Investigating the molecular guts of endoderm formation using zebrafish

Daniela M. Figiel, Randa Elsayed and Andrew C. Nelson

Abstract

The vertebrate endoderm makes major contributions to the respiratory and gastrointestinal tracts and all associated organs. Zebrafish and humans share a high degree of genetic homology and strikingly similar endodermal organ systems. Combined with a multitude of experimental advantages, zebrafish are an attractive model organism to study endoderm development and disease. Recent functional genomics studies have shed considerable light on the gene regulatory programs governing early zebrafish endoderm development, while advances in biological and technological approaches stand to further revolutionize our ability to investigate endoderm formation, function and disease. Here, we discuss the present understanding of endoderm specification in zebrafish compared to other vertebrates, how current and emerging methods will allow refined and enhanced analysis of endoderm formation, and how integration with human data will allow modeling of the link between non-coding sequence variants and human disease.

Key words: zebrafish; endoderm; development; chromatin; genomics

Introduction

A central question of biology is how a single fertilized egg can give rise to the many distinct cell types within a fully developed organism. By investigating this, we gain valuable knowledge of core principles governing metazoan development, and key distinctions that have driven speciation. In understanding how normal development occurs in model organisms, we gain a valuable blueprint allowing investigation of how genetic influences drive human developmental disorders.

Multiple classical model species have made invaluable contributions to our understanding of cell fate and organisational patterning, such as the fly Drosophila melanogaster, nematode Caenorhabditis elegans, rodents Mus musculus and Rattus norvegicus and frogs Xenopus laevis and Xenopus tropicalis. Each of these organisms gained popularity through their unique practical attributes and experimental tractability, each offering different advantages and drawbacks. Zebrafish is an increasingly attractive model organism for studying vertebrate development and disease. Among its attributes are development external to the mother, large clutch sizes providing an abundance of embryonic material, optical transparency for ease of imaging, and also increasing genetic tractability. The provision of a fully-sequenced and well-annotated genome has allowed zebrafish to transition from an outstanding organism for studying vertebrate gene function [1, 2] to an ideal model organism for studying gene regulatory processes underlying cell fate decisions during embryogenesis [3]. Strikingly, 70% of human genes and 82% of known disease-related genes have a recognizable zebrafish homolog [4]. Zebrafish is consequently an increasingly popular model organism for modeling human genetic disease. Some other model organisms share key advantages with zebrafish, such as ex utero development and ease of microinjection in Xenopus. Others offer specific advantages over zebrafish, such as mice for which many relevant cell lines can be easily cultured and manipulated. Notably, both mouse and Xenopus are also

Daniela M. Figiel holds MSc degrees from the University of Nottingham and University of Warwick, UK and is undertaking PhD studies through the Medical Research Council Doctoral Training Partnership in Interdisciplinary Biomedical Research at Warwick Medical School.

Randa Elsayed holds MSc degrees from the King’s College London and University of Warwick, UK and is undertaking PhD studies through the Medical Research Council Doctoral Training Partnership in Interdisciplinary Biomedical Research at Warwick Medical School.

Andrew C. Nelson is an associate professor at University of Warwick, UK. The Nelson lab studies chromatin-level control of cell fate and behavior with a particular focus on endoderm formation in zebrafish, and trophoblast differentiation during murine placenta development.

© The Author(s) 2021. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
more similar to humans in morphological and evolutionary terms. Nevertheless, the unique combination of attributes zebrafish possess demands their use in studies designed to best exploit their advantages.

The endoderm is the innermost germ layer of the embryo proper, and makes major contributions to the respiratory and gastrointestinal tract and all associated organs including liver, pancreas, thyroid [5] and the pituitary gland [6]. Given the prevalence of developmental and metabolic disorders afflicting these organs, both the study of endoderm formation and function in zebrafish has clear relevance [7]. Here, we discuss use of these organs, both the study of endoderm formation and function in zebrafish, similarities and evolutionary differences in vertebrate endoderm specification, and the future prospects for the application of zebrafish and functional genomics to advance our understanding of endoderm formation and human disease.

Part 1: endoderm specification—the more things change, the more they stay the same

The Nodal signaling pathway is necessary and sufficient to induce endoderm and mesoderm fates in all vertebrates, as discussed extensively elsewhere [8–11]. The process leading to endoderm specification therefore begins with maternally contributed factors that control Nodal induction and ends with stable expression of the master regulator of zebrafish endoderm fate—sox32. In this section, we discuss the current knowledge of transcriptional control of Nodal induction through to stable specification of endoderm.

