DOI:
10.1089/zeb.2015.1179

Document Version
Peer reviewed version

DANIO-CODE: Towards an encyclopedia of DNA elements in zebrafish

Tan Haihan[1], Onichtchouk Daria[2], Winata Cecilia[3,4]

[1]Randall Division of Cell and Molecular Biophysics, King's College London, London, United Kingdom.
[2]Developmental Biology, Institute Biology I, Faculty of Biology, Albert-Ludwigs-University Freiburg, Freiburg, Germany
[3]International Institute of Molecular and Cell Biology, Warsaw, Poland
[4]Max Planck Institute for Heart and Lung Research, Bad Nauheim, Germany.

To whom correspondence should be addressed:

To register interest in being involved in the DANIO-CODE consortium, please contact **Fiona Wardle** (fiona.wardle@kcl.ac.uk) or **Ferenc Müller** (f.mueller@bham.ac.uk). Please address article correspondence to **Haihan Tan** (haihan.tan@kcl.ac.uk), **Daria Onichtchouk** (daria.onichtchouk@biologie.uni-freiburg.de), or **Cecilia Winata** (cwinata@iimcb.gov.pl).

Abstract

The zebrafish has emerged as a model organism for genomics studies. The symposium "**Towards an encyclopedia of DNA elements in zebrafish**" held in London in December 2014, was co-organized by **Ferenc Müller** and **Fiona Wardle.** This meeting is a follow-up of a similar previous workshop held two years earlier and represents a push towards the formalization of a community effort to annotate functional elements in the zebrafish genome. The meeting brought together zebrafish researchers, bioinformaticians, as well as members of established consortia, to exchange scientific findings and experience, as well as to discuss the initial steps towards the formation of a DANIO-CODE consortium. Here we provide the latest updates on the current progress of the consortium's efforts, opening up a broad invitation to researchers to join in and contribute to DANIO-CODE.

Introduction

The genomics revolution has made possible rapid advances in genome annotation. Since 2007, the ENCODE project (ENCyclopedia Of DNA Elements) has been charged with the purpose of annotating functional elements in the human genome[1], and has made use of genomics technologies such as next-generation sequencing (NGS) to produce several thousand datasets on genome-wide transcription, epigenetic modifications, and binding profiles of transcription factors and RNA-binding proteins, documented in more than a hundred major publications. The modENCODE project (Model Organism ENCODE) was initiated thereafter with a similar mission in the model organisms *Drosophila melanogaster* and *Caenorhabditis elegans*. A cumulative analysis of nematode worm and fruit fly regulatory genomes was published in 2010 in two integrative publications[2, 3], and more than 40 publications by modENCODE consortium members. These large-scale analyses have deeply challenged our views on genome structure and function, and influenced multiple research directions in modern biology.

Challenges to our better understanding of human genome function include the analysis of dynamic changes in the regulatory landscape during developmental transitions, and within complex tissues of the organism[4]. Genomic features that are conserved across animal phyla can already be gleaned from small invertebrate model organisms, including *C. elegans* and *Drosophila*. However, recent cross-comparative studies of transcription and chromatin structure using ENCODE and modENCODE data[5, 6] have highlighted not only the common features, but also important differences between phyla, for example in the composition and locations of repressive chromatin. Taking this approach a step further, our understanding of the dynamism of the regulatory genome in the context of chromatin structure will greatly profit from investment into functional studies of simpler non-mammalian vertebrate model organisms, that are amenable to experimental manipulation. After mammalian species, the zebrafish has the best annotated genome[7] and is an obvious candidate for additional functional studies.

This proposition is furthered by the fact that zebrafish research has benefited greatly by riding on the wave of genomics technologies. As a model organism, the zebrafish has several unique features that makes it an ideal model for large-scale genomics studies, including its ability to produce large numbers of embryos, its short generation time, and its relatively low maintenance cost. Owing to the integration of RNA-seq data, the genome assembly and annotation of this established model organism has greatly improved over the past five years, since the release of the latest zebrafish gene build. Accordingly, an increasing number of zebrafish labs have taken the genomics high road to study multiple aspects of zebrafish biology, particularly those interested in gene regulation and comparative genomics. Pioneering zebrafish genomics studies utilized chromatin immunoprecipitation coupled to microarrays (ChIP-on-chip)[8-11] and expression microarray analyses[12-16]; these were followed shortly by an exponentially increasing number of zebrafish studies using NGS technologies, including RNA-seq of mRNAs[17-19] and long non-coding RNAs[20], ChIP-seq for chromatin modifications[21, 22], ribosomal profiling[23-25], DNA methylation[26], nucleosome organization[27], and ChIP-seq for sequence-specific transcription factors, such as Nanog and Mxtx2[28], Pou5f3 and SoxB1[29], Eomesa and Smad2[30], and Zic3[31]. Taken together, there is no doubt that the data from these and future studies hold great promise to capture the dynamic aspects of gene regulatory logic in vertebrate development, using zebrafish as a model system.

