PERSPECTIVES

Time for the zebrafish ENCODE

SRIDHAR SIVASUBBU[1]*, CHETANA SACHIDANANDAN[2]* and VINOD SCARIA[3]*

[1]*Genomics and Molecular Medicine,* [2]*Chemical and Systems Biology and* [3]*G. N. Ramachandran Knowledge Centre for Genome Informatics, CSIR Institute of Genomics and Integrative Biology (CSIR-IGIB), Mall Road, New Delhi 110 007, India*

[Sivasubbu S., Sachidanandan C. and Scaria V. 2013 Time for the zebrafish ENCODE. *J. Genet.* **92**, 695–701]

Genomics research in recent years, especially the human ENCODE project, have made great strides in understanding the genomic and epigenomic structure and organization of humans. These advances promise a new era of precision medicine, through a better understanding of the genomic correlates of human physiology and promise to offer precise and personalized preventive and therapeutic options. The translation of genome-scale maps of genomic and epigenomic markers to clinically relevant information and further to medical practice await functional validation of the genomic features identified through these large-scale efforts. Such studies must essentially be done in model systems where it is possible to model physiological and pathological processes and enquire how they could be modulated by genomic elements and epigenomic signatures. The availability of large number of personal genomes and maps of genomic variations at population scale has created an acute necessity for model systems to model phenotypic and molecular effects of variations, especially in regulatory regions. Efforts to create orthologous maps have been underway in other model systems including Caenorhabditis elegans and Drosophila through the modENCODE programe and in Mus musculus through the mouse ENCODE. We propose that the enormous wealth of disease models and excellent tools to engineer genomes in zebrafish could be effectively capitalized towards making it an effective and widely used model system for precision medicine. This would be possible only through a concerted and systematic effort to create orthologous genomic and epigenomic maps for zebrafish. We discuss how the present understanding and genome-scale methodologies available in this model organism could be effectively used towards realizing this goal.

*For correspondence. E-mail: Sridhar Sivasubbu, s.sivasubbu@igib.res.in; Chetana Sachidanandan, chetana@igib.res.in; Vinod Scaria, vinods@igib.res.in.

Recent years have been witness to dramatic improvements in our understanding of the molecular networks governing biological processes in humans and other model organisms. It is not coincidental that during the same time rapid advances have been made in the area of genomics. The previous decade was marked by global sequencing efforts to understand the genome and genomic diversity of various different species. Current advancements in technology, especially breakthroughs in the scale of nucleotide sequencing have spawned ambitious programmes to gain a functional understanding of the genome through probing the functional and regulatory capacity of genomic elements, their organization and complexity. The Encyclopedia of DNA Elements (ENCODE) consortium was established over five years ago to create a comprehensive catalogue of human functional elements in the genome (ENCODE Project Consortium 2004). The recent publication of over 30 papers cataloguing and interpreting the results of this effort is a major landmark in the understanding of genome biology of any organism. The data encompasses over 1500 experiments performed on ~140 cell types by 440 consortium members in more than 30 institutions and marks an unparalleled effort in the post Human Genome landscape (ENCODE Project Consortium 2004; Dunham *et al*. 2012). It has been widely agreed upon that the efficient organizational structure, a well-defined code of conduct and a top–down management approach contributed significantly to meeting the timelines and for ensuring the high quality of data production and analysis (Birney 2012).

The strength of the ENCODE data lies in the selection and restriction of the enquiry to a distinct set of cell lines such that the data from multiple laboratories and approaches could add to, rather than dilute the information gathered. This strategy is especially useful in distilling functional information from the number of studies on 'regulatory' sequences identified using multiple techniques. Transcription regulatory factors (TRFs) play a most important role in controlling

Keywords. DNA elements; regulation; collaboration; zebrafish; genomics.

