The zebrafish reference genome sequence and its relationship to the human genome

Kerstin Howe[1,*], Matthew D. Clark[1,2,*], Carlos F. Torroja[1,3], James Torrance[1], Camille Berthelot[4,5,6], Matthieu Muffato[7], John E. Collins[1], Sean Humphray[1,8], Karen McLaren[1], Lucy Matthews[1], Stuart McLaren[1], Ian Sealy[1], Mario Caccamo[2], Carol Churcher[1], Carol Scott[1], Jeffrey C. Barrett[1], Romke Koch[9], Gerd-Jörg Rauch[10], Simon White[1], William Chow[1], Britt Kilian[1], Leonor T. Quintais[7], José A. Guerra-Assunção[7], Yi Zhou[11], Yong Gu[1], Jennifer Yen[1], Jan-Hinnerk Vogel[1], Tina Eyre[1], Seth Redmond[1], Ruby Banerjee[1], Jianxiang Chi[1], Beiyuan Fu[1], Elizabeth Langley[1], Sean F. Maguire[1], Gavin K. Laird[1], David Lloyd[1], Emma Kenyon[1], Sarah Donaldson[1], Harminder Sehra[1], Jeff Almeida-King[1], Jane Loveland[1], Stephen Trevanion[1], Matt Jones[1], Mike Quail[1], Dave Willey[1], Adrienne Hunt[1], John Burton[1], Sarah Sims[1], Kirsten McLay[1], Bob Plumb[1], Joy Davis[1], Chris Clee[1], Karen Oliver[1], Richard Clark[1], Clare Riddle[1], David Eliott[1], Glen Threadgold[1], Glenn Harden[1], Darren Ware[1], Beverly Mortimer[1], Giselle Kerry[1], Paul Heath[1], Benjamin Phillimore[1], Alan Tracey[1], Nicole Corby[1], Matthew Dunn[1], Christopher Johnson[1], Jonathan Wood[1], Susan Clark[1], Sarah Pelan[1], Guy Griffiths[1], Michelle Smith[1], Rebecca Glithero[1], Philip Howden[1], Nicholas Barker[1], Christopher Stevens[1], Joanna Harley[1], Karen Holt[1], Georgios Panagiotidis[1], Jamieson Lovell[1], Helen Beasley[1], Carl Henderson[1], Daria Gordon[1], Katherine Auger[1], Deborah Wright[1], Joanna Collins[1], Claire Raisen[1], Lauren Dyer[1], Kenric Leung[1], Lauren Robertson[1], Kirsty Ambridge[1], Daniel Leongamornlert[1], Sarah McGuire[1], Ruth Gilderthorp[1], Coline Griffiths[1], Deepa Manthravadi[1], Sarah Nichol[1], Gary Barker[1], Siobhan Whitehead[1], Michael Kay[1], Jacqueline Brown[1], Clare Murnane[1], Emma Gray[1], Matthew Humphries[1], Neil Sycamore[1], Darren Barker[1], David Saunders[1], Justene Wallis[1], Anne Babbage[1], Sian Hammond[1], Maryam Mashreghi-Mohammadi[1], Lucy Barr[1], Sancha Martin[1], Paul Wray[1], Andrew Ellington[1], Nicholas Matthews[1], Matthew Ellwood[1], Rebecca Woodmansey[1], Graham Clark[1], James Cooper[1], Anthony Tromans[1], Darren Grafham[1], Carl Skuce[1], Richard Pandian[1], Robert Andrews[1], Elliot Harrison[1], Andrew Kimberley[1], Jane

Garnett[1], **Nigel Fosker**[1], **Rebekah Hall**[1], **Patrick Garner**[1], **Daniel Kelly**[1], **Christine Bird**[1], **Sophie Palmer**[1], **Ines Gehring**[10], **Andrea Berger**[10], **Christopher M. Dooley**[1,10], **Zübeyde Ersan-Ürün**[10], **Cigdem Eser**[10], **Horst Geiger**[10], **Maria Geisler**[10], **Lena Karotki**[10], **Anette Kirn**[10], **Judith Konantz**[10], **Martina Konantz**[10], **Martina Oberländer**[10], **Silke Rudolph-Geiger**[10], **Mathias Teucke**[10], **Kazutoyo Osoegawa**[12], **Baoli Zhu**[12], **Amanda Rapp**[13], **Sara Widaa**[1], **Cordelia Langford**[1], **Fengtang Yang**[1], **Nigel P. Carter**[1], **Jennifer Harrow**[1], **Zemin Ning**[1], **Javier Herrero**[7], **Steve M. J. Searle**[1], **Anton Enright**[7], **Robert Geisler**[10,14], **Ronald H. A. Plasterk**[9], **Charles Lee**[15], **Monte Westerfield**[13], **Pieter J. de Jong**[12], **Leonard I. Zon**[11], **John H. Postlethwait**[13], **Christiane Nüsslein-Volhard**[10], **Tim J. P. Hubbard**[1], **Hugues Roest Crollius**[4,5,6], **Jane Rogers**[1,2], and **Derek L. Stemple**[1]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB101SA, UK [2]The Genome Analysis Centre, Norwich Research Park, Norwich NR47UH, UK [3]Bioinformatics Unit, Centro Nacional de Investigaciones Cardiovasculares, 28029 Madrid, Spain [4]Ecole Normale Supérieure, Institut de Biologie de l'ENS, IBENS, 46 rue d'Ulm, Paris F-75005, France [5]INSERM, U1024, 46 rue d'Ulm, Paris F-75005, France [6]CNRS, UMR 8197, 46 rue d'Ulm, Paris F-75005, France [7]EMBL European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK [8]Illumina Cambridge, Chesterford Research Park, Little Chesterford, Saffron Walden CB10 1XL, UK [9]Hubrecht Laboratory, Uppsalalaan 8, 3584 CT Utrecht, The Netherlands [10]Max Planck Institute for Developmental Biology, Spemannstraße 35, 72076 Tübingen, Germany [11]Stem Cell Program and Division of Hematology and Oncology, Children's Hospital and Dana Farber Cancer Institute, 1 Blackfan Circle, Karp 7, Boston, Massachusetts 02115, USA [12]Children's Hospital Oakland, 747 52nd Street, Oakland, California 94609, USA [13]Institute of Neuroscience, University of Oregon, 1254 University of Oregon, 222 Huestis Hall, Eugene, Oregon 97403-1254, USA [14]Karlsruhe Institute of Technology (KIT), Campus North, Institute of Toxicology and Gentics (ITG), Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany [15]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA

Abstract

Zebrafish have become a popular organism for the study of vertebrate gene function[1,2]. The virtually transparent embryos of this species, and the ability to accelerate genetic studies by gene knockdown or overexpression, have led to the widespread use of zebrafish in the detailed investigation of vertebrate gene function and increasingly, the study of human genetic disease[3–5]. However, for effective modelling of human genetic disease it is important to understand the extent to which zebrafish genes and gene structures are related to orthologous human genes. To examine this, we generated a high-quality sequence assembly of the zebrafish genome, made up of an overlapping set of completely sequenced large-insert clones that were ordered and oriented using a high-resolution high-density meiotic map. Detailed automatic and manual annotation provides evidence of more than 26,000 protein-coding genes[6], the largest gene set of any vertebrate so far sequenced. Comparison to the human reference genome shows that approximately 70% of human genes have at least one obvious zebrafish orthologue. In addition, the high quality of this genome assembly provides a clearer understanding of key genomic features such as a unique repeat content, a scarcity of pseudogenes, an enrichment of zebrafish-specific genes on chromosome 4 and chromosomal regions that influence sex determination.

The zebrafish (*Danio rerio*) was first identified as a genetically tractable organism in the 1980s. The systematic application of genetic screens led to the phenotypic characterization of a large collection of mutations[1,2]. These mutations, when driven to homozygosity, can produce defects in a variety of organ systems with pathologies similar to human disease. Such investigations have also contributed notably to our understanding of basic vertebrate

biology and vertebrate development. In addition to enabling the systematic definition of a large range of early developmental phenotypes, screens in zebrafish have contributed more generally to our understanding of the factors controlling the specification of cell types, organ systems and body axes of vertebrates[7–9].

Although its contributions have already been substantial, zebrafish research holds further promise to enhance our understanding of the detailed roles of specific genes in human diseases, both rare and common. Increasingly, zebrafish experiments are included in studies of human genetic disease, often providing independent verification of the activity of a gene implicated in a human disease[3,5,10]. Essential to this enterprise is a high-quality genome sequence and complete annotation of zebrafish protein-coding genes with identification of their human orthologues.

The zebrafish genome-sequencing project was initiated at the Wellcome Trust Sanger Institute in 2001. We chose Tübingen as the zebrafish reference strain as it had been used extensively to identify mutations affecting embryogenesis[2]. Our strategy resembled the clone-by-clone sequencing approach adopted previously for both the human and mouse genome projects. The Zv9 assembly is a hybrid of high-quality finished clone sequence (83%) and whole-genome shotgun(WGS) sequence (17%), with a total size of 1.412 gigabases (Gb) (Table 1). The clone and WGS sequence is tied to a high-resolution, high-densitymeiotic map called the Sanger AB Tübingen map (SATmap), named after the strains of zebrafish used to make the map (Supplementary Information).

Zebrafish are members of the teleostei infraclass, a monophyletic group that is thought to have arisen approximately 340 million years ago from a common ancestor[11]. Compared to other vertebrate species, this ancestor underwent an additional round of whole-genome duplication (WGD) called the teleost-specific genome duplication (TSD)[12]. Gene duplicates that result from this process are called ohnologues (after Susumu Ohno who suggested this mechanism of gene duplication)[13]. Zebrafish possess 26,206 protein-coding genes[6], more than any previously sequenced vertebrate, and they have a higher number of species-specific genes in their genome than do human, mouse or chicken. Some of this increased gene number is likely to be a consequence of the TSD.

A direct comparison of the zebrafish and human protein-coding genes reveals a number of interesting features. First, 71.4% of human genes have at least one zebrafish orthologue, as defined by Ensembl Compara[14] (Table 2). Reciprocally, 69% of zebrafish genes have at least one human orthologue. Among the orthologous genes, 47% of human genes have a one-to-one relationship with a zebrafish orthologue. The second largest orthology class contains human genes that are associated with many zebrafish genes (the 'one-human-to-many-zebrafish' class), with an average of 2.28 zebrafish genes for each human gene, and this probably reflects the TSD. A few notable human genes have no clearly identifiable zebrafish orthologue; for example, the leukaemia inhibitory factor (*LIF*), oncostatin M (*OSM*) or interleukin-6 (*IL6*) genes, although the receptors *lifra, lifrb, osmr* and *il6r* are clearly present in the zebrafish genome. It is possible that zebrafish proteins with functionally similar activities to LIF, OSM and IL-6 exist, but that their sequence divergence is so great that they cannot be recognized as orthologues. Similarly, the zebrafish genome has no *BRCA1* orthologue, but does have an orthologue of the *BRCA1*-associated *BARD1* gene, which encodes an associated and functionally similar protein and a *brca2* gene, which plays an important role in oocyte development, probably reflecting its role in DNA damage repair[15].