However, despite its status as one of the most popular model organisms for developmental studies, pharmacological studies, and disease modelling, among others, as well as a continually expanding knowledge base of its mechanisms of gene regulation, there is still no concerted, co-operative effort to functionally annotate the zebrafish genome, which renders it lagging behind the genomes of human, *Drosophila*, and *C. elegans*. With this in mind, workshops had been conducted in previous years to bring together leading scientists in the field of zebrafish genomics, aimed at establishing a zebrafish community effort similar to that of the ENCODE project. The most recent of these events was the symposium "**Towards an encyclopedia of DNA elements in zebrafish**" co-organized by **Ferenc Müller** and **Fiona Wardle**, held in London in December 2014, which was a follow-up of a similar previous workshop[32]. However, in a decisive progression from previous meetings, this edition

finalized a strong push towards the formalization of a consortium structure by the organizers and indeed, the meeting served to nucleate the formation of a **DANIO-CODE consortium**. Here we provide a summary of the symposium's proceedings, and pertinently, provide information on DANIO-CODE's early-stage considerations and efforts so far.

Application of genomics in the study of zebrafish gene regulation

The use of genomics approaches in the study of gene regulation in zebrafish is still in relatively early days as compared to the mammalian system, but notable landmark studies have been published and this field is rapidly expanding. The zebrafish offers a unique system to study developmental gene regulation due to the convenient accessibility of early developmental stages, and several groups have exploited this advantage by applying various genomics methodologies in gene regulatory studies. Notably, ChIP-seq is increasingly being used in the zebrafish system to profile the binding sites of epigenetic marks and transcription factors.

How epigenetics affects gene expression has long intrigued scientists from many different fields. **Brad Cairns** (Howard Hughes Medical Institute, USA) opened the symposium with a talk on the epigenetic landscape of the zebrafish germline, focusing on the relationships between the chromatin landscape of germline cells and genome activation in early embryonic development. Through the profiling of different histone marks by ChIP-seq, his group has shown that a portion of the germline genome is poised for activation at a later stage during development. This poised state is particularly common at developmental gene loci, which show enrichment of both active (H3K4me2/3) and inactive (H3K27me3) marks at their promoters, as well as pronounced DNA hypomethylation. These bivalent chromatin marks are maintained until the mid-blastula transition (MBT), indicating that the chromatin at developmental gene loci reside in an open conformation while transcription is repressed, and suggesting a mechanism for rapid gene activation during zygotic genome activation (ZGA). Furthermore, through profiling of methylation patterns in both maternal and paternal genomes in egg and sperm, his group discovered that the paternal methylation pattern is retained throughout early embryonic development, and becomes a template for the maternal genome to undergo demethylation and re-methylation to match the paternal genome prior to ZGA[26]. The molecular mechanism regulating this demethylation process is currently under investigation.

Daria Onichtchouk (University of Freiburg, Germany) presented a case study of the zygotic transcriptional activators Pou5f3 and SoxB1, which are the homologs of mammalian pluripotency factors. Her group showed that these transcription factors are often found at the centers of active chromatin[29], prompting them to ask whether they are responsible for regulating chromatin accessibility. She then raised important considerations in designing good ChIP-seq studies, as well as the challenges often faced by zebrafish researchers in performing ChIP-seq, which commonly include the lack of consensus in statistical analysis parameters to define true binding sites and control methods for antibody specificity. However, the major challenge that often discourages zebrafish

researchers from harnessing a ChIP-seq approach is the difficulty in finding an antibody which works in zebrafish.

To overcome this hurdle, several labs have turned to tagged proteins in place of endogenous proteins. **Yi Zhou** and his colleagues (Harvard University, USA) injected mRNA encoding Myc-tagged Nanog-like protein into embryos and performed ChIP-seq for the Myc tag. This analysis led to the discovery of *mxtx2* as a direct transcriptional target of Nanog-like, with Mxtx2 in turn responsible for the activation of genes regulating yolk syncytial layer formation[28]. In another case study of his group's work, he showed how a combination of DNAse-seq and ChIP-seq profiling of the H3K4me3 histone mark and the Gata1 transcription factor uncovered the locus control region which regulates the expression of alternative isoforms of the zebrafish globin gene[33]. In another example of the application of ChIP-seq, **Cecilia Winata** (International Institute of Molecular and Cell Biology, Poland) also contributed insights from her study of Zic3 in gastrulation and neural patterning. Her findings revealed a highly dynamic binding pattern of this transcription factor in different cell states and developmental stages[31], which emphasized the importance of considering spatiotemporal context when annotating regulatory elements *in vivo*.

