome-level changes

To investigate the transition from an Epi to a XEN-like state, we used recombinase-mediated cassette exchange to develop ES cell lines carrying Gata4 or Gata6 cDNAs under the control of doxycycline-inducible promoters [18 and (Wamaitha et al., submitted)]. Addition of doxycycline to the culture medium induced efficient transgene expression in the majority of cells (Fig. 1A, 1B). Western blotting and quantitative mass spectrometry showed that expression levels of inducible GATA4 protein were comparable to those observed in embryo-derived XEN cells (Fig. 1C, 1D). Induction of GATA4 expression led to the rapid and simultaneous downregulation of the pluripotency marker NANOG (Fig. 1C, 1E) and upregulation of the PrE marker SOX17 (Fig. 1C). This doxycycline-inducible expression system therefore recapitulates key findings that were previously obtained using alternative GATA expression systems [11, 12].

Within three days of continuous doxycycline treatment, the ES cells gradually assumed a phenotype that closely resembled embryo-derived XEN cells (Fig. 1F; Supplementary Movie). We decided to delineate the changes in the total proteome during the early phase of this fate transition at 0, 16, 24, 48 and 72 hours after doxycycline addition. In parallel, we monitored an embryo-derived XEN cell line, representing the fully differentiated phenotype. Each sample was proteolytically digested and differentially labeled with amine-reactive Tandem Mass Tags (TMT) (Fig. 2A). This quantitative approach allowed all six samples to be simultaneously co-analyzed by liquid chromatography-mass spectrometry (LC-MS). The resulting dataset provides a dynamic account of the global proteomic changes as cells transit from a pluripotent state towards PrE differentiation.
In three independent biological replicate experiments, a total of 3,044 proteins were quantified. Of these, 2,336 proteins (76%) were robustly quantified at all time-points and in all replicate experiments (Fig. 2B; Supporting Information Tables S2, S3). The individual time-points of all three biological replicates were hierarchically clustered and confirmed excellent reproducibility between independent experiments (Fig. 2C). Furthermore, the first 24 hours of the time-course were distinct from the 48hr and 72hr time-points; while the entire time-course clustered separately to the XEN cells. We conclude that this quantitative proteomics approach can reliably track a large proportion of the proteome and capture gross changes during the transition from an ES- to a XEN-like phenotype.

**Induction of GATA4 or GATA6 expression induces similar global proteomic changes**

Previous studies have used both *Gata4* and *Gata6* overexpression to induce a XEN-like phenotype in ES cells, and these studies suggested that both proteins were equivalent as judged by marker gene expression and functional tests of the differentiated cells [11, 12]. In the mouse blastocyst however, *Gata6* and *Gata4* are expressed sequentially [2, 23], and it is possible that subtle functional differences between these factors have been previously missed. To address this question, we subjected ES cell lines expressing doxycycline-inducible GATA6 to the same differentiation time-course as the inducible GATA4 lines above. Western blot analysis using FLAG antibodies showed that both inducible GATA4 and GATA6 were expressed with similar dynamics following doxycycline addition, and triggered rapid expression of endogenous GATA factors with comparable efficiency (Fig. 3A), suggesting induced GATA4 and GATA6 have similar functions in inducing primitive endoderm-like gene expression.

TMT-labeling and mass spectrometry of the inducible GATA6 line quantified 1,765 proteins that were also identified in all GATA4-inducible replicate experiments (Supporting
Information Table S4). Dynamic changes throughout the two time-series were very similar for many proteins of interest (Supporting Information Fig. S1), including OCT4 and SOX2 (Fig. 3B). In fact, the GATA6- and GATA4-induced changes were strongly correlated when comparing the 0hr and 72hr time-point for all commonly identified proteins ($r^2 = 0.857$, Fig. 3C). Similar correlation coefficients were found when comparing each of the GATA4 replicates with the GATA6 data (Supporting Information Fig. S2), suggesting that the activity of GATA4 and GATA6 proteins in inducing a XEN-like fate are interchangeable. We therefore decided to focus on the dynamic GATA4-inducible triplicate dataset, and used multivariate bioinformatics tools to assign putative, novel markers of pluripotency and differentiation, as well as to cluster proteins which displayed similar temporal abundance patterns in response to the GATA4 stimulus.