Transcriptional control of Nodal induction

Endoderm specification occurs at the onset of gastrulation, predominantly in the first two cell tiers above the blastoderm margin [12, 13]. In zebrafish, there are three Nodal-related genes compared to one in mice and humans, two of which (ndr1/sqt and ndr2/cyc) participate in mesendoderm induction. Binding of these TGFβ family ligands to their cognate heterodimeric transmembrane receptors leads to intracellular phosphorylation and release of receptor-tethered effectors Smad2/3. Smad2/3 then associate with Smad4 and translocate into the nucleus where they associate with a range of other transcription factors at Nodal-responsive cis-regulatory modules (CRMs) to induce endoderm and mesoderm fates [8]. While maternal ndr1 mRNA is detectable within eggs and its localization predicts the embryonic dorsal axis by the four-cell stage, it is translationally repressed (reviewed [10, 14]) and zygotically expressed ndr1/2 are critical for endoderm induction.

Though Nodal signaling is key to endoderm specification, the upstream mechanisms that induce Nodal have undergone a degree of diversification during vertebrate evolution. A notable distinction between zebrafish and amniotes is the role of the yolk syncytial layer (YSL)—a teleost-specific structure suggested to be equivalent to the mouse primitive endoderm, which is the initial source of the endoderm-inducing Nodal activity. The YSL forms from collapse of marginal blastomeres into the yolk cell between the 9th and 10th cell cleavages, thus forming a syncytium that underlies the future endoderm (reviewed in [15]). Induction of ndr1/2 expression at the blastoderm margin is controlled by multiple maternally supplied factors including β-catenin and transcription factors Nanog and Eomesoderm homolog A (Eomesa). Dorsal expression of ndr1/2 and thus dorsal organizer formation is controlled by β-catenin [16–18], though a direct requirement for β-catenin in zebrafish endoderm formation is yet to be established as we discuss later. Broader expression of ndr1/2 is controlled during blastula stages by Nanog and Eomesa, which directly activate expression of mxtx2—a key regulator of YSL formation (Figure 1A) [19–22]. Eomesa and Mtxx2 themselves directly bind ndr1/2 to induce their expression [20, 22]. Of particular note is that our knowledge of the Eomesa/Nanog/Mtxx2 regulatory arc upstream of ndr1/2 was gained by exploiting the ability to express affinity-tagged transcription factors for chromatin immunoprecipitation with sequencing (ChIP-seq) analysis during early embryogenesis through microinjection of encoding mRNA, thus mitigating the paucity of commercial antibodies recognizing zebrafish proteins. Exploiting the ex utero development of zebrafish, fecundity and ease of microinjection to generate abundant embryonic material for ChIP-seq analysis during early development is virtually unparalleled and a key benefit of zebrafish.

Nanog is most notable for its role in mammalian embryonic stem cells, where it works alongside Pou5f1/Oct4, Sox2 and others to orchestrate the transcriptional network maintaining pluripotency [23]. In human embryonic stem cells (hESCs), differentiation towards definitive endoderm is initiated by positive regulation of EOMES expression by NANOG [24]. In zebrafish, Nanog functions alongside and directly regulates the closest fish homologs of Pou5f1 and Sox2 to control zygotic genome activation, while Nanog and Pou5f3 (the closest zebrafish homolog of Pou5f1) also control later aspects of endoderm specification, as we discuss later. Consequently, the role of Nanog is much broader than in endoderm induction alone. Strikingly, though Nanog has roles in YSL formation and in the equivalent mouse tissue, mechanistically these roles appear divergent since Nanog appears to non-autonomously control mouse primitive endoderm formation [25, 26] while zebrafish YSL formation is cell autonomously controlled via Mtxx2, which lacks a close mammalian homolog [20]. Similarly, Eomes has pivotal roles in endoderm formation in mice including in collaboration with Nanog, but acts later in the epiblast rather than having maternal functions in endoderm induction [27]. Consequently, initial activation of zygotic Nanog expression in zebrafish appears to be through divergent mechanisms driven by key conserved factors, rather than highly conserved mechanisms.

From Nodal induction to sox32 expression

Ndr1/2 are transcriptional targets of Nodal, therefore, initial expression in the YSL spreads to overlying marginal blastomeres forming a spatiotemporal gradient of nuclear Smad2/3 in the five tiers overlying the YSL [28, 29]. Interestingly, Nodal signals from the YSL also appear to regulate endoderm specification in marginal blastomeres by driving asymmetric nuclear localization independent of Smad2/3-mediated transcription via JNK, and promoting Smad2 nuclear translocation [30]. Both ChIP-seq and ChiP-chip approaches have been successfully applied to explore Smad2 targets at blastula and early gastrula stages, respectively [31, 32]. Strikingly, comparison of Smad2 ChiP-seq targets in zebrafish with data from Xenopus embryos and differentiated human and mouse embryonic stem cells suggests broad conservation of Nodal target genes in early embryogenesis, validating use of zebrafish to explore core principles of vertebrate germ layer formation [32–35].