Zebrafish have been used successfully to understand the biological activity of genes orthologous to human disease-related genes in greater detail[3–5]. To investigate the number

of potential disease-related genes, we compared the list of human genes possessing at least one zebrafish orthologue with the 3,176 genes bearing morbidity descriptions that are listed in the Online Mendelian Inheritance in Man (OMIM) database. Of these morbid genes, 2,601 (82%) can be related to at least one zebrafish orthologue. A similar comparison identified at least one zebrafish orthologue for 3,075 (76%) of the 4,023 human genes implicated in genome wide association studies (GWAS).

Zv9 shows an overall repeat content of 52.2%, the highest reported so far in a vertebrate. All other sequenced teleost fish exhibit a much lower repeat content, with an average of less than 30%. This result suggests that the evolutionary path leading to the zebrafish experienced an expansion of repeats, possibly facilitated by a population bottleneck. Alternatively, the repeat content of the other sequenced teleost species may be under-represented, as these assemblies are mostly WGS[16].

The majority of transposable elements found in the human genome are type I (retrotransposable elements), with more than 4.3 million placements covering 44% of the sequence, whereas only 11% of the zebrafish genome sequence is covered by type I elements in less than 500,000 instances. In contrast, the zebrafish genome contains a marked excess of type II DNA transposable elements. Indeed, 2.3 million instances of type II DNA transposable elements cover 39% of the zebrafish genome sequence (Supplementary Table 12), whereas type II repeats cover only 3.2% of the human genome.

This pronounced abundance of type II transposable elements is unique among the sequenced vertebrate genomes, and the genome sequence shows evidence of recently active type II transposable elements. The closest vertebrate species in terms of the abundance of type II transposable elements is *Xenopus tropicalis* (25% type II transposable elements), whereas the sequenced and annotated teleost fish (the pufferfish *Takifugu* and *Tetraodon*, the three-spined stickleback *(Gasterosteus aculeatus)* and the medaka (*Oryzias latipes*)) each possess type II transposable element coverage of less than 10%, which may relate to the fact that the zebrafish genome diverges basally from the other sequenced and annotated teleost genomes[17]. Zebrafish type II transposable elements are divided into 14 superfamilies with 401 repeat families in total (Supplementary Table 12). The DNA and hAT superfamilies are the most abundant and diverse in the zebrafish genome, together covering 28% of the sequence. The type II transposable element abundance of zebrafish, or lack of retrotransposable elements, may provide an explanation for the low zebrafish pseudogene content (Supplementary Table 14).

The long arm of chromosome 4 is unique among zebrafish genomic regions, owing to its relative lack of protein-coding genes and its extensive heterochromatin. Chromosome 4 is known to be late-replicating and hybridization studies suggest that genomic copies of 5S ribosomal DNA (rDNA), which are not notably present on any other chromosome, are scattered along the long arm at high redundancy[18]. Immediately after the presumed centromere at approximately 24 megabases (Mb), the sequence landscape (Fig. 1 and Supplementary Fig. A4) shows a remarkable increase in repeat content, which continues through to the telomere of the long arm. At approximately 27 Mb, the otherwise uniform presence of the satellite repeat SAT-2 on the long arm ends abruptly. This location is also the starting point of uniform MOSAT-2 distribution, a satellite repeat that is nearly absent from all other chromosomes but highly enriched on the long arm of chromosome 4. The subtelomeric region of the long arm shows a distinct distribution of repeat elements, with relatively fewer interspersed elements and an increased content of satellite, simple and tandem repeats that do not harbour 5S rDNA sequences. Moreover, the gene content is reduced on the long arm and the guanine–cytosine content is slightly increased.

The long arm of chromosome 4 also has a special structure with respect to gene orthology and synteny. Approximately 80% of the genes present have no identifiable orthologues in human. In fact, 110 genes (out of 663) have no identifiable orthologues in any other sequenced teleost genome and indeed seem to be zebrafish-specific genes. The genes in this region are highly duplicated, with 31 ancestral gene families alone providing 77.5% of the genes, the largest of which contains no less than 109 duplicates in this region. The largest of these families correspond to NOD-like receptor proteins[19] with putative roles in innate immunity and zinc finger proteins. We also observed a very high density of small nuclear RNAs (snRNAs) on chromosome 4, and in particular those that encode spliceosome components. The cohort of snRNAs carried on the long arm of chromosome 4 accounts for 53.2% of all snRNAs in the zebrafish genome. In addition, in a specific group of zebrafish derived recently from a natural population, the subtelomeric region of the long arm of chromosome 4 has been found to contain a major sex determinant with alleles that are 100% predictive of male development and 85% predictive of female development, suggesting that this chromosome may be, might have been, or may be becoming, a sex chromosome in this particular population[20].