**Novel indicators of pluripotency and XEN cell fate can be assigned using a support vector machine learning algorithm (SVM)**

Pluripotency is regulated by a core transcriptional network of three transcription factors, NANOG, SOX2, and OCT4. While we failed to reliably detect NANOG protein in the GATA4-inducible proteomics dataset due to its low abundance, extensive post-translational modification and lack of tryptic sites amenable for mass spectrometry [24], SOX2 was found to rapidly downregulate within the first 24hr following doxycycline exposure, while OCT4 expression was gradually reduced but still detectable at 72hr (Fig. 3B; Supporting Information Fig. S3). We reasoned that interrogating the proteomic dataset for proteins with dynamic changes similar to those of the core transcription factors might reveal novel markers of pluripotency and the XEN fate, as well as highlight key molecular pathways underlying the respective states.
The temporal abundance profiles of a selection of well-curated “pluripotency” and “differentiation” marker proteins were used as training data for a support vector machine learning (SVM) algorithm implemented in the pRoloc package [20, 21]. Proteins were assigned to clusters defined by these marker proteins based on their similarity to the training data. The SVM scores are a-posteriori class probability predictions: high scores are given to proteins which demonstrate similar dynamic abundance profiles to the training set. Using optimized parameters, 88 and 56 proteins were robustly classified to either the “pluripotency” or “differentiation” clusters, respectively (Supporting Information Table S2). Fig. 4A depicts the localization of all proteins in the dataset in principle component (PC) space, highlights proteins of the training sets (black circles) and the newly identified members of the pluripotency (blue) and differentiation (red) cluster. The excellent separation of the clusters in PC space suggests that we have identified sets of proteins suited to distinguish the pluripotent state of mouse ES cells from the differentiated XEN-like state.

Proteins in the pluripotency-related cluster downregulate in abundance following the onset of GATA4 expression. 18 proteins within this group were matched to the PluriNetwork, an electronic repository for regulatory networks and molecular mechanisms underlying pluripotency [25] (Supporting Information Table S5). Several other proteins in this cluster have also been linked to pluripotency, such as L1TD1 [26], LYAR [27], DPY30 [28] and TDH [29]. “Heterochromatin” was identified as the most significantly enriched Gene Ontology (GO) term associated with the pluripotency group and many additional chromatin remodeling proteins were identified just below the SVM score threshold (Supporting Information Fig. S4). The temporal abundance pattern for one example from this group, JARID2, was confirmed by western blotting and immuno-staining (Fig. 4B). JARID2 is a key regulator of embryonic stem cell development and mediates the recruitment of the Polycomb Repressive Complex (PRC2) to target genes [30-33]. Our results reaffirm that the pluripotent
state is characterized by dedicated molecular machinery that maintain a specific chromatin state, which rapidly changes upon PrE differentiation [34].

Inspection of the differentiation cluster suggested that increased deposition of basement membrane and extracellular matrix (ECM) is a hallmark of XEN cell differentiation. While our training set contained a number of ECM components that had previously been implicated in XEN cell differentiation such as COL4A1, COL4A2, LAMA1, and HSPG2 [10], functionally related proteins featured prominently amongst the newly identified components of this cluster, for example the nidogens NID1 and NID2 and the collagen biosynthetic enzymes PLOD1 and GLT25D1. The differentiation cluster also suggested that components of endocytic protein trafficking pathways are co-regulated during extra-embryonic endoderm differentiation. Our SVM training set included the endocytic receptor cubilin, which is essential for the formation of definitive and visceral endoderm [35], and the adapter protein DAB2 [10, 36]. Several interactors of cubilin and DAB2 were subsequently classified to the differentiation cluster, such as LRP2 (megalin), a major binding partner of cubilin [37, 38] which has been linked to PrE formation [39], and its binding partner LRPAP1, unconventional Myosin-6, which can direct DAB2 trafficking [40], and the LDLR chaperone MESD, which is essential for apical localization of LRP2 in the visceral endoderm [41] (Fig. 4C).

The SVM analysis also highlighted a number of proteins relating to MAPK/ERK signaling. Sprouty4 for example, inhibits FGF signaling and was identified in the pluripotency-associated group (Fig. 4D). GLIPR2 and FAM129B were both assigned to the differentiation cluster (Fig. 4D) and have recently been identified as direct transcriptional targets of ERK signaling in HK-2 cells and human melanoma cells, respectively [42, 43]. This suggests that cells upregulate FGF/MAPK signaling by shutting down negative feedback components during PrE differentiation, thereby ensuring the continuous FGF/MAPK signaling required
for PrE differentiation [44, 45]. Consistent with this idea, we found increased phosphorylation levels of the MAP kinase ERK by immunoblotting (Fig. 4E) along the differentiation time-course, although phosphorylation levels in established XEN cell lines were lower than after 72h of doxycycline exposure. Both GLIPR2 and FAM129B have been implicated in endowing cells with increased motility [42, 43], suggesting that they provide a link between the dependence of PrE differentiation on FGF/MAPK signaling [44, 45] and acquisition of a motile phenotype upon differentiation along this lineage.