Post-transcriptional control provides an additional layer of gene expression regulation at the level of mRNA. This form of regulation is prevalent in the transcriptionally quiescent state in pre-MBT stages. **Antonio Giraldez** (Yale University, USA) presented several lines of exciting research from his group, focusing on the translational regulation of maternal mRNAs and its relation to MBT. Using ribosomal profiling techniques, they sought ribosomal footprints which matched that of the translational frame, and through this identified true translational events. The application of this technique in early embryos resulted in the identification of over 100 micropeptide genes. Furthermore, his group is also interested in studying the mechanism of maternal mRNA degradation essential for MBT, and in a high-throughput sequence stability assay, they profiled the transcriptome of embryos treated with the transcriptional inhibitor α-amanitin, leading to the discovery of novel destabilizing sequences in maternal mRNAs[23]. On behalf of **Sinnakaruppan Mathavan** (Genome Institute of Singapore, Singapore), **Cecilia Winata** presented research on a similar theme of post-transcriptional control of gene expression, focusing on cytoplasmic polyadenylation as a mechanism of maternal mRNA translational regulation. Using polysome profiling to identify polysome-associated transcripts, this group attempted to define the relationship between polyadenylation status and translation of maternal mRNAs during pre-MBT development. Preliminary analysis suggests the presence of such a correlation and revealed thousands of polysome-associated maternal transcripts during early embryonic development.

In addition to regulation at the level of the mRNA molecule, microRNAs also play a central role in regulating gene expression by mediating transcript degradation. Yet while thousands of miRNAs have been identified in mammals, only a few hundred are known in zebrafish. In order to expand our knowledge of the zebrafish miRNA milieu, the Giraldez group focuses on miRNAs in the early embryo, and the Mathavan group has performed miRNA profiling in multiple adult tissues, which has revealed

a high number of previously unidentified miRNAs. These data sets are part of a growing knowledge base among an increasing number of zebrafish groups which are putting their efforts into the discovery of novel miRNAs, and this area of research promises many exciting findings in the years to come.

Physical interactions between genomic regions have long been recognized as an essential feature of gene regulatory events. **José Luis Gómez-Skarmeta** (Centro Andaluz de Biología del Desarrollo, Spain) shared an elegant study on comparative epigenomics which provided a unique insight into gene regulation from an evolutionary perspective. By comparing epigenomic profiles across distant species such as zebrafish, medaka, and amphioxus, his group identified regulatory regions with highly conserved epigenetic footprints[34]. Furthermore, using the *HoxD4* and *Six* gene loci as illuminating examples, he presented results from the application of circularized chromatin conformation capture (4C) showing that the three-dimensional architectures of these loci are well-conserved across phyla[35, 36]. This suggests that regulatory elements critical for the basic vertebrate body plan may exhibit highly conserved architecture, representing an important evolutionary constraint on gene regulatory networks.

Taken together, we were treated to several delightful perspectives on how established genomics approaches have enhanced classical gene regulatory studies in the zebrafish system. However, the wider world of genomics is a fast-moving one, with technological and methodological innovations being a constant theme, and this meeting also brought technologists into the fold, to share some of the novel techniques that are beginning to grace the laboratories of the zebrafish genomics community.

New technologies in zebrafish genomics

A challenge to successful applications of ChIP-seq is the availability of ChIP-ready antibodies against endogenous proteins of interest, and one of the ways to overcome this is the use of proteins tagged with epitopes such as Myc tags, as **Yi Zhou** had described earlier in the meeting. **Tatjana Sauka-Spengler** (University of Oxford, UK) introduced us to another tagged protein system that has recently been implemented for cell- and tissue-specific ChIP-seq, and for the isolation of cells or organelles from defined tissues - the Avi-BirA/bioChIP system. This technique is based on the ability of the bacterial biotin ligase BirA to biotinylate an Avi tag[37], which can be fused to various proteins, including markers for cellular compartments. Using Avi-tagged chromatin regulators or transcription factors in a binary combination with BirA-expressing transgenic fish lines allows for simple streptavidin-based protein pulldown, which eliminates the need for specific ChIP-quality antibodies, and greatly reduces the amount of material required for ChIP experiments due to the high affinity of streptavidin-biotin binding. Beyond ChIP applications, tissue-specific expression of BirA in transgenic fish (for example in neural crest cells) combined with Avi-tagged marker proteins of defined cellular compartments, such as the cell membrane or nuclear envelope, allows for efficient cell or organelle

