bioRxiv preprint doi: https://doi.org/10.1101/2023.10.04.559613; this version posted October 6, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Quantitative proteome dynamics across embryogenesis in a model chordate

Alexander N. Frese[1,2,*], Andrea Mariossi[1,2,*], Michael S. Levine[1,2,#], Martin Wühr[1,2,#]
1 Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, USA
2 Department of Molecular Biology, Princeton University, Princeton, NJ, USA
* equally contributed
# correspondence: msl2@princeton.edu, wuhr@princeton.edu

The evolution of gene expression programs underlying the development of vertebrates remains poorly characterized. Here, we present a comprehensive proteome atlas of the model chordate *Ciona,* covering eight developmental stages and ~7,000 translated genes, accompanied by a multi-omics analysis of co-evolution with the vertebrate *Xenopus*. Quantitative proteome comparisons argue against the widely held hourglass model, based solely on transcriptomic profiles, whereby peak conservation is observed during mid-developmental stages. Our analysis reveals maximal divergence at these stages, particularly gastrulation and neurulation. Together, our work provides a valuable resource for evaluating conservation and divergence of multi-omics profiles underlying the diversification of vertebrates.

Keywords: Proteome, Transcriptome, Development, Evolution, Chordates, Vertebrates, *Ciona*, *Xenopus*

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.04.559613; this version posted October 6, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

Urochordates are the nearest extant relatives to vertebrates and share several morphological and genomic traits (Delsuc et al. 2006). Their comparatively simple development and physiology make them attractive models to study the evolutionary origins of mechanisms in vertebrates, which have often evolved to higher complexity (Abitua et al. 2015, 2012; Berthelot et al. 2018; Christmas et al. 2023). Research on the model urochordate *Ciona* has substantially advanced the field of evolutionary and developmental biology, illuminating conserved mechanisms for body plan establishment, cellular organization, and gene regulatory networks (Abitua et al. 2015, 2012; Stolfi et al. 2010, 2015; Horie et al. 2018). These discoveries underscore the value of employing the relatively simple *Ciona* model to study processes in more complex organisms, including humans.

The assembly of the *Ciona* genome (Dehal et al. 2002) represented a significant landmark, triggering a swift surge in genome-wide analyses that included genomics, transcriptomics, epigenomics, and more recently, single-cell studies (Cao et al. 2019; Keller et al. 2016; Kubo et al. 2010; Madgwick et al. 2019; Suzuki et al. 2016; Zhang et al. 2020; Sladitschek et al. 2020). Transcriptomic studies are typically conducted to use mRNA dynamics as a surrogate for protein dynamics, the macromolecule that conveys function. Although direct measurement of protein dynamics would be more advantageous, proteome studies have been scarce, likely due to the persistent technical challenges associated with these investigations (Lopez et al. 2017; Nomura et al. 2009a; Yamada et al. 2009; Saxena et al. 2013). Technological advancements in quantitative proteomics, including multiplexing, enable sensitive and high precision measurements of protein abundances (Pappireddi et al. 2019; Johnson et al. 2021a; Thompson et al. 2003; Demichev et al. 2020; Ammar et al. 2023; McAlister et al. 2014). Proteomics studies of early development demonstrated a weak correlation between mRNA and protein levels, suggesting the importance of post-transcriptional regulatory mechanisms (Peshkin et al. 2015a; Ghaemmaghami et al. 2003; Schwanhäusser et al. 2011; Abdulghani et al. 2019; Ghazalpour et al. 2011; Vogel and Marcotte 2012; Smits et al. 2014a).

To date, proteome regulation in development remains poorly understood particularly in evolutionary contexts. Haeckel proposed that embryonic development recapitulates evolutionary trajectories (Haeckel 1866). While this theory is no longer accepted, more recent studies investigating the phylotypic period across species have generated conflicting results when based solely on transcriptome data and when limited to short phylogenetic distances within vertebrates or long distances spanning invertebrates to vertebrates (Chan et al. 2021; Uesaka et al. 2022). While some studies support the existence of a mid-embryonic period with maximal similarity in gene expression, in line with the hourglass model, others propose an inverted hourglass model, highlighting higher conservation of gene expression in the early and late developmental phases (Levin et al. 2016a; Marlétaz et al. 2018a; Hu et al. 2017a; Schrimpf et al. 2009a; Laurent et al. 2010a). To address these limitations, further investigations utilizing proteomic approaches in comparative analyses, particularly in chordates, are needed (Gil-Gálvez et al. 2022; Marlétaz et al. 2018b; Touceda-Suárez et al. 2020).

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.04.559613; this version posted October 6, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

To bridge this gap, we utilized state-of-the-art proteomics to quantify the maternal proteins deposited in unfertilized eggs, as well as proteome dynamics throughout *Ciona* embryogenesis. We then integrated this data with corresponding transcriptome information and carried out an inter-species comparison, between *Ciona* and *Xenopus,* an African clawed frog, observing evolutionary conservation and divergence of protein and mRNA dynamics during development.