**K-means clustering identifies global temporal abundance changes**

We next performed k-means clustering to obtain an overview of all global changes and to identify further clusters of proteins with dynamic profiles different from those represented by our SVM training set. Fig. 5A depicts the 20 k-means clusters in PC space. The temporal profiles for proteins within each cluster shows that they display similar abundance patterns, i.e. time of onset and magnitude of response (Fig. 5B, Supporting Information Table S2). We observed good correlation between our supervised (SVM) and unsupervised (k-means) clustering methods (Supporting Information Fig. S5).

To obtain a functional insight into the temporal order of events underlying the early phases of fate transition, we performed GO enrichment on all k-means clusters (Table 1; Supporting Information Table S6). Expression of proteins in Cluster 2 for example, were initially unaffected by GATA4 induction, but were subsequently down-regulated at 48hr to a very low level. This cluster was most enriched for the term “nuclear chromatin”. Conversely, Cluster 13 was upregulated at 48hr of differentiation and was highly enriched for “endocytosis” and “membrane trafficking”, including proteins such as clathrins, coatamers, sorting nexins, exocyst docking proteins and endosomal sorting proteins (Fig. 5C). This demonstrates that re-
organization of the endomembrane system occurs after a lag time during the transition towards a XEN-like state.

\(K\)-means analysis also revealed a number of clusters whose relative abundance was much greater in XEN cells than during the GATA4 time-course (Clusters 16 - 20). Overrepresented functional terms for these clusters highlight processes that differ between the embryo-derived XEN cells and the GATA4-induced XEN-like cell type, potentially because these processes are further downstream in the differentiation cascade and not yet activated by 72hr of GATA4 expression. Clusters 18 and 19 were highly enriched for “protein processing in the endoplasmic reticulum (ER)” and further inspection of the dataset revealed that enzymes involved in fatty acid biosynthesis, glycosylation, unfolded protein response and collagen biosynthesis were upregulated in XEN cells (Supporting Information Table S7). GO term analysis for Clusters 16 and 17 revealed enrichment of the terms “citric acid (TCA) cycle and respiratory electron transport” and “mitochondrion” (Table 1). Specifically, we found that XEN cells contained a moderate increase in components of the mitochondrial electron respiratory chain and enzymes involved in the TCA cycle (Fig. 5D). This finding supports the observation that differentiated cells have a higher reliance on oxidative phosphorylation than ES cells, and may undergo mitochondrial biogenesis in order to maintain their increased metabolic demands [46]. Consistent with the fact that ES cells rely predominantly on glycolysis, we find a gradual decrease in expression of the rate-limiting glycolytic enzyme hexokinase-2 (HK2) and SLC2A1 (GLUT1), the major facilitative glucose importer in ES cells, upon differentiation. Many other glycolytic enzymes however, are unchanged or even upregulated in our dataset (e.g. HK1, PFKL, GPI, PGK1), suggesting there may be more complex adjustments to the metabolic status of these cells.

As expected, many proteins remained unchanged throughout the experiment (e.g. Cluster 7 and 8). We also found two clusters (11 and 12) that were upregulated during the time-course
but remained at very low abundance in XEN cells. A longer time-course would be required to distinguish whether these proteins are transiently expressed during the early phase of the transition, or whether they represent effects that specifically result from the long-term exposure of cells to the doxycycline-induced GATA factors.

**Additional biologically informative clusters are revealed using the novel model-based algorithm CLUSTERnGO.**

The \(k\)-means clustering results suggested there is an almost-continuous distribution of dynamic profiles in our dataset that may not be fully resolved within 20 discrete clusters. To address this, we applied CLUSTERnGO (CnG), a model-based clustering algorithm specifically designed for analyzing dynamic time-series datasets that does not place any constraints on the size or number of clusters assigned (see [http://www.cmpe.boun.edu.tr/content/CnG](http://www.cmpe.boun.edu.tr/content/CnG) for further details and source code). In this way, we hoped to identify subsets of proteins with strongly correlated reporter ion profiles that may have been obscured in the \(k\)-means approach.