Smad2/3 genomic target site selection is in part dictated through physical interaction with sequence-specific transcription factors. Among the core Smad2/3 interacting factors across vertebrate species are FoxH1 and Eomes. There are two
Figure 1. Expression and regulation of key genes during zebrafish endoderm specification. (A) Schematic of ndr1/2 expression domains and their key regulators in late blastula stage embryos (labeled in purple and blue). The yolk cell is shown in yellow and ndr1/2 negative domain in green. Dorsal is to the right. (B) RNAseq gene expression profiles of maternally contributed transcripts, ndr1/2 and their upstream activators implicated in endoderm formation between fertilization and the end of gastrulation using data from White et al. [142]. Timepoints corresponding to the onset and complete zygotic genome activation are indicated. (C) As panel B indicating the expression of transcription factors downstream of Nodal implicated in endoderm specification. (D) Schematic indicating the known physical interactions between transcription factors implicated in zebrafish endoderm specification as discussed in the text. All factors present in panels B and C are shown. (E) Schematic indicating the likely regulatory relationships between factors present in panels B and C, based on ChIP-seq and/or genetic evidence from the cited literature.
paralogous Eomes genes in zebrafish owing to duplication of the ancestral gene. Eomesb is not expressed during early development and does not appear to participate in endoderm specification [32]. However, both Eomesa and Foxh1 are maternally contributed in zebrafish (Figure 1B) [36, 37]. ChIP-seq analyses at blastula stages strongly suggest occupancy of Eomesa at Smad2-bound sites is a hallmark of Nodal-responsive genes, highlighting a key parallel with human endoderm [24, 32]. Moreover, these sites in zebrafish are enriched for the known Foxh1 binding motif, and a more recent study exploiting microinjection of foxh1-flag mRNA followed by ChIP-seq further confirmed the co-occurrence of Smad2 and Foxh1 at Nodal responsive sites [28], as observed in cell culture experiments [35].

At the onset of zygotic genome activation, Smad2/3-containing complexes initiate expression of various transcription factor genes (Figure 1C), the products of which form complexes with Smad2/3 as part of a feedforward mechanism. Notable examples include Gata6 which has been shown to interact with Smad2 in zebrafish [31], Mixer/Mixl1 in Xenopus [38], while other factors such as Gata5 are likely to co-occupy complexes with Smad2 given they bind Smad2-interacting factors Eomesa and Mixl1 (Figure 1D) [39]. Each of Gata6, Mixl1 and Gata5 have individually been shown to influence zebrafish endoderm master regulator sox32 and other upstream factors (Figure 1D) [39–43]. However, gata6 expression is not completely Nodal-dependent since, though expression in marginal blastomeres is lost in Nodal mutants, YSL expression is not [44].

While mixl1 (also known as bon in zebrafish) has a conserved role in endoderm specification in vertebrates [45], in zebrafish another Mix family member sebox (also known as mezzo and og9x) acts in parallel with mixl1. Sebox is also a target of Nodal signaling and exhibits partial redundancy with mixl1, however, it also appears to function in mesoderm induction [46]. Though an identifiable sebox homolog exists in mammals and Xenopus, function during endoderm specification seems to be unique to fish [45, 47].

In zebrafish, Foxh1 and Eomesa are individually dispensable for endoderm formation, with eomesa mutants exhibiting only minor loss of early endoderm marker expression [48, 49]. However, Foxh1 and Eomesa are collectively required for endoderm formation, evidenced both through expression of a dominant negative Eomesa in maternal zygotic (MZ) foxh1 mutants, and Foxh1 knockdown in MZeomesa mutants [32, 49]. This is in stark contrast to mammals where Eomes is individually required for definitive endoderm formation [24, 27]. However, recent analyses have revealed the likely reason for the different severities of Eomes mutant endoderm defects between zebrafish and mouse. Eomes is a member of the T-box family of transcription factors, which has undergone changes in complement in different lineages during vertebrate evolution [50]. Notably, this includes tbx16 which is present in fish, frogs (where it is known as veg), birds, marsupials and monotremes but lost in placental mammals [51]. Functional genomics analyses of Eomesa and Tbx16 in zebrafish blastulas and gastrulas, respectively, suggests similar genomic target site selection, including at key regulator of endoderm formation mixl1 [52, 52]. Expression of several mouse Eomes isoforms in zebrafish through mRNA microinjection also suggests functional conservation of Eomes across vertebrate evolution, while knockdown of Tbx16 in eomesa mutants leads to profound loss of mixl1 [53]. The different severities of endoderm phenotypes observed between mouse and zebrafish may therefore be due to T-box factor functional redundancy in zebrafish rather than differences in Eomes molecular function. Given the loss of tbx16 expression seen in foxh1 mutants, it also seems likely that the loss of endoderm observed on abrogation of both Eomesa and Foxh1 function is partially attributable to loss of tbx16 [49]. The consistencies between zebrafish endoderm formation with that of other vertebrates, and the conservation of the Nodal transcriptional network therefore suggest that zebrafish is an ideal model organism to advance our knowledge of endoderm formation.