In addition to the chromosome 4 sex determinant, three other separate genomic regions have been identified as influencing sex determination, and these vary between the strains and even within the families studied[20,21]. Our meiotic map, SATmap, which was generated to anchor the genomic sequence, provided an opportunity to examine whether there are any strong signals for sex determination. To generate SATmap we took advantage of the fact that it is possible to create double haploid individuals that contain only maternally derived DNA, that are homozygous at every locus and that can be raised until they are fertile[22] (Fig. 2a). To investigate the interesting finding that SATmap F_1 fish could be either male or female while being genetically identical and heterozygous at every polymorphic locus, we sought a genetic signal for sex determination in the F_2 generation, in which these polymorphisms segregate. Using morphological secondary sexual traits, we were able to score the sex of 332 genotyped F_2 individuals. Although most chromosomes showed no significant genetic bias for a particular sex, we found that most of chromosome 16 carried a strong signal ($P = 9.1 \times 10^{-7}$) with a broad peak around the centromere (Fig. 2b, c). Homozygotes for the Tübingen (grandmaternal) allele had a very high probability of being female, whereas homozygotes for the AB (grandpaternal) allele were very unlikely to be female (Fig. 2).

The number of protein-coding genes among vertebrates is relatively stable, although even closely related species may show great disparities in the nature of their protein-coding gene content. We carried out a four-way comparison between the proteome of two mammals (human and mouse), a bird (chicken) and the zebrafish to quantify the fraction of shared and species-specific genes present in each genome (Fig. 3a). A core group of 10,660 genes is found in all four species and probably approximates an essential set of vertebrate protein-coding genes. This number is somewhat less than the core set of 11,809 vertebrate genes identified previously as being common to three fish genomes (*Tetraodon*, medaka, zebrafish) and three amniotes (human, mouse, chicken)[16], but the discrepancy probably reflects the improved annotation of these genomes that often results in fusing fragmented gene structures. Each taxon has between 2,596 and 3,634 species-specific genes. The notable excess observed in zebrafish may be a consequence of the WGD, because pairs of duplicated genes that arose from the WGD, but with no orthologue in amniotes, are counted as two specific genes. Furthermore, 2,059 genes are found in human, mouse and zebrafish but not in chicken, and this number is two times higher than the number of genes that are found in all amniotes but not in zebrafish (892). It is unclear whether these genes have been lost along the evolutionary branch leading to the chicken, or whether this is due to annotation or orthology assignation errors in the chicken genome.

We identified double-conserved synteny (DCS) blocks between all sequenced tetrapods and four fish genomes (zebrafish, medaka, stickleback and *Tetraodon*). DCS blocks are defined as runs of genes in the non-duplicated species that are found on two different chromosomes in the species that underwent aWGD[23], although the genes may not be adjacent in the duplicated species[24]. The DCS between zebrafish and human are represented on either side of each human chromosome (Supplementary Fig. 15). Using DCS blocks, we identified zebrafish paralogous genes that are part of DCS blocks and consistent with the locally alternating chromosomes, hence with an origin at the TSD. We identified 3,440 pairs of such ohnologues (26% of the all genes), for a total of 8,083 genes when subsequent duplications are taken into account. It is notable that although true pairs of ohnologues may exist within the same chromosome owing to post-TSD rearrangements, we excluded such cases as we cannot reliably distinguish them from segmental duplications. This number of ancestral genes retained as duplicates in zebrafish is higher, both in absolute number and in proportion, than in other fish genomes (chi-squared test, all $P<3 \times 10^{-5}$).

We compared the 8,083 zebrafish TSD ohnologues with human ohnologues originating from the two rounds of WGD that are common to all vertebrates and find that the two sets overlap strongly (chi-squared test, $P<2 \times 10^{-16}$). In general, zebrafish ohnologous pairs are enriched in specific functions (neural activity, transcription factors) and are orthologous to mammalian genes under stronger evolutionary constraint than genes that have lost their second copy.

A circular representation of ohnologue pairs (Fig. 3b) highlights chromosomes, or parts of chromosomes, that descended from the same pre-duplication ancestral chromosome (for example, chromosomes 3 and 12, 17 and 20, 16 and 19).Among zebrafish chromosomes, chromosome 16 and chromosome 19 are unique in their one-to-one conservation of synteny. Consistent with the conservation of synteny, chromosome 16 and chromosome 19 possess clusters of orthologues of genes associated with the mammalian major histocompatibility complex (MHC) as well as the *hoxab* and *hoxaa* clusters, respectively, which are each orthologous to the human *HOXA* cluster[25].

Since the earliest whole-genome shotgun-only assembly became public in 2002, the zebrafish reference genome sequence has enabled many new discoveries to be made, in particular the positional cloning of hundreds of genes from mutations affecting embryogenesis, behaviour, physiology, and health and disease. Moreover, the annotated reference genome has enabled the generation of accurate whole-exome enrichment reagents, which are accelerating both positional cloning projects and new genome-wide mutation discovery efforts[26,27].Although the zebrafish reference genome sequencing is complete, a few poorly assembled regions remain, which are being resolved by the Genome Reference Consortium (http://genomereference.org).

METHODS SUMMARY

We generated cloned libraries of large fragments of genomic DNA, assembled a physical map of large-insert clones and completely sequenced a set of minimally overlapping clones. In addition, we generated WGS sequences by end-sequencing a mixture of large- and short-insert libraries. Overlapping clone sequences were combined with WGS sequences and tied to the meiotic map, SATmap, which enabled independent placement and orientation of clones in the genome sequence. The sequence data can be found in the BioProject database, under accession number PRJNA11776.