Using optimized parameters, this analysis revealed 206 distinct multi-protein clusters (average of 16 proteins per cluster) with only 30 proteins assigned to single-member clusters (Supporting Information Tables S2, S8). Cluster 54 (Fig. 6A) containing 23 proteins, was initially unchanged in abundance by the differentiation signal, but subsequently was rapidly down-regulated at 48 hours. All proteins in this cluster were assigned to the larger \(k\)-means cluster 2 containing 66 proteins. Cluster 54 includes ZSCAN10, a transcription factor known to play a key role in maintaining pluripotency [47], \(\alpha\)-catenin, \(\beta\)-catenin and cadherin-1 (E-cadherin) which are core components of the adherens junctions, as well as AJUBA, a LIM-domain protein which co-localizes with cadherins and catenins [48, 49]. We confirmed this pattern of temporal abundance changes for \(\beta\)-catenin and cadherin-1 by western blotting and
immuno-staining (Fig. 6B). The cell adhesion molecules EPCAM, BCAM and integrin-α6 (ITGA6) display a similar pattern of expression although they were not assigned to the same cluster (Fig. 6C). Interestingly, the time-point at which we begin to detect downregulation of these junctional proteins (48hr) coincides with distinct morphological changes such as breakdown of cell-to-cell contact and dispersal of the tightly adherent ES cell colonies (Supplementary movie). This congruence between the dynamics of phenotypic change and the behavior of proteins within one specific cluster suggests that this clustering approach allows us to infer a temporal order of events during the early stages of differentiation.

CnG Cluster 102 is composed of 9 tightly co-regulated proteins and contains all six components of the minichromosome maintenance complex (MCM2-7) which is essential for DNA replication initiation, as well as DNA topoisomerase 2-alpha, an important regulatory enzyme for DNA replication and transcription (Fig. 6D). Seven of the nine proteins in this cluster had been assigned to the k-means Cluster 3, which was itself enriched for the GO term “DNA replication”. Taken together, clustering by k-means and CnG suggests that a group of proteins involved in the control of transcription and maintaining genomic integrity is gradually down-regulated as cells differentiate towards extra-embryonic fates, reflecting a slight decline in proliferation over the course of the experiment.

Partitioning the dataset into a larger number of clusters using CnG has identified small cohorts of tightly co-regulated proteins, some of which had been overlooked in the k-means analysis. The majority of CnG groups however, were found to be subsets of those previously identified with k-means, indicating that the major biological findings from this dataset are not dependent on the clustering method used.
DISCUSSION

Understanding the mechanisms that underlie the functional specialization of stem cells as they differentiate is of paramount importance for the field of stem cell biology and regenerative medicine. In this study, we have made use of the fact that ES cells can be differentiated in an efficient and synchronous manner towards PrE-like lineages through the induced expression of GATA transcription factors [11, 12], and used this system to investigate with high temporal resolution the dynamic changes that occur in the cell’s proteome as they alter their fate and function.

Isobaric labeling and mass spectrometry-based proteomic profiling coupled with multivariate data analysis techniques allowed us to capture the dynamic and sequential response of over 2,000 proteins following GATA4 induction. SVM classification prioritized proteins of interest from a large list of candidates by identifying proteins whose expression dynamics are highly correlated to well-curated markers of pluripotency or differentiation. Functional validation experiments will be necessary to further refine this group and to establish their relationship to each other and to the core pluripotency transcription factors.

The high temporal resolution of our dataset allowed us to define the sequences of change between the pluripotent ES-cell state and extra-embryonic differentiation. Both $k$-means clustering and a novel, model-based clustering approach identified temporally co-regulated proteins. For example, we found that one of the first global changes following GATA4 induction is alteration of the chromatin-regulatory network. Later along the differentiation time-course, we detected loss of the expression of the core components of adherens junctions and cell adhesion molecules and a concomitant increase in the deposition of ECM and basement membrane proteins, corresponding with alterations in the morphology of the cells and a coincident increase in cell motility. Together, these observations suggest that our differentiation time-course captures an epithelial-mesenchymal transition (EMT). ES cells
form adherent cell sheets in culture, and the primitive endoderm of the pre-implantation blastocyst and the visceral endoderm of the post-implantation embryo display distinct epithelial characteristics whereas the parietal endoderm is a dispersed cell population specialized in ECM production. Because embryo-derived XEN cells and XEN-like cells obtained by induced GATA factor expression mostly contribute to the parietal endoderm [12, 50], it appears likely that the EMT observed in our dataset recapitulates events underlying the emergence of the parietal endoderm in the embryo. Importantly, EMTs are a recurrent feature of cell fate conversions during embryonic development, and also play an important role in tumor progression and metastasis. The EMT-like changes detected in our dataset occur simultaneously with an up-regulation in membrane trafficking and defined metabolic alterations. It will be interesting to investigate whether this is a general theme in different EMTs during differentiation and disease.