sorting, and allows for the effective purification of biological material from specific cell populations. Besides improvements in ChIP methods, **Yi Zhou** and **José Luis Gómez-Skarmeta** also reported technical advances in several newer flavours of NGS-based methods, such as ATAC-seq and chromatin architectural capture techniques (4C and Hi-C), while many speakers also touched upon their groups' successful implementations of TALEN- and CRISPR-Cas9-based genome editing techniques for convenient zebrafish mutagenesis.

Moving on to innovations in zebrafish genome annotation, we heard **Eivind Valen** (University of Bergen, Norway) presenting previous research performed in **Alexander Schier**'s group (Harvard University, USA), focusing on the discovery of novel protein coding transcripts in zebrafish embryos, and the use of this protein annotation data to improve genome annotation. The results of ribosome profiling in early embryos revealed that translation is far more pervasive than anticipated and occurs for many transcripts previously assumed to be non-coding. The resulting improvements in the accuracy of annotations distinguishing between coding and non-coding RNAs may lead to identification of novel proteins, as open reading frames can be extracted from transcripts that had not been annotated as coding. Some of these newly discovered translated transcripts encode short, functional proteins that had been missed out in prior screens, an example being the recent discovery of the functional embryonic signaling molecule Toddler[38], which may be the first of a family of uncharacterized developmental signals.

Keeping along the lines of genome annotation, **John Collins**, from the group of **Derek Stemple**, (Sanger Institute, UK) presented a novel pipeline for the analysis of data obtained by transcript counting. This method utilizes polyA transcript pulldown to enrich for the 3' end of fragmented transcripts, followed by NGS to produce data on transcript counts. The main innovation in this pipeline is the use of unique molecular identifier barcodes to flag PCR duplicates, allowing the removal of NGS reads that are likely PCR duplicates and thus reducing the number of false positives in the final transcript count results[39]. Traditionally, transcript annotation work performed on the zebrafish reference genome has been carried out by RNA-seq. Compared to RNA seq, transcript counting is relatively cheap and simple, and can be used for large samples, although we were reminded that each has their advantages depending on the type of research question asked.

The vast majority of zebrafish genomics projects so far have concentrated on biological material derived from whole embryos, with few utilizing limited material gathered from specific tissues and cell populations. In a refreshing approach, **Steve Harvey** (Sanger Institute, UK) and **Andrea Pauli** (Harvard University, USA) demonstrated quantitative, single cell transcriptomics in describing the transcriptome of different subpopulations of cells in the embryo. In an example of solving old biological questions with new genomics techniques, **Steve Harvey**, reporting work from the group of **Derek Stemple** (Sanger Institute, UK) spoke about the molecular characterization of the embryonic shield, otherwise known as an organizer region, the earliest morphologically defined inducing center critical to proper embryonic patterning in vertebrate embryos. In the zebrafish gastrula, the deep and superficial layers of the shield differ in their inductive properties[40], but the molecular nature of this

difference is not completely understood. He performed the abovementioned transcript counting from small pool of cells and single cells to compare the transcriptomes of the deep and superficial layers of the fish organizer. This approach estimated an average of 123,660 mRNA molecules per cell, and identified genes with regional specific expression differences, with only half of these previously implicated in the embryonic patterning.

A different approach was presented by **Andrea Pauli**, who reported work from the group of **Alexander Schier** in using RNA-seq to spatially map cellular transcriptomes in the embryo at the early gastrula stage. Single cell RNA-seq was used to expression profile isolated cells from different positions in the early gastrula, and an algorithm was simultaneously developed to map these single-cell profiles back onto specific spatial locations in the embryo by harnessing previously reported expression of known genes in zebrafish expression databases[41]. The spatial position of any one cell could therefore be predicted from the levels of expression for 30 known genes, as was validated by transplantation experiments. This exciting new approach allows for the finer cataloguing of cell types, the generation of a spatial expression database, and the development of predictive algorithms for the expression patterns of novel genes without the requirement for exhaustive *in situ* hybridizations. Whilst this approach is limited in scope thus far, the development of similar methods for older, more complex embryos is currently being explored.