To obtain evidence for a more complete description of protein-coding genes, we used high-throughput short-read complementary DNA sequencing and obtained a deep-coverage data

set for messenger RNAs expressed in zebrafish at various stages of development and in adult tissues[6]. Finally, a standard Ensembl gene build, incorporating filtered elements from the complementary DNA sequencing gene build, was merged with the manually curated gene models to produce a comprehensive annotation in Ensembl version 67 (http://may2012.archive.ensembl.org/Danio_rerio/Info/Index). Detailed descriptions of all the methods used for this project are available in the Supplementary Information.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

1. Driever W, et al. A genetic screen for mutations affecting embryogenesis in zebrafish. Development. 1996; 123:37–46. [PubMed: 9007227]

2. Haffter P, et al. The identification of genes with unique and essential functions in the development of the zebrafish, *Danio rerio*. Development. 1996; 123:1–36. [PubMed: 9007226]

3. Golzio C, et al. KCTD13 is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant. Nature. 2012; 485:363–367. [PubMed: 22596160]

4. Panizzi JR, et al. CCDC103 mutations cause primary ciliary dyskinesia by disrupting assembly of ciliary dynein arms. Nature Genet. 2012; 44:714–719. [PubMed: 22581229]

5. Roscioli T, et al. Mutations in ISPD cause Walker-Warburg syndrome and defective glycosylation of alpha-dystroglycan. Nature Genet. 2012; 44:581–585. [PubMed: 22522421]

6. Collins JE, White S, Searle SM, Stemple DL. Incorporating RNA-seq data into the zebrafish Ensembl genebuild. Genome Res. 2012; 22:2067–2078. [PubMed: 22798491]

7. Talbot WS, et al. A homeobox gene essential for zebrafish notochord development. Nature. 1995; 378:150–157. [PubMed: 7477317]

8. Gritsman K, et al. The EGF-CFC protein one-eyed pinhead is essential for nodal signaling. Cell. 1999; 97:121–132. [PubMed: 10199408]

9. Ober EA, Verkade H, Field HA, Stainier DY. Mesodermal Wnt2b signalling positively regulates liver specification. Nature. 2006; 442:688–691. [PubMed: 16799568]

10. Tobin DM, et al. Host genotype-specific therapies can optimize the inflammatory response to mycobacterial infections. Cell. 2012; 148:434–446. [PubMed: 22304914]

11. Amores A, Catchen J, Ferrara A, Fontenot Q, Postlethwait JH. Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted gar, an outgroup for the teleost genome duplication. Genetics. 2011; 188:799–808. [PubMed: 21828280]

12. Meyer A, Schartl M. Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. Curr.Opin. Cell Biol. 1999; 11:699–704. [PubMed: 10600714]

13. Wolfe K. Robustness–it's not where you think it is. Nature Genet. 2000; 25:3–4. [PubMed: 10802639]

14. Vilella AJ, et al. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. 2009; 19:327–335. [PubMed: 19029536]

15. Rodríguez-Mari A, et al. Roles of *brca2* (*fancd1*) in oocyte nuclear architecture, gametogenesis, gonad tumors, and genome stability in zebrafish. PLoS Genet. 2011; 7:e1001357. [PubMed: 21483806]

16. Kasahara M, et al. The medaka draft genome and insights into vertebrate genome evolution. Nature. 2007; 447:714–719. [PubMed: 17554307]

17. Postlethwait JH. The zebrafish genome in context: ohnologs gone missing. J. Exp. Zool. B. 2007; 308:563–577.

18. Sola L, Gornung E. Classical and molecular cytogenetics of the zebrafish, *Danio rerio* (Cyprinidae, Cypriniformes): an overview. Genetica. 2001; 111:397–412. [PubMed: 11841183]

19. Stein C, Caccamo M, Laird G, Leptin M. Conservation and divergence of gene families encoding components of innate immune response systems in zebrafish. Genome Biol. 2007; 8:R251. [PubMed: 18039395]

20. Anderson JL, et al. Multiple sex-associated regions and a putative sex chromosome in zebrafish revealed by RAD mapping and population genomics. PLoS ONE. 2012; 7:e40701. [PubMed: 22792396]

21. Bradley KM, et al. An SNP-based linkage map for zebrafish reveals sex determination loci. G3 (Bethesda). 2011; 1:3–9. [PubMed: 21949597]

22. Streisinger G, Walker C, Dower N, Knauber D, Singer F. Production of clones of homozygous diploid zebra fish (*Brachydanio rerio*). Nature. 1981; 291:293–296. [PubMed: 7248006]

23. Kellis M, Birren BW, Lander ES. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature. 2004; 428:617–624. [PubMed: 15004568]

24. Jaillon O, et al. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature. 2004; 431:946–957. [PubMed: 15496914]

25. Amores A, et al. Developmental roles of pufferfish Hox clusters and genome evolution in ray-fin fish. Genome Res. 2004; 14:1–10. [PubMed: 14707165]

26. Kettleborough RNW, et al. A systematic genome-wide analysis of zebrafish protein-coding gene function. Nature. (in the press).

