Fish genomes : a powerful tool to uncover new functional elements in vertebrates

Stupka, E.

Chapter 1: Introduction

Introduction

Fish as model organisms

Over the last twenty years fish have rapidly emerged as key model organisms utilized in a variety of research fields. This is owing to their position within the vertebrate subphylum, which provides them with a molecular and body make-up that shares many aspects with that of humans, combined with unparalleled capacity to perform genetic screens and visualize phenotypes, especially in the most widely studied fish species, zebrafish. The latter has enjoyed unsurpassed popularity because of its many enticing features as a model organism such as the ease of maintenance, its transparent embryos which allow powerful visualization of phenotypes, the availability of its genome, as well as a large industry which quickly developed around it to serve the needs of biologists [4-5]. Despite that the emergence of zebrafish was more by accident than by design and it is becoming quickly apparent that many other fish species are equally or even more attractive, depending on the biological question at hand [reviewed in 3]. Until recently it would have been a very large endeavour to begin work on a new model organism species, requiring the co-ordinated action of many laboratories. The development of next-generation sequencing technologies, however, makes it feasible to embark on new species, because information on the genomes, transcriptomes and proteomes can be gained with much less effor than in the past. Thus, for example, species such as *Macropodus opercularis* or *Betta splendens* (which have very compact genomes but display complex behaviour), could be investigated with greater ease, thus connecting complex phenotypes to molecular networks. Although initially great emphasis was placed on mouse and

rat as models for human disease, it is now apparent that fish can be just as good (and sometimes better) models for human disease. Zebrafish is now a well-accepted model organism for the study of complex diseases such as cancer [7], and traits such as ageing [8].

Genome sequencing and assembly

Over 40 years ago the first sequencing was achieved using the Sanger method to allow the deciphering of the sequence of a virus in the 1970s, and later allowing cloning and sequencing of human genes in subsequent years. The human genome project spurred further automation of the same process, allowing (over several years and using hundreds of millions of dollars), the sequencing of the human genome by using a BAC cloning approach (in the publicly funded project) as well as a shotgun approach (in the privately funded Celera project) using long (>500bps) high quality sequence reads. A radical step forward introduced in recent years was the development of next-generation sequencing technologies such as those from Roche 454, Illumina Solexa and ABI SOLID, which now allow a single laboratory on a single machine to obtain 300Gbs of sequence in 10 days from shorter lower quality sequence reads (up to 150bps with current Illumina technology). The data produced by this type of sequencers generates new methodological challenges in genome assembly, which, in turn, have recently pushed the development of new algorithms (discussed in depth in chapter 5 and 6).

Fish genomes

The sequencing and assembly of several fish genomes has greatly enhanced the potential of these organisms, both owing to more accurate identification of

important human orthologs and because they have enabled the discovery of other important vertebrate functional elements of the genome, beyond characterized protein-coding genes. The characteristics of fish genomes had been studied in depth long before genome sequencing was even conceivable. Extensive work by R Hinergardner (1-2) based on simple fluorometric methods had provided genome size estimates for over 200 species of fish, both teleosts and non-teleosts, providing an in-depth investigation of genome sizes throughout the evolutionary branches of this very diverse group. His studies were able to show that more evolved, specialized fishes tended to have smaller genome sizes, and that teleosts have smaller genomes than non-teleost fishes. It is based also on these results that a preliminary characterization was made by in the early 1990s by Nobel Laureate Sydney Brenner of the pufferfish genome, showing that it was likely to be one of the most compact model vertebrate genomes which could be studied [9]. Eventually five years after this initial characterization the pufferfish genome was indeed the first fish genome (and second vertebrate genome after the human genome) to be sequenced, assembled and annotated in our lab[10]. This pivotal study was followed by two more fish genomes, a very close relative of Fugu, *Tetraodon nigroviridis* [11], and a freshwater teleost, medaka (*Oryzias latipes*) [12]. With the advent of next-generation sequencing technologies dozens if not hundreds of fish genomes are now either planned for sequencing or being sequenced already.

Comparative Genomics

The ability to obtain fairly complete and accurate genome sequences for several fish species has allowed the emergence of the field of comparative genomics, i.e. the alignment and comparison of genome sequences and genome structure from

different species. The available genomes allowed comparisons on both shorter evolutionary distances (such as 20MYS between Tetraodon and Fugu), intermediate distances (such as 75MYS between Fugu and Medaka, and 100MYS between Zebrafish and Medaka) and long evolutionary distances (such as 450MYS between human and Fugu). It quickly became apparent that comparative genomics in general, and the Fugu genome in particular were a very powerful tool to detect non-genic functional elements in the genome, such as regulatory elements, which were conserved across the vertebrate lineage. This had been shown much earlier on a smaller scale in Sidney Brenner's lab [13], but the availability of full genomes brought the entire field to a new scale [reviewed in 14]. The field spurred the development of many novel bioinformatics tools, approaches and databases which further refined and optimized the basic task of aligning sequences to be able to detect and score conserved non-coding sequences to distinguish significant conservation from background noise. A variety of acronyms were created for various "classes" of conserved elements, based on the bioinformatics pipeline utilized to identify them, such as HCNEs [15] identified by using MegaBLAST between the human and Fugu genomes, and SCEs, identified using a more complex pipeline focused on shuffled elements, discussed in depth in this thesis [16]. On a larger scale the comparison of these genomes shed light on the complexities of genome duplication genome re-arrangements during vertebrate evolution, showing clearly that while large blocks of synteny are common in short distance comparisons such as those between the mouse and human genome, they are few and far apart when comparing fish to human [10-12].