the transcriptional activity in cells. ChIP-sequencing experiments using antibodies against over 100 transcription factors on a panel of cell lines provides us with a bird's eye view of the dynamic combinations and occupancy of TRFs on the DNA that brings about the tissue specific expression profiles (Gerstein *et al.* 2010). However, this study is limited to a small number of known transcriptional regulators; a more comprehensive hunt for protein-bound regions in the genome using DNAaseI protection as a proxy for occupancy revealed that genomic elements nearly twice the size of the exome may be occupied by regulatory factors expanding the regions of interest vastly (Neph *et al.* 2012). DNAaseI hypersensitivity on the other hand, indicates regions that are accessible or 'unpacked' in the genome. A profiling of such sites in over 100 cell lines and tissues indicates accessible regions both proximal as well as distal to the trancriptional units (John *et al.* 2013). The relationship between accessibility, methylation (Varley *et al.* 2013) and expression is further influenced by modifications of histones, which influence the chromatin conformation (Ferreri *et al.* 2010). The nucleosome modifications and accessibility profiles are simpler to interpret when they are proximal to the transcriptional units. However, it is well accepted now that distal regions, sometimes kilobases or megabases apart can have pivotal influence on gene expression. These distal regions and the regulatory factors bound to them interact physically with the transcriptional start site machinery by bending and folding of the chromatin. However, because of their lack of proximity, it has been difficult to deduce the interactions by studying accessibility patterns or chromatin marks across the genome. Chromatin conformation capture has emerged as a useful tool to map such long range topological interactions and has been applied to three different cell lines in the ENCODE effort to reveal patterns in such interactions (Sanyal *et al.* 2012). All these different strategies for identifying and analysing regulatory sequences offer possibilities for discovering as yet unexplored complexities in the genome.

Another rich harvest of the ENCODE project has been the large transcriptomic data from 15 cell lines with localization information for them (Djebali *et al.* 2012). This study revealed that nearly 75% of the human genome is transcribed and that splicing is a widespread phenomenon, which occur cotranscriptionally. A significant number of the transcripts were nonprotein-coding in nature such as long-noncoding RNAs, microRNAs and processed pseudogenes. Although the human lncRNAs have similar trancriptional and epigenetic signatures as coding transcripts, they are more likely to be localized in the nucleus and less likely to be translated (Banfai *et al.* 2012). The trancriptome data also enabled a systematic annotation of the pseudogenes revealing evidence that some pseudogenes may be reutilized as functioning noncoding RNAs (Pei *et al.* 2012). In view of the hitherto unexpected number of transcribed regions and regions with potential regulatory role that have been identified by the ENCODE project, it is now important to revisit the previously identified polymorphisms, signals from genomewide association studies and genomic variations in populations in this context. Such efforts are beginning to shed light on new areas of interest that may be relevant to diseases and to personalized medicine (Boyle *et al.* 2012).

The enormous amount of data generated through these projects is expected to be the foundation upon which all future efforts to understand gene function, organization and regulation will be built. It is estimated that over 100 manuscripts that use the ENCODE consortium data have already been published. However, the current data set should be used as a reference to begin understanding the significance of the potentially interesting regions in the genome. The success of the ENCODE project in deciphering human biology will depend heavily on experimental dissection of function in each of these DNA elements in biological systems, most importantly in model organisms. Drawing parallels and extrapolating discoveries across species rests on the availability of comparative data sets and studies on model organisms (Ecker *et al.* 2012).

Following the establishment of human ENCODE consortium, similar efforts have been initiated for other model systems. Notable initiatives include the mouse ENCODE project and the modENCODE project which focusses on the *C. elegans* and *Drosophila* genomes (Ecker *et al.* 2012; Roy *et al.* 2010; Gerstein *et al.* 2010). The collaborations are distinct in their organization and funding patterns. The mouse ENCODE project aims to be based on a loosely knit open collaboration model (Stamatoyannopoulos *et al.* 2012), while the modENCODE project is led and organized directly by the National Human Genome Research Institute in the United States. Both initiatives have been successful in capitalizing on protocol and data sharing and setting standards for analysis and communication. This has enhanced the speed of data generation, analysis and discovery, as evidenced by a number of recent publications based on modENCODE initiatives. The modENCODE project has involved transcriptome profiling identifying both novel coding and noncoding transcripts in *Drosophila* (Gerstein *et al.* 2010; Roy *et al.* 2010). Nuclease sensitivity mapping (Henikoff *et al.* 2011) and ChIP sequencing of histone modifications (Kharchenko *et al.* 2011; Riddle *et al.* 2011; Liu *et al.* 2011), have revealed tissue specific and developmental stage specific dynamics in the regulation of gene expression (Roy *et al.* 2010).