As implied above and summarized in Table 1, while a similar complement of transcription factors appears to control endoderm specification in all vertebrates, there are some differences in either the temporal expression of these factors or the alteration of the factors involved within gene families. For example, while functionally important Eomesa is maternally contributed in zebrafish, in other vertebrates it is zygotically expressed [27, 48, 54]. Similarly, recent evidence suggests that in Xenopus maternally contributed Otx1, Veg and Foxh1 orchestrate endoderm formation through coordinated binding to selected regulatory elements [55]. In contrast, in fish, the sebox orthologue tbx16 is zygotically expressed in response to Nodal, and there is no evidence for otx1 involvement in endoderm specification, though otx2b has been implicated [40]. Interestingly, in mice, Otx1 and Otx2 exhibit a degree of functional equivalence in neuroectoderm formation [56]. While not directly tested, it therefore seems likely that otx1 and otx2 orthologues may have similar functional potential with respect to endoderm formation, and that these evolutionary substitutions are effectively neutral. Furthermore, two-hybrid screening has suggested physical interaction between mammalian OTX2 and MIXL1 [57], raising the possibility that involvement of otx2b and mixl1 in zebrafish endoderm formation may be conserved mechanism through direct physical association at target loci.

A notable difference between teleost and mammalian endoderm formation is the role of the zinc finger transcription factor Osr1, which is induced by Nodal [32] and limits endoderm differentiation from germ ring mesendoderm in zebrafish [58], though a similar role during mammalian endoderm induction has not been revealed.

On sox32 and the establishment of endoderm fate

Integrated analyses of ChIP-seq and expression data in the early embryo, as well as classical analyses have revealed a complex gene regulatory network controlling sox32 expression (discussed in [52]). Though sox32 is a fish-specific member of the SOX subgroup of SOX family transcription factors, it appears to have arisen from a small-scale tandem duplication of an ancestral gene hence it is adjacent to sox17 in the zebrafish genome [59]. Sox17 is a critical regulator of endoderm fate in other vertebrates, and capable of inducing endoderm fate on overexpression in embryonic stem cells [60, 61]. However, in zebrafish, sox32 appears to act upstream of sox17, and sox32 expression is sufficient to drive endoderm fate [62–64]. Furthermore, ChIP-seq analyses suggest the regulatory inputs into zebrafish sox32 reflect those of human SOX17, evidenced by proximal binding of Eomes and Smad2 in both species [24, 52, 65]. Sox32 therefore appears to have adopted the location of ancestral Sox17 in the regulatory hierarchy. Importantly, while loss of Sox17 in mice leads to substantial reduction of endoderm [66], Sox17 depletion in zebrafish leads to laterality defects but not dramatic loss of endoderm [67]. Conversely, Sox17 overexpression, while rescuing knockdown-induced laterality defects did not lead to increased pancreatic endoderm [67].
### Table 1. Transcription factors implicated in zebrafish endoderm specification

<table>
<thead>
<tr>
<th>Zebrafish factor</th>
<th>Role</th>
<th>Maternal expression</th>
<th>Early zygotic expression</th>
<th>ChIP-chip or ChIP-seq in zebrafish</th>
<th>Implicated in endoderm formation in:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Xenopus</strong></td>
</tr>
<tr>
<td>Nanog</td>
<td>Activates mxtx2 expression to induce YSL formation and Nodal induction</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>NA</td>
</tr>
<tr>
<td>Pouf3</td>
<td>Interacts with Sox32 and required for Sox32-mediated endoderm induction</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>NA</td>
</tr>
<tr>
<td>Mxtx2</td>
<td>Induces YSL formation and directly activates ndr1/2</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>NA</td>
</tr>
<tr>
<td>Smad2</td>
<td>Main effector transcription factor activated by Nodal to induce Nodal target genes</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Foxh1</td>
<td>Constitutive Smad2 interactor redundantly required for endoderm gene expression with Eomesa</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Eomesa</td>
<td>Activates mxtx2 expression and interacts with Smad2, Mixl1 and Gata5 to drive endoderm gene expression</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Osr1</td>
<td>Negative feedback regulator of Nodal-induced endoderm development</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Tbxta</td>
<td>Partially redundant with Tbx16 in regulating mixl1 expression and thus endoderm induction</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Tbx16</td>
<td>Directly activates mixl1 redundantly with Tbtxa and Eomesa to promote endoderm fate</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Otx2b</td>
<td>Regulator of transcription factor expression downstream of Nodal in endoderm specification</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Gata5</td>
<td>Interacts with Mixl1 and Eomesa to drive sox32 expression</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Gata6</td>
<td>Regulator of transcription factors downstream of Nodal in endoderm specification</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Mixl1</td>
<td>Interacts with Gata5, Smad2 and Eomesa to drive sox32 expression. Regulates other genes implicated in endoderm formation including dusp4.</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Sebox</td>
<td>Nodal induced regulator of sox32, partially redundant with Mixl1</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Sox32</td>
<td>Master regulator necessary and sufficient for zebrafish endoderm fate</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>NA</td>
</tr>
</tbody>
</table>