Global cellular changes of the transition from an ES to a XEN-like state have recently been investigated at the transcriptomic level using inducible Sox17 expression [50]. This study revealed transcript-level alterations in several functional processes that were also implicated in our GATA4-inducible proteomics dataset. Inducible expression of Sox17 however appears to trigger extra-embryonic differentiation with a different dynamic signature to that observed in the embryo and upon inducible GATA expression, making it difficult to directly compare the results of this study with ours.

While we have delineated the major proteomic changes that occur as cells differentiate and restrict their lineage potential, a recent publication examined the temporal proteomic dynamics of the reverse process, i.e. when cells acquire pluripotency through the reprogramming of mouse embryonic fibroblasts (MEFs) to induced pluripotent stem cells (iPSCs) [51]. Surprisingly, the functional processes that underlie the intermediate phase of this transition are very similar to those identified in our study. Hansson et al., demonstrate
that cellular reprogramming involves changes in the stoichiometry of electron transport chain complexes, a switch from mitochondrial oxidation in MEFs to glycolysis in iPSCs, chromatin remodeling, altered vesicle-mediated transport, protein processing in the ER, and an EMT-like process involving changes in ECM and cell adhesion molecules.

Conclusion:

In summary, the temporal dissection of our proteomic dataset has revealed sequential changes in the differentiating proteome that may represent crucial hallmarks of the transition from a pluripotent to an extra-embryonic, differentiated state. This dataset provides a valuable insight into lineage determination during early embryogenesis and will serve as an important resource for the stem cell community.
Data Submission

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the dataset identifier PXD001901 and 10.6019/PXD001901 [52, 53].

Acknowledgements

This work was supported by the European Union 7th Framework Program (PRIME-XS project grant number 262067 to K.S.L., L.G and C.M.M), the Biotechnology and Biological Sciences Research Council (BBSRC grant number BB/L002817/1 to K.S.L and L.G.), as well as a HFSP grant (RGP0029/2010) and a European Research Council (ERC) Advanced Investigator grant to A.M.A.. C.S was supported by an EMBO long term fellowship and a Marie Curie IEF. L.T.Y.C. and K.K.N. were supported by the Medical Research Council (MRC, UK, MC_UP_1202/9) and the March of Dimes Foundation (FY11-436). We also thank Professor Steve Oliver and Dr. A.K. Hadjantonakis for helpful discussions and advice.

Author Contributions

C.M.M., C.S., A.M.A. and K.S.L. designed and planned the experiments. C.M.M. performed the proteomics sample preparation, mass spectrometry, data analysis and western blotting. C.S. performed the cell culture, transfections, FACs analysis and immuno-staining. L.G. performed the multivariate data analysis and submission to the ProteomeXchange repository. D.D. and I.B.F. performed the CLUSTERnGO analysis, A.C. assisted with data analysis and M.J.D. assisted with mass spectrometry. L.T.Y.C. and K.K.N. generated the GATA4 and GATA6 cell lines and performed initial characterization of the cell culture system. C.M.M., C.S., A.M.A. and K.S.L. wrote the manuscript and all authors approved the final manuscript.

Conflict of Interest
The authors declare that there are no conflicts of interest.
Figure Legends

Figure 1. Extra-embryonic differentiation induced by doxycycline-controlled Gata4 expression. (A) Immuno-staining for the FLAG-tag in ES cells carrying a doxycycline-inducible Gata4-FLAG transgene 24hr after addition of doxycycline. The majority of cells express the inducible GATA4-FLAG protein. (B) Flow cytometry of cells treated as in (A) and immuno-stained for the FLAG tag. 80% of cells express the inducible GATA4-FLAG protein. (C) Western blot analysis of GATA4, NANOG and SOX17 expression during three days of induced GATA4 expression. A strong GATA4 signal, comprising both inducible and endogenous protein is detected within 16 hours of doxycycline-induction, followed by rapid downregulation of the pluripotency marker NANOG and upregulation of SOX17, a marker of primitive endoderm differentiation. Alpha-tubulin was used as a loading control (TUBA1C). (D) GATA4 protein expression measured in the TMT-proteomics dataset (dots indicate normalized reporter ion intensities from independent experiments, bar indicates mean). GATA4 expression reaches a similar level in ES cells carrying the inducible transgene as compared to embryo-derived XEN cells. Note that both western blot analysis and TMT ion intensities report both inducible GATA4-FLAG and endogenous GATA4 expression; see Fig. 3A for a distinction between induced and endogenous GATA expression. (E) Immunostaining of for NANOG expression in ES cells carrying the doxycycline-inducible Gata4-FLAG transgene at different time-points after doxycycline addition. NANOG expression is rapidly and near-simultaneously down-regulated in the population. (F) Phase contrast microscopy of cells undergoing the three day time-course of GATA4 expression show the morphological changes of cells acquiring a XEN-like phenotype. Abbreviation: XEN, extra-embryonic endoderm.
Figure 2. Quantitative proteomics analysis of ES cell differentiation following inducible GATA4 expression. (A) Experimental workflow: Five time-points of GATA4 expression (0, 16, 24, 48 and 72 hour) were used to investigate early differentiation events; XEN cells were used to represent fully differentiated primitive endoderm (PrE) cells. Cells were harvested and processed for TMT labeling as described in the Methods. Downstream bioinformatics analysis was performed with SVM classification, k-means clustering, model-based clustering and GO annotation (see Methods). (B) Venn diagram representing the overlap in proteins that were robustly quantified in the three replicate experiments. A total of 2,336 proteins were quantified in all replicates. (C) Hierarchical clustering demonstrates excellent reproducibility between biological replicate experiments. Abbreviations: DOX, doxycycline; GO, gene ontology; HCD, Higher-energy collisional dissociation; LC-MS, liquid chromatography-mass spectrometry; mESC, mouse embryonic stem cells; SVM, support vector machine; TMT, tandem mass tags; XEN, extra-embryonic endoderm.