Introduced as a model organism for developmental genetics in the 1980s, zebrafish (*Danio rerio*) has contributed heavily towards our understanding of the vertebrate embryonic development (Basu and Sachidanandan 2013). The forward genetic screens designed in the 1980s and 1990s yielded hundreds of mutants with interesting embryonic phenotypes, which upon identification of the mutant locus led to the discovery of novel genes and a compendium of genotype-to-phenotype relationships. Technologies such as transposon-based gene trapping, morpholinos antisense oligonucleotide based-knock-down and zinc finger nucleases and transcription activator like effector nucleases (TALENs)-based targeted mutations have allowed the interrogation of

gene function at a whole organism level. Lately, TALENs have been used to induce homologous recombination in the zebrafish genome inserting, deleting and modifying bases in a targeted fashion.

One of the striking features of zebrafish embryos is their optical transparency, which enables whole animal imaging at a resolution impossible with other vertebrates and fluorescent reporter based transgenic animals have been extensively used to understand real-time events in the embryo. Over the years, the refinement of genetic manipulation tools and growing knowledge of new genes and their role in development has led to the use of zebrafish in modelling human organ systems and diseases (Basu and Sachidanandan 2013). In this context, an effort to understand the regulatory elements in the zebrafish genome would be valuable in extracting functional information from the human genome and applying these insights into modelling disease processes.

However, unfortunately, international collaborative efforts towards creating a comprehensive and comparable encyclopaedia of regulatory elements in the zebrafish genome are still in its infancy. The zebrafish research community is rightly poised to reap the benefits of both the breakthroughs in technology and scale as well as the standardized protocols presently available in the post Human ENCODE environment. The zebrafish reference genome has been available recently (Howe *et al.* 2013) and has been part of the Genomic References Consortium, which provides a much-needed base for starting to understand functional elements in the genome. Multiple strains of zebrafish including the AB, Tübingen and wild-type strains have recently been sequenced, providing a first look at the genomic variability within the organism (Patowary *et al.* 2013). Many other species of *Danio* are presently being sequenced, which will provide a rich dataset to begin to understand genome organization and functional elements within an evolutionary framework.

Some of the earliest attempts to annotate gene expression patterns in zebrafish were two large-scale whole mount RNA *in situ* hybridization screens performed in the 2001 (Kudoh *et al.* 2001) that created a developmental stage-specific expression pattern database for hundreds of genes in zebrafish (www.zfin.org). More recently, a number of datasets annotating the transcriptome of zebrafish at different developmental time-points and various tissues have been made available, revealing a hitherto unknown subset of nonprotein-coding RNAs, including long noncoding RNAs (Aanes *et al.* 2011; Ulitsky *et al.* 2011; Vesterlund *et al.* 2011; Pauli *et al.* 2012; Wei *et al.* 2012; Kaushik *et al.* 2013). Deep sequencing of the genome has been used for detecting variations and mutations within populations. A number of histone modifications have also been mapped to the zebrafish genome, which includes H3K4me1 H3K4me3, H3K36me3, H3K27me3, in addition to ChIP-seq datasets for a small number of transcription factors currently available (Lindeman *et al.* 2010; Vastenhouw *et al.* 2010; Aanes *et al.* 2011; Pauli *et al.* 2012). To overcome the limitation of zebrafish reactive antibodies, systematic efforts to generate antibodies specific to the zebrafish proteins is essential and this gap is already being filled by various commercial companies.

This windfall of genome-scale data has been accompanied by advances in methodology for assaying and quantifying phenotypes at much higher resolutions and scale (Clemons 2004). For example, using transposon-based DNA cassettes it is possible to 'trap' hundreds of enhancers and create reporter-based transgenic lines (Balciunas *et al.* 2004) in addition to generating mutations in protein-coding genes (Sivasubbu *et al.* 2006, 2007). These reporters have been used for tracking expression of protein coding and noncoding genes leading to the identification of regulatory elements with interesting expression paradigms. High resolution live microscopy for tracking cellular (and intracellular) movements has enabled the community to create spatiotemporal maps of zebrafish gene expression (Clark *et al.* 2012). The results of such efforts are being made available in open source portals such as zfishbook, ZTrap and ZeTrap (Kawakami *et al.* 2010; Clark *et al.* 2011; Kondrychyn *et al.* 2011). Very recently, a genomewide effort to identify and create expression profiles of more than 3000 transcription regulatory factor has led to a dataset of RNA *in situ* hybridizations of more than 1500 transcription regulatory factors in the embryo (Armant *et al.* 2013). These spatiotemporal maps for expression of genes would be a key resource towards understanding the dynamics and specificity of gene expression and synthesizing this knowledge with the corresponding changes in the genome landscape would enable us to build a more complete picture of the dynamic nature of gene regulation in the whole organism.