27. Varshney GK, et al. A large-scale zebrafish gene knockout resource for the genome-wide study of gene function. Genome Res. 2013; 23:727–735. [PubMed: 23382537]

28. Freeman JL, et al. Definition of the zebrafish genome using flow cytometry and cytogenetic mapping. BMC Genomics. 2007; 8:195. [PubMed: 17597531]

29. The 1000 Genomes Project Consortium An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491:56–65. [PubMed: 23128226]

30. Krzywinski M, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009; 19:1639–1645. [PubMed: 19541911]
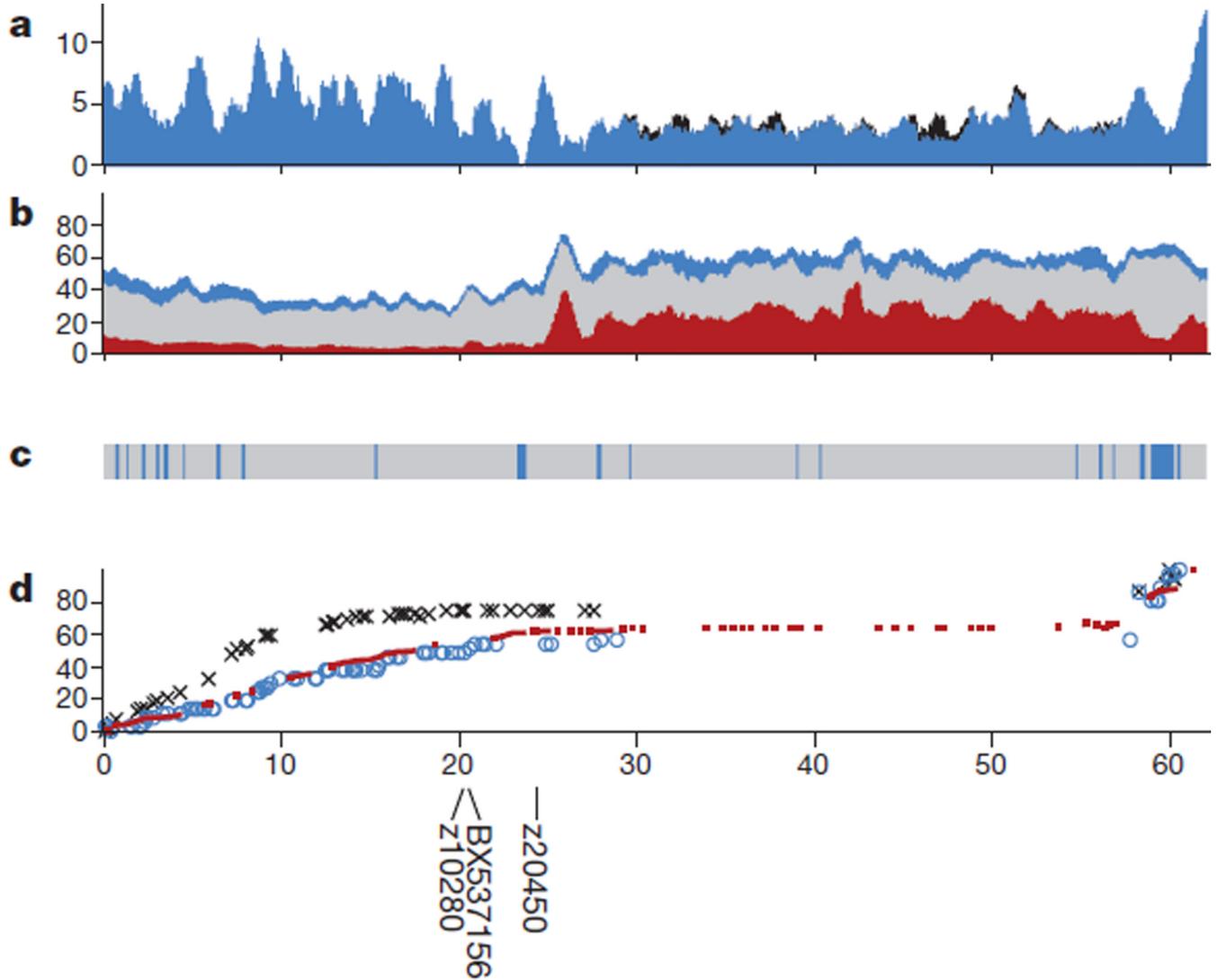
Figure 1. Landscape of chromosome 4