Figure 3. Overexpression of GATA6 induces the same phenotype as GATA4. (A) Comparison of the expression dynamics of inducible and endogenous GATA factors in cell lines carrying an inducible Gata4-FLAG (left) or Gata6-FLAG (right) transgene. FLAG antibodies (upper panels) specifically detect the inducible protein. Because of its slightly lower molecular weight, simultaneous detection of FLAG-tagged and total GATA protein using secondary antibodies labeled with distinct fluorophores reveals the endogenous protein as a distinct band (red) below the signal from the FLAG-tagged protein in the overlay image (fourth panel from above). Expression of endogenous GATA4 and GATA6 occurs with similar dynamics following doxycycline addition in both cell lines. (B) Left: Scatter plots of the relative protein abundance for SOX2 and OCT4 in each of the three GATA4 replicates
(blue circles) and the single GATA6 experiment (red diamond) as determined by mass spectrometry. Right: Western blotting of SOX2 and OCT4 in total lysates obtained from the GATA4 and the GATA6-inducible time-courses. The dynamics of downregulation for these proteins is similar following inducible GATA4 and GATA6 expression. (C) Linear regression analysis for the 1,765 common proteins quantified in the GATA4 triplicate study and the GATA6 experiment. The ratio between the 72hr time-point and 0hr time-point (72hr / 0hr) for the average of the three GATA4 replicates versus the GATA6 experiment was plotted ($r^2 = 0.8574$). The full GATA6 dataset can be found in Supporting Information Table S4. Abbreviations: DOX, doxycycline; XEN, extra-embryonic endoderm.

**Figure 4. Support vector machine learning algorithm (SVM) identifies novel markers of pluripotency and differentiation.** (A) Principle Components Analysis (PCA) plot of all proteins identified reproducibly upon inducible GATA4 expression. Eight well-curated markers each of pluripotency and differentiation were used as training data for a SVM analysis (circled and labeled in black). Panels to the left and right show the temporal abundance profiles of the 8 pluripotency (blue, left) and differentiation markers (red, right), respectively. Colored spots on the PCA plot indicate proteins assigned to the pluripotency cluster (blue, 88 proteins classified using a threshold at the 95th percentile) and the differentiation cluster (red, 56 proteins classified using a 90th percentile cutoff); see Supporting Information Table S2. Unclassified proteins are shown in grey circles. (B) TMT reporter ion distribution, western blot and immuno-staining (right), for JARID2, a novel putative marker for the pluripotent state. See Fig. 3B for western blot loading control (C) TMT reporter ion distributions of MESD, LRPAP1 and LRP2, which were classified to the differentiation cluster. (D) Dynamic changes in proteins related to the FGF/MAPK signaling pathway; upper panels: TMT reporter ion distribution for GLIPR2 (left) and FAM129B (right); lower panels: Reporter ion distributions and western blot for SPRY4. Data points
represent the protein abundance in individual experiments, bars indicate mean. See Fig. 3B for western blot loading control (E) Western blot analysis for p42/p44 ERK phosphorylation and total ERK expression levels following induced GATA4-FLAG expression. Abbreviations: ERK, extracellular-signal-regulated kinases; PC, principle component; ppERK, doubly phosphorylated ERK; XEN, extra-embryonic endoderm.