Results and Discussion

Adapting proteomics for the analysis of *Ciona* eggs and embryos

Mass spectrometry-based proteomics is a versatile tool for studying biological systems containing proteins, although new model systems often require method adaptations. Key areas needing adaptation include sample preparation and the reference proteome. Analyzing eggs and early embryos is often challenging due to the high yolk content. For instance, in *Xenopus*, yolk makes up ~90% of egg protein content, limiting proteomics analysis depth (Goldberger 1980). Researchers usually remove yolk through centrifugation after egg or embryo lysis (Wühr et al. 2014; Gupta et al. 2018; Sonnett et al. 2018a; Baxi et al. 2018). However, when we analyzed *Ciona* egg lysates via Coomassie-stained gels, we found no exceptionally dominant protein band (Supplementary Fig. 1a), allowing us to analyze *Ciona* lysates without yolk removal by mass spectrometry. Another concern is a high-quality protein reference database. For widely used models like humans, mice, or yeast, this is derived from the genome. However, the quality of the genome for non-canonical model organisms is often poor and a better reference database can be generated based on mRNA-seq data (Wühr et al. 2014; Evans et al. 2012).

To evaluate the quality of the *Ciona* genome for proteomics analysis, we assembled a reference proteome using publicly available RNA-seq datasets (Supplementary Fig. 1b) (Wühr et al. 2014). When using this RNA-seq based reference database, it clearly outperformed the Uniprot and KY genome (Satou et al. 2019), but only increased peptide coverage by 5% compared to the recent KY21 proteome (Satou et al. 2021) (Supplementary Fig. 1c). We decided to accept the modest decrease in identified peptides for the ease of annotation offered by the genome assembly and proceeded to use the KY21 proteome as a reference for the remainder of this study. Further examination of peptides identified only by our genome-free database reveal mis-annotated gene coding sequences, mispositioned intercistronic regions, and discrepancies in selenoprotein sequences present in the KY21 proteome (Supplementary Fig. 1d,e) (Santesmasses et al. 2017; Tsagkogeorga et al. 2012a; Satou et al. 2006). We believe that our analysis will help improve the accuracy and completeness of the *Ciona* genome annotation.

Together, our analysis reveals that the sequenced *Ciona* genome, combined with the characteristics of its eggs and embryos, is highly suitable for proteomics studies, supporting *Ciona*'s potential as a valuable model system for investigating evolutionarily conserved mechanisms among chordates.

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.04.559613; this version posted October 6, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Absolute abundance measurement in the unfertilized egg

Using the KY21 reference database, we were able to analyze the proteome of *Ciona* eggs and embryos. First, we estimated the absolute abundance of the proteins in the *Ciona* egg. Maternally deposited proteins in vertebrate eggs play a crucial role in fertilization and shape embryonic development (Wan et al. 2008; Messerschmidt et al. 2012; Bultman et al. 2006). These proteins include a range of components, including metabolic factors, transcription factors (TFs), and signaling molecules (SMs), acting as organizers and guides to ensure the proper formation of body axes and structures. We quantified 6,102 genes, after collapsing protein isoforms (Fig. 1a, Supplementary Table 1); as expected, the most abundant protein is vitellogenin (yolk protein) followed by ATP synthase subunits, actin, and a 60S ribosome subunit (Nomura et al. 2009b). Overall, our analysis spans ~8 orders of magnitude covering 95 TFs and 46 SMs (Fig. 1b,c). The median protein concentration is 22 nM (Fig. 1c), while the median concentrations of TFs and SMs is comparatively lower, at 5.4 nM and 3.5 nM, respectively. Among the TFs identified in the egg, are known maternal factors Gata.a, Prd-B/Prdtun2, and Zinc Finger (C2H2)-33 (Imai et al. 2004a). Among the SMs are known maternal factors β-catenin, Eph.a/Eph1, Eph.b/Eph2, Raf/Raf1, and Tll/Tolloid, Notch and Numb (Imai et al. 2004b; Ahn and Kim 2012; Walton et al. 2006). Finally, we investigated whether different subunits within the same protein complex were found at expected stoichiometric ratios. To this end, we mapped the proteins identified in the egg to known stable complexes from the CORUM database (Fig. 1d) (Tsitsiridis et al. 2022). While proteomics is comparatively weak in estimating absolute concentrations, the similar value derived from subunits of the same protein complexes suggests that our study provides reliable information for absolute protein abundances in the *Ciona* egg, which can act as a valuable resource of maternally deposited proteins for *Ciona* researchers.

A high-quality multi-omics atlas of *Ciona* development

Next, we measured the change of protein and mRNA abundances while the egg developed into a swimming larva. To this end, we combined accurate multiplexed proteome analysis (TMTproC) (Johnson et al. 2021b), with RNA-seq on matching samples at eight key time points spanning early embryonic development, including the maternal/zygotic transition, gastrulation, neurulation, tail elongation, and hatching of swimming larvae (Supplementary Table 2). These proteomics datasets contain 7,095 proteins (with a median of 5 peptides per protein) (Fig. 2a,b). The quantified proteins cover 38% of protein-coding genes annotated in the latest *Ciona* genome and account for approximately 50% of the expressed genes detected in RNA-seq analysis (Satou et al. 2022). This is a substantial increase compared with the only previously published proteome of *Ciona* embryogenesis which identified 695 proteins over three sampled stages using two-dimensional gel electrophoresis and MALDI-TOF/MS (Nomura et al. 2009a).

To assess the relationship between protein and transcript dynamics during development, we identified 6,998 corresponding genes from a combined transcript and protein quantification (Fig. 2b, Supplementary Table 3). We found that the overall Pearson correlation between mRNA and corresponding protein dynamics was very poor (median r = 0.05), consistent with previous reports (Fig. 2c) (Peshkin et al. 2015b; Smits et al. 2014b; Wegler et al. 2019; Edfors et al.