Genomic resources for the zebrafish community

With the broad adoption of DANIO-CODE projects in the foreseeable future, the first primary data will be generated rapidly, raising the issue of a centralised repository where the data should be deposited for consortium access. With regards to this, the Zebrafish Information Network (ZFIN) database has provided an online repository of integrated zebrafish genetic, genomic, and developmental information since 1994. Containing about 75,000 gene expression patterns, 82,000 phenotypes, and 7,000 registered researchers, ZFIN links to major genome annotation databases, and also features the data mining tool ZebrafishMine, similar to the modMINE resource of the modENCODE initiative. **Monte Westerfield** (University of Oregon, USA) gave an update on ZFIN's resources and how they can integrate into the DANIO-CODE initiative, explaining for instance that ZFIN will take over zebrafish genome annotation from the Genome Reference Consortium, and offering to host a collective genomic track hub that links into the ZFIN and ZebrafishMine resources. In addition, with additional sources of funding, it would be possible for ZFIN to host supplementary projects such as curating track data associations with ZFIN mutants, transgenes, phenotypes, and disease information, as well as performing centralised DANIO-CODE data analysis. It was most heartening for us to know that the community has an immediate repository where data may be stored, facilitating the rapid roll-out of a pilot initiative.

Another relevant point of discussion about genome-wide resources was brought up by **Shawn Burgess** (National Human Genome Research Institute, USA), whose group has generated a stable NHGRI-1 zebrafish line with a deeply sequenced genome at high coverage. In the current climate of convenient *in vivo* genome engineering and editing, such a well-characterized and readily available genomic background is critically important for genome editing efforts. In particular, a data resource of all GG/GA sites in the NHGRI-1 background - potential target sites for CRISPR reagent design - has been generated, and a genomic track containing this information is now readily available. It is increasingly clear that alongside our genome annotation efforts, analysis of gene function will occur in parallel through the use of mutant and/or transgenic animals. For the moment, established zebrafish lines are freely available through the Zebrafish International Resource Center (ZIRC), but as the DANIO-CODE project begins to be fully implemented, we will start to turn our thoughts towards how genome annotation and mutant/transgenic data may be synchronised.

Lessons learned from previous consortium projects

Having whetted appetites with a brief perspective on recent genomics-focused investigations and current technological trends and resources, the meeting then turned to broad discussions and presentations on the reality of large-scale genome annotation projects, aimed at informing the initial steps in the DANIO-CODE project. Led by investigators experienced in the nuances of consortium-

based projects, such as **Ben Brown** (University of California, Berkeley, USA), **Laura Clarke** (Sanger Institute, UK), **Carsten Daub** (Karolinska Institute, Sweden), **Jen Harrow** (Sanger Institute, UK), **Boris Lenhard** (Imperial College, UK), and **Piero Carninci** (RIKEN, Japan), we enjoyed many interesting discussions about what lessons could be taken from previous consortium efforts like the ENCODE project. From various informative overviews of procedural considerations in ENCODE and modENCODE, we learned that the success of a big-biology project involving global participants relies heavily on well-formed structural foundations, and that by retrospectively looking at the ways in which prior projects had been run, we can borrow from those efforts in constructing our own program. Here we provide an outline of some of these discussions and considerations.

One of the earliest and most important factors to be considered is the delineation of standards across all participating groups. These standards include, but are not limited to, a defined list of core biological assays for data generation, common experimental protocols, standardized quality control measures for all data produced, and a standard set of metadata attributes used to describe experiments and analyses. Thus even before any data generation is undertaken, the establishment of these conditions immediately allows disparate investigators to share a common vocabulary for ease of subsequent data management. These standards also would be applied to already published zebrafish genomics data in retrospect, in order to determine which data sets have already met the minimal requirements for inclusion in the project, and to standardize their data definitions to ones comparable with future data. A point was raised that all relevant experimental procedures could be uploaded onto the ZFIN Protocol hub in an attempt to facilitate experimental standardization, yet the difficulty of performing certain procedures, especially the very intricate ones, invariably requires the experimenter to learn them in person at the originating laboratory, implying that procedural equality may not be simply achieved through written means.

Another consideration is the co-ordination, sharing, and analysis of data that has been produced by individual labs. This is a considerable task and one that is likely to have to be re-visited over time, as was the experience with ENCODE and modENCODE. It was suggested that quality control and primary analyses of data will initially be performed by individual groups, so that there could be rapid experimental validations in a first-pass measure of biological veracity of the data. Upon passing this stage, the data could then be transferred to a centralized data co-ordination center for secondary or meta-analyses combining multiple data sets from different data production groups. In this respect, it may be advantageous to model the analysis pipeline on the ENCODE workflow, in which the pipeline was actively and regularly assessed, particularly in the case of the emergence of novel methods of analysis, to determine if changes in the pipeline were necessary to improve the analysis of all data sets. This raises the issue of how the proposed data co-ordination center should be established. Other consortia like ENCODE and modENCODE, by virtue of having at least one funding source covering the entire project, were able to create a dedicated data co-ordination center staffed with bioinformaticians tasked with defined responsibilities. At the moment, our program effort lacks suitable centralized funding and hence an exclusive agency for data co-ordination remains unfeasible, but this will be something to aim for going forward.