**Figure 5. K-means clustering reveals distinct clusters in the GATA4-induced dataset.** (A) PCA plot of the entire dataset, where each spot represents the TMT distribution profile for a protein across the time-course. Each spot is colored by membership to one of twenty clusters identified by k-means clustering. (B) Temporal changes of reporter ion intensities for proteins in each cluster. Replicate time-points are represented next to each other. (C) Averaged temporal abundance profiles of selected proteins functionally associated with endocytic trafficking (identified in k-means cluster 13). (D) Averaged temporal abundance profiles of selected proteins functionally associated with the citrate acid (TCA) cycle and the respiratory electron transport chain (identified in k-means clusters 10, 14, 16, 17, 18).

**Figure 6. CLUSTERnGO reveals additional biologically relevant clusters.** (A) Averaged temporal abundance profiles of proteins in Cluster 54 of the CnG analysis. This cluster contained 23 proteins, including the core adherens junction components alpha-catenin, beta-catenin and cadherin-1 (E-cadherin). (B) Confocal microscopy and western blotting of beta-catenin and cadherin-1 confirms that they are reduced rapidly in abundance from 48 hours of GATA4 induction. (C) TMT reporter ion profiles for BCAM, EPCAM and ITGA6 (integrin-alpha6). Spots represent the protein abundance in three replicate experiments, bars indicate mean. (D) CnG Cluster 102, containing 9 proteins, six of which are components of the MCM2-7 complex. Shown are the averaged temporal abundance profiles for the three
GATA4 replicate experiments combined. Abbreviations: BCAM, basal cell adhesion molecule; CDH, cadherin-1; CTNNB, beta-catenin; EPCAM, epithelial cell adhesion molecule; ITGA6, integrin-a6; XEN, extra-embryonic endoderm.

**Table 1. GO annotation enrichment of k-means clusters.** ClueGO annotation enrichment analysis was used to identify functional processes associated with each cluster. The ClueGO results for each cluster can be found in Supporting Information Table S6.
Supporting Information

Supporting Information Figure Legends

**Figure S1.** The temporal abundance patterns for a selection of proteins of interest demonstrate that the GAT4 and GATA6 inducible expression systems affect expression of these proteins in a similar manner. The normalized TMT abundance profiles for the GATA4 replicates are shown as blue spots, whereas the GATA6 experiment is shown as red diamonds. Also shown are western blots for SPRY4, beta-catenin and JARID2 abundance during the GATA6-induced time-course, together with the tubulin control blot for this experiment.

**Figure S2.** Linear regression analysis to assess variability between experimental replicates and between induced GATA4 and GATA6 expression. Each graph compares two independent experiments by plotting the log2 change in abundance for each protein between 0hr and 72hr of doxycycline treatment in the two experiments. The r² values are shown on the plots and demonstrate excellent reproducibility between replicates, as well as a high degree of similarity in the proteomic response to induced GATA4 and GATA6 expression.

**Figure S3.** Immuno-stainings for NANOG, SOX2, OCT4 and GATA6 during the time-course of inducible GATA4 expression.

**Figure S4.** A selection of proteins involved in histone modification and chromosome remodeling, which are down-regulated in abundance following onset of GATA4 induction.

**Figure S5.** Upper panel: PCA plot illustrating cluster membership. Point sizes are proportional to SVM scores, i.e. cluster membership probabilities. Lower panel: The boxplots illustrate the SVM scores for the 20 clusters, ordered by their median score, showing, towards the right, clusters with most reliable membership.

**Figure S6.** Uncropped images for all western blots presented in this manuscript.
Supporting Information Tables:

**Table S1.** Antibody information for western blotting and immuno-staining

**Table S2.** Protein level spreadsheet for the GATA4 triplicate experiment. Columns are included for normalized reporter ion intensities, k-means cluster numbering, SVM scores (a score of 1 indicates that the protein was used in the training set), SVM (the class to which the score relates), SVM classification (whether a protein’s SVM score was above the threshold for inclusion in the class) and CLUSTERnGO (rep.ave) numbering. Columns are also presented for the mascot score, coverage, number of peptides and peptide-spectral-matches (PSMs) for each replicate experiment individually (A, B, C) as well as the sum of all replicates combined.

**Table S3.** Peptide-spectral-match (PSM) and peptide level data for 75,550 PSMs and 19,797 peptides corresponding to the 2,336 proteins quantified in the GATA4 triplicate dataset.

**Table S4.** Protein level data for 1,765 quantified proteins in the GATA6 single replicate experiment.

**Table S5.** Enriched functional annotations (ClueGO) for the Pluripotency and Differentiation SVM clusters.

**Table S6.** ClueGO results for all 20 k-means clusters (expanded information for Table 1).

**Table S7.** Proteins involved in sterol and fatty acid metabolism and biosynthesis, protein glycosylation, unfolded protein response and collagen biosynthesis ad metabolic processes, are upregulated in XEN cells.
Table S8. CLUSTERnGO table, showing the list of proteins assigned to each cluster with functional GO terms.