We are well aware that the zebrafish genomics community is small when compared with the respective community of researchers working on human and other model organisms such as mouse and fly. Nevertheless, this small community has been amenable to adapt to the rapidly changing landscape of genome biology and technology. We argue that even such a small community of researchers with an appropriate well planned strategy of operation and an effective organization can make a huge difference to our understanding of the regulatory power of genomic elements in zebrafish and by extension in humans. Building on our previous experience in genomics and other areas of biology, we discuss briefly how a strategy based on open collaboration and open access could potentially be employed for such a goal.

Recent years have seen the explosive growth of large-scale cocreative collaborative efforts on the one end and large manpower intensive analytic and curation projects through massive crowdsourcing on the other; both fuelled by the increased accessibility and connectivity afforded by the Internet (Oprea *et al.* 2009). Studies show that such activities have been able to significantly infuse innovation, perform humongous tasks while cutting costs dramatically in comparison to conventional approaches. Our previous involvement with organizing large-scale focussed analytic tasks as part of the open source drug discovery (OSDD)

initiative, both in terms of manual curation and high-end technical analyses suggests that with streamlined protocols, constant support and monitoring and a stringent quality control would enable large-scale people-intensive genome analysis tasks through crowdsourcing (Singh 2008). The OSDD effort involves a number of researchers at graduate and undergraduate levels led by a team of scientists curating, annotating and analysing the vast amount of data on *Mycobacterium tuberculosis* towards a goal of identifying the best targets for effective and efficient drug-design and also in diverse areas of big-data analysis including building cheminformatics models based on high-throughput screens (Periwal *et al*. 2011; Jamal *et al*. 2012, 2013; Periwal *et al*. 2012). We have also explored the possibility of using Wiki tools for annotating information in standard formats from literature evidence for the zebrafish genome (http://fishwiki.igib.res.in). The annotation involved graduate and undergraduate students performing extensive curation of information from unstructured literature transforming it into structured data for over 600 zebrafish genes for which literature evidence exist. Similar annotation efforts have been tried for bacteriophage identification and also annotation of metagenomic datasets (Hingamp *et al.* 2008; Caruso *et al.* 2009). Collaborative cocreation has become one of the norms in organizing large-scale mutually beneficial resources, with ample examples in areas as diverse as genomics, drug discovery, design, software development and urban cartography. Unlike typical large-scale collaborative projects elsewhere in the field of biological sciences and genomics, which involves a top–down approach, we argue that the bottom–up approach involving people with diverse expertise and large number of young researchers and students would be the most suitable model for creating a zebrafish ENCODE. Such an open model could necessarily harness the full potential of individual independence and motivation for scientific exploration with full utilization of one's scientific skills balanced by rigorous standardized protocols for data sharing and analysis towards creating a comprehensive map of zebrafish regulatory elements. Such an open collaboration approach has the potential to tap into focussed innovations in protocol development and data analysis on a large scale. It has not escaped our notice that such an approach could potentially run into issues with nonstandard protocols for data generation and analysis, the long term vision of such small focussed projects in software development prove otherwise, as it also allows for comparison and development of novel technologies and a faster advancement of knowledge and development. The ready availability of data for cross-comparison, integration and analysis has been one of the driving forces in most of the large-scale collaborative initiatives. In the human genome and model organism genome era, this has been primarily served by nucleotide databases maintained by the National Center for Biotechnology Information (NCBI), European Bioinformatics Institute (EBI) and elsewhere that constituted a networked set of resources, which made data easily and readily available for comparison, and analysis (Hubbard *et al.* 2002). The availability of networked data resources for raw and processed data for postgenomic datasets has been made possible with generous support from funding agencies including the National Institute of Health and the European Union. These agencies support large-scale data sharing resources such as the Short Read Archive (SRA) and the European Nucleotide Archive (ENA) (Leinonen *et al.* 2011; Benson *et al.* 2013). For easy visualization and comparison of postgenomic data and annotations, several resources exist for zebrafish. This includes the Zebrafish Information Network (ZFIN), which hosts information on gene annotations and mutants, Zebrafish GenomeWiki, which has information culled from literature and other primary resources. In addition, genome browsers hosted at Ensembl, UCSC Bioinformatics site and FishMap provide a visual browsing interface to postgenomic datasets on zebrafish (Meli *et al.* 2008).