Another issue is how data sharing between groups should be treated in the academic publishing environment. The stance of ENCODE was to enforce a publication embargo until an agreed flagship article was published, which would serve as an initial publicity of the project as a whole. However, we noted that without centralized funding for the program, it may be unreasonable to apply a blanket embargo on the publication of data. Hence, this meeting's participants were in general agreement that individual groups shall publish the primary data that they generate, but that the collated data across groups will be integrated into a broader flagship publication that serves to signpost the consortium's program. With regards to already published data sets, it was suggested that these could be utilized as part of a pilot DANIO-CODE project. With this in mind, it was agreed to establish a public track hub hosted at ZFIN, where current published genomics data from individual labs can be deposited. This will be a convenient, all-in-one resource for the community in data visualization and manipulation. It was envisaged that this resource would grow in the future and that unpublished data could also be hosted and shared within the consortium. However, it was recognized that the curation and description for track annotation and visualization is not trivial, and in the absence of centralized project funding, this will be a significant operational consideration for the future.

To kick start our zebrafish consortium effort, a proposal was made to set up several initial working groups to oversee the crucial initial phases of establishing the project's structural foundations. Investigators who were involved with ENCODE advised us that its original working hierarchy project was almost immediately re-structured upon initiation, and recommended that we be open to flexibility in the working groups and not be constrained by rigid working group structures. Whether we adopt a centralized decision-making process modelled on the ENCODE consortium, or a more working group-centric approach, remains to be seen. Nevertheless, the meeting's participants agreed that a pilot project would inform us of what leadership approach would work for us, and beyond that, what data standards, workflows, and policies will fit into the structure.

DANIO-CODE's growth to date

Following the successful conclusion of the symposium, several working groups were established to provide a foundational basis to the DANIO-CODE consortium. Through discussions at the symposium and a series of further virtual discussions, three levels of activities have been identified in order to facilitate the growth of the project.

At the first stratum of activity, the the creation of a track hub to collate as many published zebrafish genomic and epigenomic data sets as possible is the priority, allowing for immediate community access to already available data that are scattered across publication space. Since this first level of action is geared towards a rapid roll-out of the DANIO-CODE program and immediately workable, the participants have agreed that all data will first be included without strict quality filtering, although there remains a basic requirement of tagging data with universal metadata standards for ease of curation and comparison. We are pleased to note that the track hub has now been generated and is

kindly hosted by the ZFIN team, and that a pilot set of metadata standards have been established that are currently being implemented as published data is uploaded to the hub. This effort relies heavily on data producers uploading their own data, and as such calls will be going out to the community for their assistance in uploading data.

The second level of action will build upon the data co-ordination and collation currently being established; this will involve the filtering and combined analysis of the uploaded, already published data sets to discover high-quality novel biological observations that would have been missed by mere piecemeal data analysis. Further discussions will confirm the directions and aims of this re-analysis, with the overall goal of outputting a collective publication as a bellwether announcement of the DANIO-CODE consortium, much in the spirit of the flagship ENCODE and modENCODE publications.

The third and final level of action will then progress towards the sharing of unpublished data between groups, and the collaborative generation of new data sets together. This will see the maturing of the consortium project into a phase where groups will identify major themes for collaboration and co-ordinate research into well-defined biological problems that zebrafish genomics can make important contributions to, thereby fostering community connectedness on a global scale.

Conclusion

The symposium "**Towards an encyclopedia of DNA elements in zebrafish**" represented a landmark event in the zebrafish community, in which advances in the annotation of the zebrafish genome were shared by experts in the field, and more importantly, in which the touchpaper for the DANIO-CODE consortium was lit. The consensus agreement of the meeting's participants was the first important step towards the formation of a concrete project structure, which will channel and fortify our efforts into advancing the understanding of gene regulatory mechanisms in our model organism of choice, and in vertebrates in general. To date, the DANIO-CODE consortium has already initiated the provision of a one-stop repository of zebrafish genomic data tagged with comparable and universal metadata. In the near future, it is expected that this will progressively develop into a comprehensive database for the collective effort to annotate zebrafish genomic elements, providing an invaluable community resource for investigators, a route through which collaborations can easily occur, and a means of strengthening zebrafish genomics research.