a, Exon coverage (blue), stacked with coverage by snRNA exons (black). **b**, Stacked repeat coverage, divided into type I transposable elements (red), type II transposable elements (grey) and other repeat types (blue), including dust, tandem and satellite repeats. **c**, Sequence composition (grey bars, clones; blue bars, WGS contigs). **d**, Genetic marker placements (red, SATmap markers; blue, heat shock meiotic map markers; black, Massachusetts General Hospital meiotic map markers). Marker placements have been normalized so that the maps can be compared. Near-centromeric clones are positioned at 20Mb (BX537156), 20.2Mb (Z10280) and 24.4Mb (Z20450)[28]. The *x* axis shows the chromosomal position in Mb. **a** and **b** were calculated as percentage coverage over 1-Mb overlapping windows (*y* axis), with a 100-kb shift between each window. **c** and **d** were calculated over 100-kb windows. The *y* axis for **d** shows the normalization of marker positions relative to the span of the individual map. Similar graphs for the other chromosome are provided in the Supplementary Information.
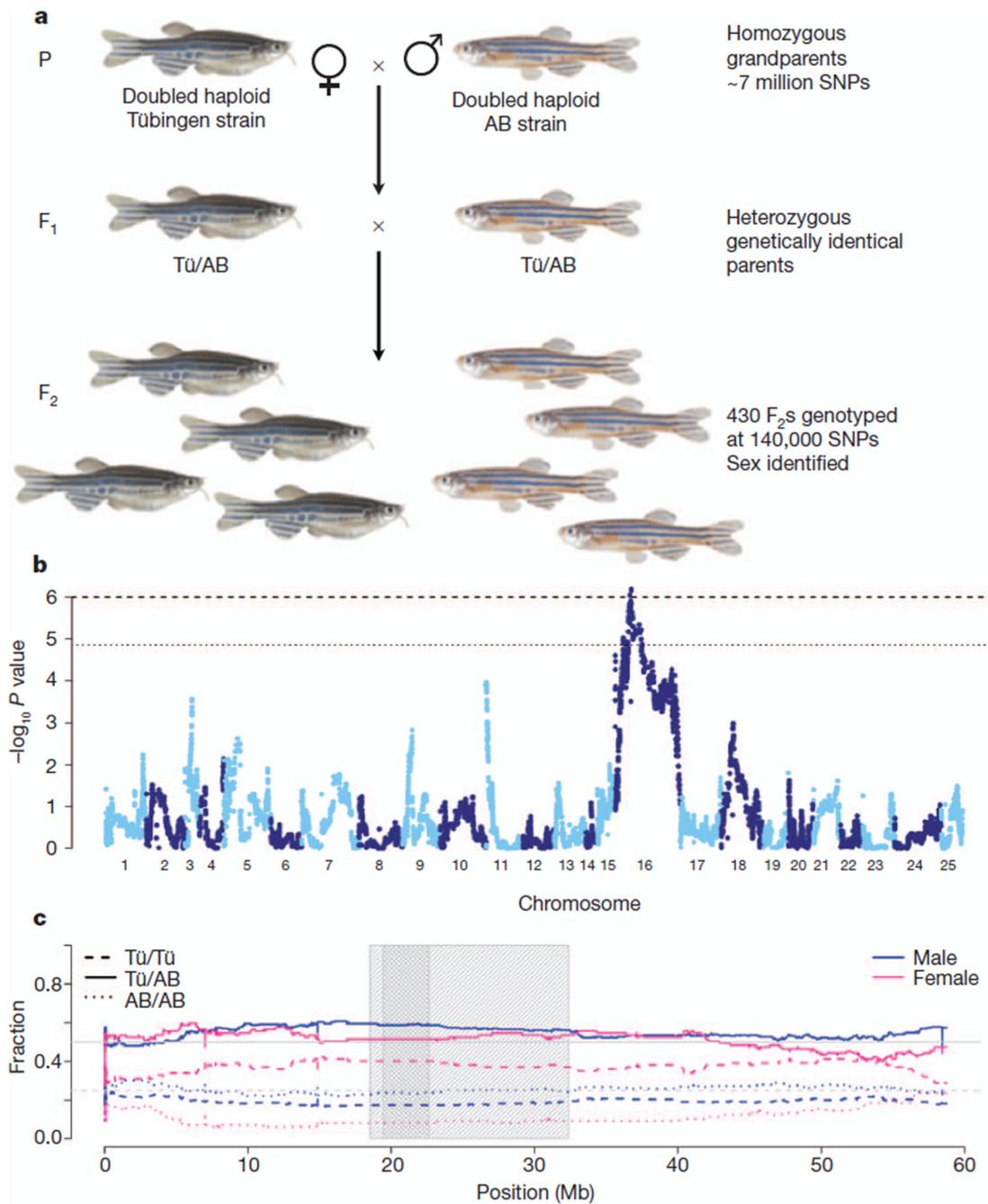
Figure 2. Sex determination signal on chromosome 16

a, Breeding scheme for SATmap. Double haploid generation zero (G_0) founders were sequenced to approximately 40× depth using Illumina GAII technology. We found approximately 7 million SNPs between the two SATmap founders. This number of SNPs between just two homozygous zebrafish individuals is far in excess of that seen between any two humans and is nearly one-fifth of all SNPs measured among 1,092 human diploid genomes[29]. Genetically identical, heterozygous F_1 fish of both sexes resulted from crossing the founders. The F_1 individuals were crossed to generate a panel of F_2 individuals, each with its own unique set of meiotic recombinations between AB and Tübingen (Tü)