**Note:** other transcription factors implicated in endoderm formation are expressed in endoderm progenitors during gastrulation, but do not have identified/characterized phenotypes prior to segmentation stages: e.g. foxa2, foxa3, foxd3, etv5a, cdx4, grhl3, her5, lmo4a, rara, rarab, mmx2b, phlb, tp53. NA—No clear homolog is present.

<sup>a</sup>Otx1 is implicated in endoderm formation in Xenopus.

<sup>b</sup>Otx2 appears to act in the visceral endoderm in mouse rather than definitive.

---

How Sox32 acts to specify endoderm is not understood, however, there are key indicators suggesting significant parallels with mammalian and Xenopus Sox17. A key feature of Sox family transcription factors is conservation of a short peptide (EF/EQYL) in the C-terminal transactivation domain which interacts with β-catenin [64, 68]. Indeed, a recent ChIP-seq study showed Xenopus Sox17 co-occupies Wnt-responsive endoderm CRMs with β-catenin [69]. Notably, though zebrafish sox17 shows markedly higher conservation with Sox17 orthologues compared to sox32, the β-catenin interacting peptide shows greater conservation in sox32 than sox17 [64, 68]. Moreover, deletion of this peptide abrogates Sox32 ability to induce ectopic endoderm [68, 70]. However, though β-catenin is required to induce dorsal ndr1/2 as discussed above, no evidence for β-catenin requirement in conjunction with Sox32 has yet been presented. Zebrafish have two β-catenin genes—ctnnb1 and ctnnb2, exhibiting considerable co-expression and sharing >92% sequence identity. It is therefore possible they have redundant function and perturbation of both may be needed to reveal a requirement in Sox32-mediated endoderm induction.

Sox32 also physically interacts with Pou5f3, the closest zebrafish homolog of mammalian Pou5f1/Oct4 (Figure 1D) [70, 71]. Quantitative imaging approaches have also revealed that Sox32 can associate with mouse Oct4 [72], mirroring interaction between Sox17 and Oct4 observed in mammals [73]. Maternal Pou5f3 is required for endoderm formation, and Sox32...
overexpression in maternal zygotic Pou5f3 (MZpog) mutants fails to result in endoderm induction suggesting Sox32-Pou5f3 interactions are critical to endoderm fate [74, 75]. Furthermore, Pou5f1 mRNA microinjection can rescue the loss of endoderm in MZpog mutants [76], though Pou5f3 cannot rescue mouse Pou5f1 mutant embryonic stem cell self-renewal [77]. Taken together, our molecular and functional knowledge of zebrafish sox32 and sox17 strongly suggest that in teleosts sox32 is the de facto equivalent of Sox17 in other vertebrates.

An outstanding question is precisely how Pou5f3 mediates Sox32 function and what the chromatin targets of Sox32–Pou5f3 complexes are likely to be. While key clues may be present in existing Pou5f3 ChIP-seq datasets, it is notable that widespread changes are observed in Pou5f3 genomic binding between 512 cell and late blastula stages [78]. Pou5f3 target site selection during endoderm formation is likely to be dependent on Sox32, which will only be present in a minority of Pou5f3+ cells contributing to existing ChIP-seq data [78]. Furthermore, use of heatshock-inducible dominant negative Pou5f3 indicates stage-specific pleiotropic effects on mesendoderm formation [79]. That Pou5f3 target genes appear to be timepoint- and cell type-specific means effectively exploring Pou5f3 targets during endoderm formation will require further functional genomics analysis in a pure endoderm population. However, this should be achievable by exploiting the potency of Sox32-mediated endoderm induction in zebrafish. Thus, co-injection of mRNAs encoding affinity tagged Sox32 and Pou5f3 would allow cell type-specific ChIP-seq interrogation of target genes during endoderm formation.