Acknowledgments

The organizers gratefully acknowledge the generous support from ZF-Health, an integrating project of the European Commission, and COST Action BM0804 (EuFishBiomed). The organizers would also like to thank Jana Maier for help with organization of the workshop. CW is supported by the EU FP7 Grant FishMed GA No. 316125, DO is supported by A7-DFG-EXC294.

Disclosure statement

No competing financial interests exist.

1. Consortium EP: An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57-74.
2. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, et al.: Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science. 2010;330:1775-87.
3. Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, et al.: Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science. 2010;330:1787-97.
4. Ecker JR, Bickmore WA, Barroso I, Pritchard JK, Gilad Y, Segal E: Genomics: ENCODE explained. Nature. 2012;489:52-5.
5. Gerstein MB, Rozowsky J, Yan KK, Wang D, Cheng C, Brown JB, et al.: Comparative analysis of the transcriptome across distant species. Nature. 2014;512:445-8.
6. Ho JW, Jung YL, Liu T, Alver BH, Lee S, Ikegami K, et al.: Comparative analysis of metazoan chromatin organization. Nature. 2014;512:449-52.
7. Wilming LG, Gilbert JG, Howe K, Trevanion S, Hubbard T, Harrow JL: The vertebrate genome annotation (Vega) database. Nucleic Acids Res. 2008;36:D753-D60.
8. Vastenhouw NL, Zhang Y, Woods IG, Imam F, Regev A, Liu XS, et al.: Chromatin signature of embryonic pluripotency is established during genome activation. Nature. 2010;464:922-6.
9. Morley RH, Lachani K, Keefe D, Gilchrist MJ, Flicek P, Smith JC, et al.: A gene regulatory network directed by zebrafish No tail accounts for its roles in mesoderm formation. Proc Natl Acad Sci U S A. 2009;106:3829-34.
10. Wardle FC, Odom DT, Bell GW, Yuan B, Danford TW, Wiellette EL, et al.: Zebrafish promoter microarrays identify actively transcribed embryonic genes. Genome Biol. 2006;7:R71. Epub 2006 Aug 4.
11. Lindeman LC, Andersen IS, Reiner AH, Li N, Aanes H, Ostrup O, et al.: Prepatterning of Developmental Gene Expression by Modified Histones before Zygotic Genome Activation. Dev Cell. 2011.
12. Mathavan S, Lee SG, Mak A, Miller LD, Murthy KR, Govindarajan KR, et al.: Transcriptome analysis of zebrafish embryogenesis using microarrays. PLoS Genet. 2005;1:260-76.
13. Giraldez AJ, Mishima Y, Rihel J, Grocock RJ, Van Dongen S, Inoue K, et al.: Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. Science. 2006;312:75-9.
14. Ferg M, Sanges R, Gehrig J, Kiss J, Bauer M, Lovas A, et al.: The TATA-binding protein regulates maternal mRNA degradation and differential zygotic transcription in zebrafish. Embo J. 2007;26:3945-56.
15. Okuda Y, Ogura E, Kondoh H, Kamachi Y: B1 SOX coordinate cell specification with patterning and morphogenesis in the early zebrafish embryo. PLoS Genet. 2010;6:e1000936.
16. Onichtchouk D, Geier F, Polok B, Messerschmidt DM, Mossner R, Wendik B, et al.: Zebrafish Pou5f1-dependent transcriptional networks in temporal control of early development. Mol Syst Biol. 2010;6:354.
17. Aanes H, Winata CL, Lin CH, Chen JP, Srinivasan KG, Lee SG, et al.: Zebrafish mRNA sequencing deciphers novelties in transcriptome dynamics during maternal to zygotic transition. Genome Res. 2011;21:1328-38.
18. Vesterlund L, Jiao H, Unneberg P, Hovatta O, Kere J: The zebrafish transcriptome during early development. BMC Dev Biol. 2011;11:30.