Organizing a large scale international collaborative effort to fine-annotate the zebrafish genome has no doubt, its own challenges. However, there have been ample examples of large-scale open collaborations between hundreds of laboratories spread across the world, thanks to the world wide web and other tools that enable real time collaboration and networking (Adams 2012). One of the major challenges, which have been brought about by the breakthrough improvements in the throughput and speed of genome-scale data generation has been the challenge of storing, retrieving and analysing datasets from diverse experiments, otherwise commonly termed as the Big Data challenge. Organizations and consortia have been trying to meet this challenge by integrating large-scale datasets with easy user-friendly interfaces and compute back ends for easy dissemination and communication of information, while making it possible for large number of people to actively access and analyse data sets with minimal hurdles. Such resources for large-scale data integration, analysis and visualization have been spearheaded by the Generic Model Organism Database (GMOD) consortium (http://www.gmod.org/). The GMOD tools have been extensively used for the modENCODE projects.

It needs to be emphasised that the standard protocols for analysis and integration of data is one of the cornerstones of a successful effort to integrate data derived from multiple laboratories and investigators, especially in the case of an open collaborative effort. Though well thought-out protocols need to be in place, it is also imperative that this should not be imposed, rather derived and be amenable to evolution from best practices in the field, which are themselves fast evolving. We propose a novel strategy for data sharing integration where the investigators are free to publish using their favourite protocol. However, the integrated data is processed through a standard common pipeline(s) and made available to all participants. To accelerate innovation and evolve the standards with the developments in analysis methods, we propose to bring in confidence tool developers to make available their resources as analytic apps

for the entire dataset, which would also incentivize the developers in exchange of providing a well-annotated compiled dataset that can be readily used to develop and test the tool/algorithm.

The immediate and major outcome of understanding the genome and epigenome organization in zebrafish could be in its serving as the baseline for understanding and modelling the phenotypic and in-depth molecular correlates of human genomic variations. This is all the more relevant at this point of time, where a number of personal human genomes now available (Venter *et al.* 2001; Levy *et al.* 2007; Bentley *et al.* 2008; Wang *et al.* 2008; Wheeler *et al.* 2008; McKernan *et al.* 2009; Pushkarev *et al.* 2009; Drmanac *et al.* 2010; Gonzaga-Jauregui *et al.* 2012; Patowary *et al.* 2012; Salleh *et al.* 2013) and availability of population-scale genome sequences as part of the 1000 Genome Project (Mu *et al.* 2011). Modelling effects of variations in noncoding regions, especially regulatory regions would be a major challenge in the coming years. The availability of highly efficient *in vivo* genome editing tools like TALENs (Bedell *et al.* 2012) provide a unique opportunity to model human variations in zebrafish system, while the zebrafish ENCODE would provide the baseline data to understand the molecular correlates of the genomic variation.

We do appreciate that sustainability of a research programme in the long term essentially depends on the number of motivated young individuals who take up a research problem in the specific area. Understanding the genomic biology of zebrafish such that it enables to efficiently capitalize the advances in human disease biology and genome biology, would only be possible with a successful outreach programme to inform, educate and motivate young researchers. The InSciEdOut programme is an example that pioneers science education at school level and efficiently uses zebrafish as a model system.

Acknowledgements

The author acknowledges Ashok Patowary and Shruti Kapoor for helping in collecting data and formatting; and members of the SSB, CS and VS labs for discussions. The authors acknowledge funding support from Council of Scientific and Industrial Research (CSIR), India (grants BSC0122, MLP1202 and BSC0123).

References

Aanes H., Winata C. L., Lin C. H., Chen J. P., Srinivasan K. G., Lee S. G. *et al.* 2011 Zebrafish mRNA sequencing deciphers novelties in transcriptome dynamics during maternal to zygotic transition. *Genome Res.* **21**, 1328–1338.

Adams J. 2012 Collaborations: the rise of research networks. *Nature* **490**, 335–336.