chromosomes, which were uncovered by dense genotyping with a set of 140,306 SNPs covering most of the genome. **b**, Genome-wide *P* values for tests of genotype difference between sexes, arranged by chromosome. The dotted line corresponds to differences that are expected once in 100 random genome scans, and the dashed line corresponds to differences expected once in 1,000 random genome scans. The only locus that is statistically significant at these levels is on chromosome 16. **c**, Genotype frequencies for males and females on chromosome 16. The grey line at 0.5 corresponds to expectation for heterozygotes (solid lines) and the grey line at 0.25 corresponds to expectation for homozygotes (dashed and dotted lines). The light grey shaded box corresponds to the region in which empirical *P* < 0.01, the dark grey shaded box corresponds to the region in which *P* < 0.001.
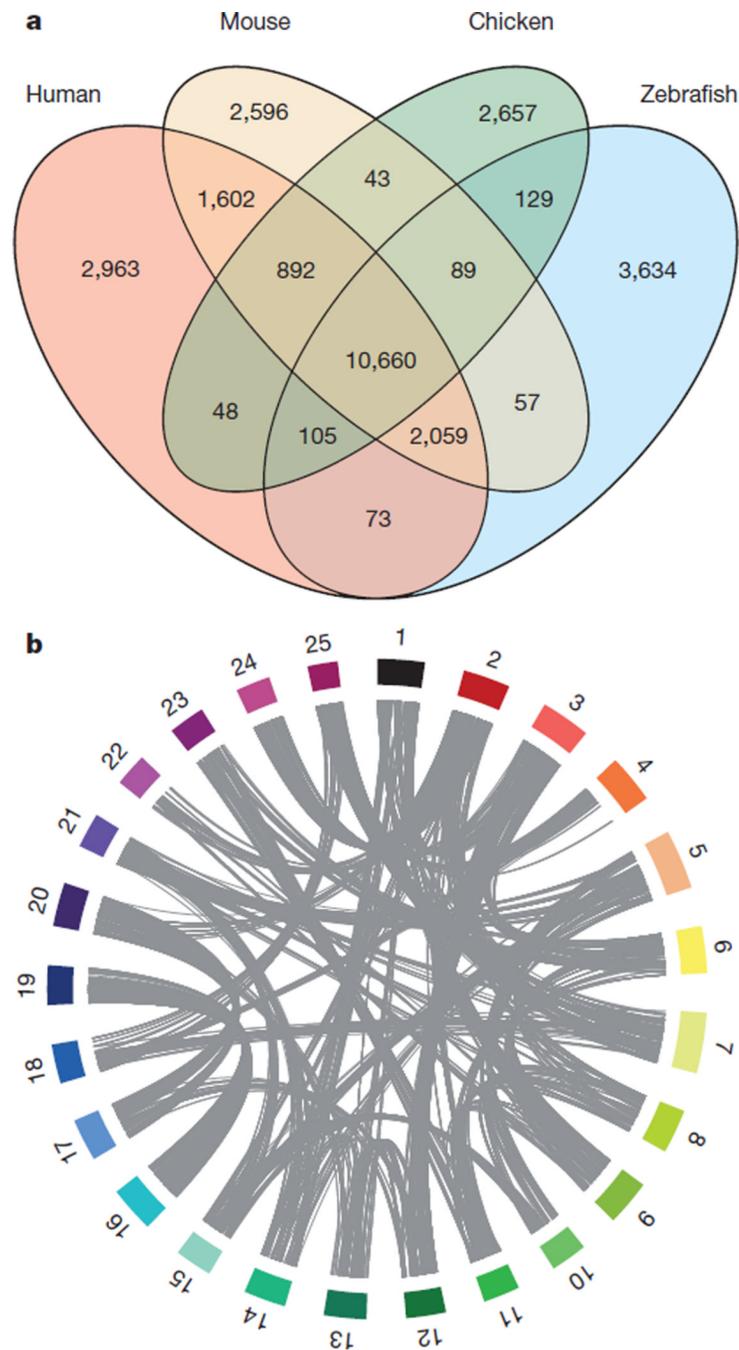
Figure 3. Evolutionary aspects of the zebrafish genome
a, Orthologue genes shared between the zebrafish, human, mouse and chicken genomes, using orthology relationships from Ensembl Compara 63. Genes shared across species are considered in terms of copies at the time of the split. For example, a gene that exists in one copy in zebrafish but has been duplicated in the human lineage will be counted as only one shared gene in the overlap. **b**, The ohnology relationships between zebrafish chromosomes. Chromosomes are represented as coloured blocks. The position of ohnologous genes between chromosomes are linked in grey (for clarity, links between chromosomes that share less than 20 ohnologues have been omitted). The image was produced using Circos[30].

Table 1

Assembly and annotation statistics for the Zv9 assembly

Assembly		Annotation	
Total length (bp)	1,412,464,843	Protein-coding genes	26,206
Total clone length (bp)	1,175,673,296	Pseudogenes	218
Total WGS31 contig length (bp)	234,099,447	RNA genes	4,556
Placed scaffold length (bp)	1,357,051,643	Immunoglobulin/T-cell receptor gene segments	56
Unplaced scaffold length (bp)	55,413,200	Total transcripts	53,734
Maximum scaffold length (bp)	12,372,269	Total exons	323,599
Scaffold N50 (bp)	1,551,602	-	-
No. of clones	11,100	-	-
No of WGS31 contigs	26,199	-	-
No. of placed scaffolds	3,452	-	-
No. of unplaced scaffolds	1,107	-	-

Data are based on Ensembl version 67. N50, the scaffold size above which 50% of the total length of the sequence assembly can be found.

Table 2

Comparison of human and zebrafish protein-coding genes and their orthology relationships

Relationship type	Human	Core relationship	Zebrafish	Ratio
One to one	-	9,528	-	-
One to many	3,105	-	7,078	1:2.28
Many to one	1,247	-	489	2.55:1
Many to many	743	233	934	1:1.26
Orthologous total	14,623	13,355	18,029	1:1.28
Unique	5,856	-	8,177	-
Coding-gene total	20,479	-	26,206	-

Data and orthology relationship definitions are based on Ensembl Compara version 67 (http://www.ensembl.org/info/docs/compara/homology_method.html).