19. Harvey SA, Sealy I, Kettleborough R, Fenyes F, White R, Stemple D, et al.: Identification of the zebrafish maternal and paternal transcriptomes. Development. 2013;140:2703-10.
20. Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, Levin JZ, et al.: Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res. 2012;22:577-91.
21. Aday AW, Zhu LJ, Lakshmanan A, Wang J, Lawson ND: Identification of cis regulatory features in the embryonic zebrafish genome through large-scale profiling of H3K4me1 and H3K4me3 binding sites. Dev Biol. 2011;357:450-62.
22. Bogdanovic O, Fernandez-Minan A, Tena JJ, de la Calle-Mustienes E, Hidalgo C, van Kruysbergen I, et al.: Dynamics of enhancer chromatin signatures mark the transition from pluripotency to cell specification during embryogenesis. Genome Res. 2012;22:2043-53.
23. Lee MT, Bonneau AR, Takacs CM, Bazzini AA, Divito KR, Fleming ES, et al.: Nanog, Pou5f1 and SoxB1 activate zygotic gene expression during the maternal-to-zygotic transition. Nature. 2013.
24. Bazzini AA, Lee MT, Giraldez AJ: Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. Science. 2012;336:233-7.
25. Chew GL, Pauli A, Rinn JL, Regev A, Schier AF, Valen E: Ribosome profiling reveals resemblance between long non-coding RNAs and 5' leaders of coding RNAs. Development. 2013;140:2828-34.
26. Potok ME, Nix DA, Parnell TJ, Cairns BR: Reprogramming the maternal zebrafish genome after fertilization to match the paternal methylation pattern. Cell. 2013;153:759-72.
27. Zhang Y, Vastenhouw NL, Feng J, Fu K, Wang C, Ge Y, et al.: Canonical nucleosome organization at promoters forms during genome activation. Genome Res. 2014;24:260-6.
28. Xu C, Fan ZP, Muller P, Fogley R, Dibiase A, Trompouki E, et al.: Nanog-like Regulates Endoderm Formation through the Mxtx2-Nodal Pathway. Dev Cell. 2012;22:625-38.
29. Leichsenring M, Maes J, Mossner R, Driever W, Onichtchouk D: Pou5f1 transcription factor controls zygotic gene activation in vertebrates. Science. 2013;341:1005-9.
30. Nelson AC, Cutty SJ, Niini M, Stemple DL, Flicek P, Houart C, et al.: Global identification of Smad2 and Eomesodermin targets in zebrafish identifies a conserved transcriptional network in mesendoderm and a novel role for Eomesodermin in repression of ectodermal gene expression. BMC biology. 2014;12:81.
31. Winata CL, Kondrychyn I, Kumar V, Srinivasan KG, Orlov Y, Ravishankar A, et al.: Genome wide analysis reveals Zic3 interaction with distal regulatory elements of stage specific developmental genes in zebrafish. PLoS Genet. 2013;9:e1003852.
32. Tan H, Zsigmond Á: Zebrafish genomics comes of age. Zebrafish. 2013;10:422-4.
33. Ganis JJ, Hsia N, Trompouki E, de Jong JL, DiBiase A, Lambert JS, et al.: Zebrafish globin switching occurs in two developmental stages and is controlled by the LCR. Dev Biol. 2012;366:185-94.
34. Tena JJ, Gonzalez-Aguilera C, Fernandez-Minan A, Vazquez-Marin J, Parra-Acero H, Cross JW, et al.: Comparative epigenomics in distantly related teleost species identifies conserved cis-regulatory nodes active during the vertebrate phylotypic period. Genome Res. 2014;24:1075-85.
35. Gehrke AR, Schneider I, de la Calle-Mustienes E, Tena JJ, Gomez-Marin C, Chandran M, et al.: Deep conservation of wrist and digit enhancers in fish. Proc Natl Acad Sci U S A. 2015;112:803-8.
36. Gomez-Marin C, Tena JJ, Acemel RD, Lopez-Mayorga M, Naranjo S, de la Calle-Mustienes E, et al.: Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders. Proc Natl Acad Sci U S A. 2015;112:7542-7.
37. Kim J, Chu J, Shen X, Wang J, Orkin SH: An extended transcriptional network for pluripotency of embryonic stem cells. Cell. 2008;132:1049-61.
38. Pauli A, Norris ML, Valen E, Chew G-L, Gagnon JA, Zimmerman S, et al.: Toddler: an embryonic signal that promotes cell movement via apelin receptors. Science. 2014;343:1248636.
39. Collins JE, Wali N, Sealy IM, Morris JA, White RJ, Leonard SR, et al.: High-throughput and quantitative genome-wide messenger RNA sequencing for molecular phenotyping. BMC Genomics. 2015;16:578.

40. Saude L, Woolley K, Martin P, Driever W, Stemple DL: Axis-inducing activities and cell fates of the zebrafish organizer. Development. 2000;127:3407-17.
41. Satija R, Farrell JA, Gennert D, Schier AF, Regev A: Spatial reconstruction of single-cell gene expression data. Nature biotechnology. 2015;33:495-502.