Armant O., März M., Schmidt R., Ferg M., Diotel N., Ertzer R. *et al.* 2013 Genome-wide, whole mount in situ analysis of transcriptional regulators in zebrafish embryos. *Dev. Biol.* **380**, 351–362.

Balciunas D., Davidson A. E., Sivasubbu S., Hermanson S. B., Welle Z. and Ekker S. C. 2004 Enhancer trapping in zebrafish using the Sleeping Beauty transposon. *BMC Genomics* **5**, 62.

Banfai B., Jia H., Khatun J., Wood E., Risk B., Gundling Jr W. E. *et al.* 2012 Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res.* **22**, 1646–1657.

Basu S. and Sachidanandan C. 2013 Zebrafish: a multifaceted tool for chemical biologists. *Chem. Rev.* **113**, 7952–7980.

Bedell V. M., Wang Y., Campbell J. M., Poshusta T. L., Starker C. G., Krug R. G. 2nd *et al.* 2012 In vivo genome editing using a high-efficiency TALEN system. *Nature* **491**, 114–118.

Benson D. A., Cavanaugh M., Clark K., Karsch-Mizrachi I., Lipman D. J., Ostell J. *et al.* 2013 GenBank. *Nucleic Acids Res.* **41**, D36–D42.

Bentley D. R., Balasubramanian S., Swerdlow H. P., Smith G. P., Milton J., Brown C. G. *et al.* 2008 Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59.

Birney E. 2012 The making of ENCODE: lessons for big-data projects. *Nature* **489**, 49–51.

Boyle A. P., Hong E. L., Hariharan M., Cheng Y., Schaub M. A., Kasowski M. *et al.* 2012 Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22**, 1790–1797.

Caruso S. M., Sandoz J. and Kelsey J. 2009 Non-STEM undergraduates become enthusiastic phage-hunters. *CBE Life Sci. Edu.* **8**, 278–282.

Clark K. J., Balciunas D., Pogoda H. M., Ding Y., Westcot S. E., Bedell V. M. *et al.* 2011 In vivo protein trapping produces a functional expression codex of the vertebrate proteome. *Nat. Methods* **8**, 506–515.

Clark K. J., Argue D. P., Petzold A. M. and Ekker S. C. 2012 zfishbook: connecting you to a world of zebrafish revertible mutants. *Nucleic Acids Res.* **40**, D907–D911.

Clemons P. A. 2004 Complex phenotypic assays in high-throughput screening. *Curr. Opin. Chem. Biol.* **8**, 334–338.

Djebali S., Davis C. A., Merkel A., Dobin A., Lassmann T., Mortazavi A. *et al.* 2012 Landscape of transcription in human cells. *Nature* **489**, 101–108.

Drmanac R., Sparks A. B., Callow M. J., Halpern A. L., Burns N. L., Kermani B. G. *et al.* 2010 Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81.

Dunham I., Kundaje A., Aldred S. F., Collins P. J., Davis C. A., Doyle F. *et al.* 2012 An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74.

Ecker J. R., Bickmore W. A., Barroso I., Pritchard J. K., Gilad Y. and Segal E. 2012 Genomics: ENCODE explained. *Nature* **489**, 52–55.

ENCODE Project Consortium 2004 The ENCODE (ENCyclopedia of dna elements) project. *Science* **306**, 636–640.

Ferreri A. J., Illerhaus G., Zucca E. and Cavalli F. 2010 Flows and flaws in primary central nervous system lymphoma. *Nat. Rev. Clin. Oncol.* **7** (doi:10.1038/nrclinonc.2010.9-c1).

Gerstein M. B., Lu Z. J., Van Nostrand E. L., Cheng C., Arshinoff B. I., Liu T. *et al.* 2010 Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* **330**, 1775–1787.

Gonzaga-Jauregui C., Lupski J. R. and Gibbs R. A. 2012 Human genome sequencing in health and disease. *Annu. Rev. Med.* **63**, 35–61.

Henikoff J. G., Belsky J. A., Krassovsky K., MacAlpine D. M. and Henikoff S. 2011 Epigenome characterization at single base-pair resolution. *Proc. Natl. Acad. Sci. USA* **108**, 18318–18323.