Supplementary Movie Time-lapse microscopy of cells carrying an inducible *Gata4-FLAG* transgene following addition of 500 ng/ml doxycycline to the culture medium.
REFERENCES


FIGURE 1.
FIGURE 2.
FIGURE 3.
FIGURE 4.
FIGURE 5.
FIGURE 6.
<table>
<thead>
<tr>
<th>Cluster No.</th>
<th>No. proteins / cluster</th>
<th>Enriched term</th>
<th>Term ID</th>
<th>No. proteins / term</th>
<th>Benjamini-Hochberg corrected p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>41</td>
<td>female pronucleus</td>
<td>GO:0001939</td>
<td>3</td>
<td>9.81E-06</td>
</tr>
<tr>
<td>2</td>
<td>66</td>
<td>nuclear chromatin</td>
<td>GO:0000790</td>
<td>7</td>
<td>1.22E-07</td>
</tr>
<tr>
<td>3</td>
<td>90</td>
<td>DNA replication</td>
<td>KEGG:03030</td>
<td>11</td>
<td>5.20E-16</td>
</tr>
<tr>
<td>4</td>
<td>177</td>
<td>S Phase</td>
<td>REACTOME:5416652</td>
<td>15</td>
<td>8.25E-11</td>
</tr>
<tr>
<td>5</td>
<td>196</td>
<td>RNA transport</td>
<td>KEGG:03013</td>
<td>19</td>
<td>7.77E-13</td>
</tr>
<tr>
<td>6</td>
<td>236</td>
<td>Ribosome biogenesis in eukaryotes</td>
<td>KEGG:03008</td>
<td>22</td>
<td>1.33E-21</td>
</tr>
<tr>
<td>7</td>
<td>392</td>
<td>Gene Expression</td>
<td>REACTOME:5416766</td>
<td>176</td>
<td>6.09E-116</td>
</tr>
<tr>
<td>8</td>
<td>250</td>
<td>M Phase</td>
<td>REACTOME:5417115</td>
<td>31</td>
<td>3.28E-19</td>
</tr>
<tr>
<td>9</td>
<td>160</td>
<td>Membrane Trafficking</td>
<td>REACTOME:5417343</td>
<td>11</td>
<td>1.02E-05</td>
</tr>
<tr>
<td>10</td>
<td>136</td>
<td>mitochondrial inner membrane</td>
<td>GO:0005743</td>
<td>16</td>
<td>2.51E-11</td>
</tr>
<tr>
<td>11</td>
<td>66</td>
<td>cytoskeleton organization</td>
<td>GO:0007010</td>
<td>10</td>
<td>2.38E-05</td>
</tr>
<tr>
<td>12</td>
<td>16</td>
<td>Clathrin derived vesicle budding</td>
<td>REACTOME:5417507</td>
<td>2</td>
<td>0.00169532</td>
</tr>
<tr>
<td>13</td>
<td>80</td>
<td>Endocytosis</td>
<td>KEGG:04144</td>
<td>17</td>
<td>1.30E-15</td>
</tr>
<tr>
<td>14</td>
<td>84</td>
<td>Protein processing in endoplasmic reticulum</td>
<td>KEGG:04141</td>
<td>9</td>
<td>4.13E-07</td>
</tr>
<tr>
<td>15</td>
<td>17</td>
<td>collagen type IV</td>
<td>GO:0005587</td>
<td>2</td>
<td>2.36E-04</td>
</tr>
<tr>
<td>16</td>
<td>144</td>
<td>Mitochondrion</td>
<td>GO:0005739</td>
<td>48</td>
<td>7.03E-21</td>
</tr>
<tr>
<td>17</td>
<td>66</td>
<td>The citric acid (TCA) cycle and respiratory electron transport</td>
<td>REACTOME:5416705</td>
<td>16</td>
<td>1.27E-16</td>
</tr>
<tr>
<td>18</td>
<td>71</td>
<td>Protein processing in endoplasmic reticulum</td>
<td>KEGG:04141</td>
<td>17</td>
<td>6.49E-21</td>
</tr>
<tr>
<td>19</td>
<td>33</td>
<td>Protein processing in endoplasmic reticulum</td>
<td>KEGG:04141</td>
<td>5</td>
<td>2.27E-05</td>
</tr>
<tr>
<td>20</td>
<td>16</td>
<td>laminin-10 complex</td>
<td>GO:0043259</td>
<td>2</td>
<td>3.97E-05</td>
</tr>
</tbody>
</table>

TABLE 1.