Transcriptomics

While other –omics technologies such as transcriptomics using microarrays, have been pervasive in the study of human disease and in studies utilizing mouse models, these have not yet achieved their full potential in studies using fish. For the past ten years this was mainly due partly to the limited genome assembly and annotation of the zebrafish genome as well as to the scarce investment made by companies to produce accurate and complete microarray platforms for fish species. This initially lead groups to resort to cDNA arrays, such as the one we used in a study presented in this thesis [17], although these clearly suffered from incomplete coverage and technological limitations. Eventually commercial microarrays became available and started being used and a microarray-based study [18] is discussed in depth in this thesis. The advent of next-generation sequencing is completely revolutionizing the field, owing to techniques such as RNA-Seq [19], which remove the requirement of accurate a priori annotation of the transcriptome, and thus open the door to complete and highly quantitative measurement of transcripts in any species, even those for which the genome has not been sequenced. As shown in the last chapter of this thesis, combining next-generation sequencing of genomic DNA and RNA-Seq nowadays allows the genomic and transcriptomic exploration of a species for which no genome-wide information was available, such as the common carp.

Organization of the thesis

The results presented in this thesis are based on several publications in international peer-reviewed scientific journals. Below is an overview of the chapters presented in this thesis and their related publications.

Chapter 2 focuses on genome sequencing and annotation. I was privileged and honoured to be part of the team which published the first fish genome, i.e. the *Fugu rubripes* genome, and thus this chapter presents the results from that pivotal study, of which I lead the annotation effort. The chapter focuses on the main features of the Fugu genome, and the first basic comparative analyses which were conducted between the Fugu genome and the human genome. The results were published in the following paper:

- Aparicio S et al. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* 2002;297(5585):1301-10

Chapter 3 focuses on comparative genomics. While working on the Fugu genome I was intrigued by the fact that gene order between mammals and fish had hardly been retained at all. Knowing that regulatory elements usually have even less constraints on their position and orientation I hypothesized that in order to identify a complete set of vertebrate enhancers one would have to develop a methodology that allows for shuffling during evolution to different genomic locations. Based on this hypothesis we developed a pipeline for the detection of over 20,000 SCEs (shuffled conserved elements), which we showed to be functional enhancers. The results were published in the following paper:

- Sanges R. et al. Shuffling of cis-regulatory elements is a pervasive feature of the vertebrate lineage. *Genome Biology* 2006; 7(7):R56

Chapter 4 focuses on the use of transcriptomics technologies in fish to answer biological questions. We focused on the degradation of maternal RNA, using

microarray-based gene expression profiling, which were published in this paper:

- Ferg M. et al. The TATA-binding protein regulates maternal mRNA degradation and differential zygotic transcription in zebrafish. *EMBO J* 2007; 26(17): 3945-3956

Chapter 5 focuses on the assembly of the carp genome and transcriptome from next-generation sequencing data. This is a manuscript under preparation.

Chapter 6 provides a discussion of the results presented, proposes future directions and conclusions. In this chapter a short summary of thesis in Dutch is also provided.

Bibliography

1. Hinegardner R. Evolution of cellular DNA content in teleostean fishes. *Am Naturalist* 1968;102:517–523.
2. Hinegardner R. The cellular DNA content of sharks, rays and some other fishes. *Comp Biochem Physiol B* 1976;55:367–370.
3. Muller F. Comparative Aspects of Alternative Laboratory Fish Models. *Zebrafish* 2005;2(1):47-54
4. Zebrafish—the canonical vertebrate. *Science* 2001;294:1290–1291.
5. Grunwald DJ, Eisen JS. Headwaters of the zebrafish— emergence of a new model vertebrate. *Nat Rev Genet* 2002;3:717–724.
6. Special issue devoted to Medaka, *Mech Dev* 2004;121: 629–637.
7. Cancer genetics and drug discovery in the zebrafish. *Nat Rev Cancer* 2003;3:533–539
8. Gerhard GS, Cheng KC. A call to fins! Zebrafish as a gerontological model. Aging Cell 2002;1:104–111.45
9. Brenner S, Elgar G, Sandford R, Macrae A, Venkatesh B, Aparicio S Characterization of the pufferfish (Fugu) genome as a compact model vertebrate genome *Nature* 1993; 366:265 - 268
10. Aparicio S et al. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* 2002;297(5585):1301-10
11. Jaillon O. et al. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature* 2004; 431: 946-957
12. Kasahara M. et al. The medaka draft genome and insights into vertebrate genome evolution. Nature 2007; 447:714-719

13. Aparicio S et al. Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, Fugu rubripes. *PNAS 1995*; 92:1684-1688
14. Boffelli D, Nobrega MA, Rubin EM. Comparative genomics at the vertebrate extremes. *Nat Rev Genet* 2004;5:456–465
15. Woolfe A et al. Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development. *PLOS Biology* 2005; 3(1):e7
16. Sanges R. et al. Shuffling of cis-regulatory elements is a pervasive feature of the vertebrate lineage. *Genome Biology* 2006; 7(7):R56
17. Yang Li et al. Comparative analysis of the testis and ovary transcriptomes in zebrafish by combining experimental and computational tools. *Comparative and Functional Genomics* 2004; 5:403-418
18. Ferg M. et al. The TATA-binding protein regulates maternal mRNA degradation and differential zygotic transcription in zebrafish. *EMBO J* 2007; 26(17): 3945-3956
19. Wang Z et al. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009 10(1):57-63
20. Yamamoto Y, Stock DW, Jeffery WR. Hedgehog signaling controls eye degeneration in blind cavefish. *Nature* 2004; 431:844–847
21. Shapiro MD, Marks ME, Peichel CL, Blackman BK, Nereng KS, Jonsson B, Schluter D, Kingsley DM. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature* 2004; 428:717-723