It is notable that Sox32 can also physically interact with Nanog in vivo, disrupting Nanog–Pou5f3 interactions (Figure 1D) [72]. So far the functional importance of this has been considered in terms of spatially restricting Nanog–Pou5f3 complexes to the ventrolateral mesendoderm, rather than whether Sox32–Nanog complexes have direct functional importance [72]. Analysis of Mnanog and M2nanog mutants by Veil et al. [80] revealed that endoderm marker expression is substantially reduced during gastrulation compared to wild type animals. However, Veil et al. also confirmed the loss of mtx2 previously observed in morpholino knockdown experiments, therefore, it is currently unclear whether reduction in endoderm in M2nanog mutants is solely attributable to loss of upstream regulators of Nodal, or a combinatorial role with Sox32 [20, 80]. A recent study has revealed a role for Nanog in attenuating β-catenin-mediated transcriptional programs by competing for binding of β-catenin co-factor TCFs [81]. It is tempting to speculate that the role of Sox32–Nanog complex formation is to similarly attenuate Sox32–β-catenin interactions or Sox32–Pou5f3 interactions, allowing Nanog to both induce endoderm via the Mtx2/Nodal pathway and restrict endoderm formation by disrupting Sox32-containing complexes. Such a model, however, requires further investigation including through mapping of Sox32–Nanog target sites, which could be achieved as proposed for Pou5f3 above.

In summary, while many of the transcription factors involved in endoderm formation in zebrafish and mammals are the same, there are nevertheless evolutionary changes in the complement of factors involved or the spatiotemporal expression patterns of common factors. The majority of these changes, however, appear to be largely neutral—either representing functional substitution or additional levels of redundancy. This is supported by studies exploiting functional genomics analyses and serves to highlight that evolutionary differences in the transcriptional programs leading to endoderm specification essentially cement the status quo.

Part 2: key challenges and future prospects for studying zebrafish endoderm formation

One of the major challenges of functional genomic and transcriptomic analyses of endoderm formation is that the cells of interest constitute a minor proportion of the embryo. As an indicative estimate, in recent single-cell RNA-seq analyses of the early embryo 3±1.4% of cells at the six developmental stages profiled between specification at ~6 h post-fertilization (h.p.f.) and the onset of organogenesis at 24 h.p.f. were annotated as endoderm [82]. Consequently, endoderm–specific signals are likely to be masked by ensemble averaging across total cell populations in whole embryo analyses typically used to identify functionally active CRMs, such as ChIP-seq analyses of histone modifications or ATAC-seq [83]. As such, though published datasets have been invaluable resources for enhancer analysis during zebrafish embryogenesis [84], there is a clear need for enrichment for cells of interest to further probe the cis-regulatory landscape of endoderm formation. Similarly, while transcription factor ChIP-seq has been invaluable in uncovering the gene regulatory relationships underpinning endoderm specification, the majority of factors discussed above have additional expression domains beyond endoderm. Consequently, it is challenging to definitively ascribe signals to individual cell populations, while key endoderm-specific signals may be absent or below threshold.

Multiple possibilities exist to enrich for endoderm based on expression of specific markers such as sox17. The GFP reporter Tg(sox17:GFP) is a transgenic line [85, 86] widely used for imaging endoderm in a range of contexts, as well as enrichment by fluorescence activated cell sorting (FACS). Notable examples include FACS enrichment followed by expression analyses to probe mechanisms controlling endoderm migration and proliferation during gastrulation [87, 88], and at later stages to explore the role of prostaglandin E2 in liver versus pancreas specification [89], and sphingosine-1-phosphate/Yap1 signaling in endoderm cell survival and interactions with adjacent cardiac mesoderm [90]. This evidences that a more detailed characterization of endoderm gene expression and the cis-regulatory landscape is achievable using sorted populations combined with approaches such as single-cell RNA-seq and ATAC-seq [91–95]. This would offer great advantages, both in terms of providing a more comprehensive characterization of the endoderm [82, 96, 97], and by allowing exploration of how and when perturbation of known regulators of endoderm fate effect the chromatin landscape and gene expression.