Hingamp P., Brochier C., Talla E., Gautheret D., Thieffry D. and Herrmann C. 2008 Metagenome annotation using

a distributed grid of undergraduate students. *PLoS Biol.* **6**, e296.

Howe K., Clark M. D., Torroja C. F., Torrance J., Berthelot C., Muffato M. *et al.* 2013 The Zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496**, 498–503.

Hubbard T., Barker D., Birney E., Cameron G., Chen Y., Clark L. *et al.* 2002 The Ensembl genome database project. *Nucleic Acids Res.* **30**, 38–41.

Jamal S., Periwal V., Consortium O. and Scaria V. 2012 Computational analysis and predictive modeling of small molecule modulators of microRNA. *J. Cheminform.* **4**, 16.

Jamal S., Periwal V., Consortium O. and Scaria V. 2013 Predictive modeling of anti-malarial molecules inhibiting apicoplast formation. *BMC Bioinformatics* **14**, 55.

John S., Sabo P. J., Canfield T. K., Lee K., Vong S., Weaver M. *et al.* 2013 Genome-scale mapping of DNase I hypersensitivity. *Curr. Protoc. Mol. Biol.* Chapter 27: Unit 21.27. doi: 10.1002/0471142727.mb2127s103.

Kaushik K., Vincent E. L., Shamsudheen K. V., Lalwani M. K., Jalali S., Patowary A. *et. al.* 2013 Dyamic expression of long non-coding RNAs (IncRANs) in adult zebrafish. *PLoS One* (in press).

Kawakami K., Abe G., Asada T., Asakawa K., Fukuda R., Ito A. *et al.* 2010 zTrap: zebrafish gene trap and enhancer trap database. *BMC Dev. Biol.* **10**, 105.

Kharchenko P. V., Alekseyenko A. A., Schwartz Y. B., Minoda A., Riddle N. C., Ernst J. *et al.* 2011 Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. *Nature* **471**, 480–485.

Kondrychyn I., Teh C., Garcia-Lecea M., Guan Y., Kang A. and Korzh V. 2011 Zebrafish Enhancer TRAP transgenic line database ZETRAP 2.0. *Zebrafish* **8**, 181–182.

Kudoh T., Tsang M., Hukriede N. A., Chen X., Dedekian M., Clark C. J. *et al.* 2001 A gene expression screen in zebrafish embryogenesis. *Genome Res.* **11**, 1979–1987.

Leinonen R., Sugawara H. and Shumway M. 2011 The sequence read archive. *Nucleic Acids Res.* **39**, D19–D21.

Levy S., Sutton G., Ng P. C., Feuk L., Halpern A. L., Walenz B. P. *et al.* 2007 The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254.

Lindeman L. C., Reiner A. H., Mathavan S., Aleström P. and Collas P. 2010 Tiling histone H3 lysine 4 and 27 methylation in zebrafish using high-density microarrays. *PLoS One* **5**, e15651.

Liu T., Rechtsteiner A., Egelhofer T. A., Vielle A., Latorre I., Cheung M. S. *et al.* 2011 Broad chromosomal domains of histone modification patterns in *C. elegans*. *Genome Res.* **21**, 227–236.

McKernan K. J., Peckham H. E., Costa G. L., McLaughlin S. F., Fu Y., Tsung E. F. *et al.* 2009 Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* **19**, 1527–1541.

Meli R., Prasad A., Patowary A., Lalwani M. K., Maini J., Sharma M. *et al.* 2008 FishMap: a community resource for zebrafish genomics. *Zebrafish* **5**, 125–130.

Mu X. J., Lu Z. J., Kong Y., Lam H. Y. and Gerstein M. B. 2011 Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. *Nucleic Acids Res.* **39**, 7058–7076.

Neph S., Vierstra J., Stergachis A. B., Reynolds A. P., Haugen E., Vernot B. *et al.* 2012 An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83–90.

Oprea T. I., Bologa C. G., Boyer S., Curpan R. F., Glen R. C., Hopkins A. L. *et al.* 2009 A crowdsourcing evaluation of the NIH chemical probes. *Nat. Chem. Biol.* **5**, 441–447.

Patowary A., Purkanti R., Singh M., Chauhan R. K., Bhartiya D., Dwivedi O. P. *et al.* 2012 Systematic analysis and functional annotation of variations in the genome of an Indian individual. *Hum. Mutat.* **33**, 1133–1140.

Patowary A., Purkanti R., Singh M., Chauhan R., Singh A. R., Swarnkar M. *et al.* 2013 A sequence-based variation map of zebrafish. *Zebrafish* **10**, 15–20.

Pauli A., Valen E., Lin M. F., Garber M., Vastenhouw N. L., Levin J. Z. *et al.* 2012 Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res.* **22**, 577–591.

Pei B., Sisu C., Frankish A., Howald C., Habegger L., Mu X. J. *et al.* 2012 The GENCODE pseudogene resource. *Genome Biol.* **13**, R51.

Periwal V., Rajappan J. K., Consortium O., Jaleel A. U. and Scaria V. 2011 Predictive models for anti-tubercular molecules using machine learning on high-throughput biological screening datasets. *BMC Res. Notes* **4**, 504.

Periwal V., Kishtapuram S., Consortium O. and Scaria V. 2012 Computational models for in-vitro anti-tubercular activity of molecules based on high-throughput chemical biology screening datasets. *BMC Pharmacol.* **12**, 1.

Pushkarev D., Neff N. F. and Quake S. R. 2009 Single-molecule sequencing of an individual human genome. *Nat. Biotechnol.* **27**, 847–850.

Riddle N. C., Minoda A., Kharchenko P. V., Alekseyenko A. A., Schwartz Y. B., Tolstorukov M. Y. *et al.* 2011 Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin. *Genome Res.* **21**, 147–163.

Roy S., Ernst J., Kharchenko P. V., Kheradpour P., Negre N., Eaton M. L. *et al.* 2010 Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* **330**, 1787–1797.

Salleh M. Z., Teh L. K., Lee L. S., Ismet R. I., Patowary A., Joshi K. *et al.* 2013 Systematic pharmacogenomics analysis of a Malay whole genome: proof of concept for personalised medicine. *PLoS One* **8**, e71554.

Sanyal A., Lajoie B. R., Jain G. and Dekker J. 2012 The long-range interaction landscape of gene promoters. *Nature* **489**, 109–113.

Singh S. 2008 India takes an open source approach to drug discovery. *Cell* **133**, 201–203.

Sivasubbu S., Balciunas D., Davidson A. E., Pickart M. A., Hermanson S. B., Wangensteen K. J. *et al.* 2006 Gene-breaking transposon mutagenesis reveals an essential role for histone H2afza in zebrafish larval development. *Mech. Dev.* **123**, 513–529.

Sivasubbu S., Balciunas D., Amsterdam A. and Ekker S. C. 2007 Insertional mutagenesis strategies in zebrafish. *Genome Biol.* **8** Suppl 1, 9.

Stamatoyannopoulos J. A., Snyder M., Hardison R., Ren B., Gingeras T., Gilbert D. M. *et al.* 2012 An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol.* **13**, 418.

Ulitsky I., Shkumatava A., Jan C. H., Sive H. and Bartel D. P. 2011 Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147**, 1537–1550.

Varley K. E., Gertz J., Bowling K. M., Parker S. L., Reddy T. E., Pauli-Behn F. *et al.* 2013 Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res.* **23**, 555–567.

Vastenhouw N. L., Zhang Y., Woods I. G., Imam F., Regev A., Liu X. S. *et al.* 2010 Chromatin signature of embryonic pluripotency is established during genome activation. *Nature* **464**, 922–926.

Venter J. C., Adams M. D., Myers E. W., Li P. W., Mural R. J., Sutton G. G. *et al.* 2001 The sequence of the human genome. *Science* **291**, 1304–1351.

Vesterlund L., Jiao H., Unneberg P., Hovatta O. and Kere J. 2011 The zebrafish transcriptome during early development. *BMC Dev. Biol.* **11**, 30.

Wang J., Wang W., Li R., Li Y., Tian G., Goodman L. *et al.* 2008 The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65.

Wei C., Salichos L., Wittgrove C. M., Rokas A. and Patton J. G. 2012 Transcriptome-wide analysis of small RNA expression in early zebrafish development. *RNA* **18**, 915–929.

Wheeler D. A., Srinivasan M., Egholm M., Shen Y., Chen L., McGuire A. *et al.* 2008 The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876.

Received 23 April 2013, in revised form 22 July 2013; accepted 25 July 2013
Published on the Web: 11 December 2013