While more complete exploitation of the Tg(sox17:GFP) line has the potential to revolutionize investigation of endoderm formation in zebrafish, it also has drawbacks. Since sox17 is a downstream target of Sox32, reporter expression is not observed prior to endoderm specification. An alternative reporter is therefore required to enrich for presumptive endoderm for analysis of endoderm specification itself. Furthermore, though GFP expression in the endoderm persists in Tg(sox17:GFP) fish beyond gastrulation (when endogenous sox17 is silenced), during segmentation stages sox17 is appreciably expressed in mesoderm lineages including endothelial cells and hematopoietic precursors [98]. However, all such issues are surmountable using alternative transgenic lines. For example, Tg(seox:GFP) marks the mesendoderm and has been used for imaging studies as early as 4 h.p.f. [99, 100]. It therefore offers a promising option for cell enrichment for detailed study of chromatin-level regulation of endoderm specification. Alternatively, for analysis of endoderm formation during late segmentation stages and beyond,
Part 3: exploiting zebrafish to model human endoderm disorders and gene regulation

Not only is zebrafish an attractive model for studying fundamental aspects of vertebrate endoderm development, but also for modeling human genetic diseases [114, 115]. Furthermore, it is increasingly realized that many Mendelian and complex genetic disorders are driven by mutations in CRMs such as enhancers, rather than merely in protein coding sequence [116]. Understanding the extent to which cis-regulatory logic is conserved between zebrafish and human will inform what aspects of human genetic disease biology may be reliably modeled in zebrafish, and analysis of how mutations in CRMs influence phenotype. There is consequently a clear imperative both to relate CRMs identified in zebrafish to human genomic data, and to study consequences of human disease-associated mutations in a tractable experimental system such as zebrafish.

While the difficulty of studying minor cell populations can be overcome through enrichment procedures such as FACS and Biotagging, approaches combining the advantages of multiple model systems offer valuable alternatives. Though one of the strengths of zebrafish is the ease of transgenesis, a relative weakness is our inability to expand and manipulate pluripotent cells in vitro. Where the primary motivation is to interrogate broader vertebrate or human biology, it is possible to exploit directed differentiation of hESCs to produce abundant endodermal populations for conventional functional genomics analysis. Since functional analysis of the identified CRMs cannot be performed in vivo in human embryos for technical and ethical reasons, transgenic reporter assays in zebrafish are a powerful alternative. Great progress has been made towards refining the efficient and representative application of transgenic reporter assays, such as development of the Tol2 system, targeted integration methods and insulination of transgenes [117–121]. The wealth of zebrafish fluorescent reporter lines also allows precision analysis of reporter expression patterns with reference to known anatomical landmarks (Figure 2).

There are multiple examples of studies where hESCs have been used to produce differentiated endoderm subpopulations combined with functional genomics analysis, e.g. [121–124]. Such studies identified putative enhancers of interest based on ChIP-seq analysis of commonly studied histone modifications such as H3K4me1 and H3K27ac [125, 126]. For example, Loh et al. [123] differentiated hESCs to definitive endoderm, and then to three distinct subpopulations along the anterior–posterior axis of the developing endoderm–anterior foregut (AFG), posterior foregut (PFG) and midgut/hindgut (MHG). Potential use of resulting functional genomics data could involve comparison with multispecies genomic conservation data, cross-referencing with emerging zebrafish functional genomics data, and ultimately functional investigation though transgenic reporter analysis in zebrafish.

While published zebrafish endoderm functional genomics datasets from gastrulation onwards are currently scarce, undoubtedly exploitation of methods discussed above will lead to complementar human and zebrafish functional data for direct comparison. However, analysis of zebrafish-human conservation within CRMs identified in differentiated hESCs can currently easily be performed using multiple databases. One example is ANCORA [127]. ANCORA provides sets of highly conserved non-coding elements (HCNEs) between multiple species pairs and genome assemblies, allowing identification of conserved components of CRMs at a range of size and percentage conservation cut-offs. While the relationship between sequence and functional conservation of CRMs is still emerging, as discussed elsewhere [128, 129], putative CRMs can be prioritized for further study based on human-zebrafish sequence conservation as an indicator of potential conserved function. For example, comparison of H3K27ac ChIP-seq data from Loh et al. [123] with human-zebrafish HCNE data from ANCORA reveals conservations within a range of putative human CRMs including at genes implicated in disease such as FOXA2, BMPR1A and TCF7L2 (Figures 2 and 3). Foxa2 is implicated in liver and pancreas formation in zebrafish [130], and human genetic variants in CRMs proximal to FOXA2 have been shown to affect glucose regulation and risk of type 2 diabetes mellitus (T2D) [131, 132]. TCF7L2 has also been implicated in T2D with a profound effect on disease susceptibility [133]. Analysis of tcf7l2 mutant zebrafish recapitulates aspects of the human pathology indicating a highly useful model for studying the pleiotropic effects of TCF7L2 loss-of-function [134]. Other genes associated with human disease with conserved CRMs, such as BMPR1A which is implicated in polyposis syndromes of the gut, are yet to be completely studied in zebrafish in the context of disease...
Investigating the molecular guts using zebrafish reporter assays. Genomic regions of interest can be identified by functional genomics approaches such as ChIP-seq in human samples either from directed differentiation of embryonic stem cells or primary tissue. Sub-selection can be performed based on attributes such as presence of highly conserved non-coding elements (HCNEs) and cloned into mCherry fluorescent reporter constructs for microinjection into transgenic zebrafish embryos e.g. Tg(sox17:GFP) at the 1-cell stage followed by imaging of both mCherry and GFP throughout development to relate novel reporter expression patterns to the established reporter line. Example epifluorescence images of Tg(sox17:GFP) larvae at 28 and 48 h.p.f. are shown with anatomical structures labeled equivalently to cartoons.

Use of functional genomics data from human and/or zebrafish combined with knowledge of sequence conservation and/or genetic data will allow us to study how CRM mutations influence gene expression and phenotype in zebrafish disease models.

In a recent example of the integration of human genetic and functional genomic data and zebrafish reporter assays, Eufrásio et al. [137] cross-referenced single nucleotide polymorphisms (SNPs) associated with T2D identified in genome-wide association studies with putative pancreatic enhancers identified through transcription factor ChIP-seq in human endocrine pancreas samples. The resulting 10 human enhancers were examined in transient reporter assays in zebrafish, confirming both endocrine pancreatic enhancer activity for six, and revealing the functional consequences of numerous identified SNPs [137]. Notably, refined methods for analysis of wild type and mutant CRMs are being developed such as Q-STARZ, which uses a dual-CRM dual-reporter system combined with targeted integration to allow comparative analysis of CRMs without technical bias [138]. The continued development of zebrafish genetic tools to
study gene and CRM function will therefore greatly facilitate disease modeling and the progression from functional genomics analysis to functional understanding of individual CRMs, and the consequences of variation and mutation to phenotype.

A further notable example of endoderm disease-causing long-range enhancer mutations is pancreatic agenesis, a congenital disease where the pancreas fails to develop during embryogenesis [139]. Linkage and whole-genome sequencing of individuals with isolated pancreas agenesis and no causal protein-coding mutations uncovered six different mutations in a previously uncharacterized putative enhancer region location ∼25 kb downstream of PTF1A (pancreas-specific transcription factor) in 10 families with pancreatic agenesis [140]. A recent unpublished study suggests that this human putative pancreatic enhancer is able to drive expression in acinar, duct and pancreas progenitor cells in zebrafish reporter assays [141]. Strikingly, the same study identified a novel zebrafish ptf1a enhancer that was capable of driving the equivalent expression pattern in reporter assays. CRISPR deletion of this enhancer also led to pancreatic agenesis in zebrafish suggesting a conserved requirement. Importantly, though this zebrafish enhancer was not highly conserved in human, the functionally equivalent human and zebrafish enhancers appeared to contain similar transcription factor binding sites such as for key pancreatic regulators Foxa2 and Pdx1 [141]. This strongly suggests that integrated analysis of human and zebrafish functional genomics data combined with interrogation of disease associated CRMs can reveal and confirm the causal link between enhancer mutations and developmental disorders, independent of a high degree of sequence conservation.

**Conclusions**

Though there are notable differences, current evidence indicates that gene regulatory control of early endoderm formation is largely conserved between zebrafish and other vertebrates. The logical exploitation of transgenic lines combined with emerging technologies will allow more detailed analysis of gene regulatory control of endoderm formation throughout ontogenesis, providing a blueprint for understanding normal development and disease. Integration of functional genomic and sequence data from human and zebrafish studies will allow the further development of disease modeling strategies in zebrafish to understand how mutations and sequence variants in non-coding regions of the genome influence developmental and metabolic defects.

**Key Points**

- The core gene regulatory program controlling endoderm specification is largely conserved between teleosts and other vertebrates
- An expanding toolkit of transgenic lines and methods will allow more detailed analyses on minor endoderm cell populations
- Zebrafish is emerging as an effective organism to study the relationship between non-coding variants/-mutations and human disease

**Funding**

Daniela Figiel and Randa Elsayed are funded by the Medical Research Council Doctoral Training Partnership in
Interdisciplinary Biomedical Research (MR/N014294/1). Andrew Nelson’s research on endoderm formation in zebrafish is funded by a Wellcome Seed Award (210177/Z/18/Z).

References

67. Aamar E, Dawid IB. Sox17 and chordin are required for formation of Kupffer’s vesicle and left-right asymmetry determination in zebrafish. Dev Dyn 2010;239:2980–8.
Investigating the molecular guts using zebrafish


