Methods and Supplementary Information

Table of Contents

Figures

Tables

1 Mapping, Sequencing and Assembly

From the start of the zebrafish genome project it was planned to provide a clone based genome sequence of gold standard quality with finished contiguous sequence of >= 99.99% accuracy. Clone sequencing, finishing and assembly is a detailed and time consuming process, yet the zebrafish research community was in need of an accessible genome sequence as soon as possible. Hence at the same time, whole genome shotgun (WGS) sequencing was undertaken, with the aim of combining both approaches into integrated assemblies to provide the best possible genome sequence approximation at varying stages of the genome project[1]. Three solely WGS-based assemblies and six integrated genome assemblies, Zv1 to Zv9 were released to the public and from Zv3 onwards annotated in Ensembl[2]. The strategy to build these assemblies changed slightly with the amount and quality of data available. Supplementary Table 1 gives an overview of the releases and their features.

Assembly Version	Date	Size (Gb)	Fragments	Clone Content	Major Developments
Zv1	2002	1.17	158,689	WGS only	• first zebrafish genome sequence release • assembled using Phusion[3]
Zv2	2003	1.31	83,470	WGS only	
Zv3	2003	1.46	58,339	WGS only	• first assembly to be annotated in Ensembl
Zv4	2004	1.56	21,333	36 %	• first integrated assembly tied to genetic maps (priority given to T51 map) • WGS derived from mixed library capillary reads
Zv5	2005	1.63	16,214	43 %	• WGS28 assembly used for integration (mixed library capillary reads assembly)
Zv6	2006	1.63	6,653	63 %	• WGS28 used for integration
Zv7	2007	1.44	5,036	71 %	• WGS derived from one double-haploid fish (WGS29) • extensive removal of haplotypic duplication • tied to meiotic maps
Zv8	2008	1.48	11,632	77 %	• WGS derived from mixed library to reinstate genes missing in Zv7 (WGS28) • no restriction to addition of small unplaced WGS contigs • priority given to meiotic maps for scaffold allocation
Zv9	2010	1.41	4,560	83 %	• SATmap used for scaffold allocation • only finished clones used • WGS31 used for integration • restriction to addition of small unplaced WGS contigs back in place

Supplementary Table 1. Zebrafish assembly releases and properties

Note that the first Ensembl release of zebrafish genome data in spring 2002 was almost entirely WGS, containing only 16 Mb sequence from 152 large insert clones. (www.sanger.ac.uk/Projects/D_rerio/wgs.shtml).

1.1 SATmap

For the reference genome up through Zv8, the quality of the physical fingerprint contig (FPC) map hampered our attempts to produce assemblies with long-range as well as short range accuracy. This is due to haplotypic duplication and the high level of polymorphism among the zebrafish used to make the FPC map, which led to the production of many very short FPC contigs. Additionally the meiotic maps available at the time were too low resolution to order and orient most FPCs. There was thus a clear need for a reliable high-resolution map with high marker density to order and orient FPCs accurately.

The SATmap is a high-density meiotic map that provides genomic-clone sized genetic resolution. We took a novel approach to generating the map. Firstly, we took advantage of the fact that it is possible to create double haploid (DH) individuals, which contain only maternally derived DNA, are homozygous at every locus, can be raised to fertile adults and can be either male or female[4]. We created a panel of DH males and females from both AB and Tübingen (Tü) strain zebrafish. Secondly, we mated a single DHAB male with a single DHTü female to generate a family of genetically identical male and female F1 fish that were heterozygotes at every single nucleotide polymorphism (SNP) for which the parents differed (Figure 1). From these F1s, we generated a large panel of F2 grandchildren, and have maintained generations of subsequent inter-crossed families, which we call the Sanger-AB-Tübingen (SAT) strain. Thirdly, we sequenced each of the original DHAB and DHTü founders to over 40X coverage by Illumina sequencing. We identified 6,995,534 SNPs and selected 201,917 SNPs to ensure FPCs were covered by sufficient number of SNPs to be placed in a genetic map in the proper orientation. Finally, we genotyped the parental DHAB and DHTü fish, several F1, and 459 F2 fish at the 201,917 loci and assembled a meiotic map using a custom implementation of a genetic map assembly algorithm called MSTmap[5]. This new map, called SATmap (Supplementary Figure 3), has an average density of 1 SNP/10kb mapped at a resolution of 0.1cM (~60kb), which is an increase of over 30-fold in marker density and represents a 10-

fold increase in resolution over previous zebrafish meiotic maps and makes it the densest *de novo* meiotic map of any animal.

1.1.1 Generation of the doubled haploid hybrid mapping cross

To maximise single nucleotide polymorphism (SNP) numbers, genome coverage and meiotic mapping conversion rates we adopted a new meiotic map-making strategy. This approach also generated haplotypes for of the most common zebrafish strains and a new genetically defined hybrid strain. We created sets of homozygous AB and Tübingen strain individuals by fertilising eggs with UV inactivated sperm then treating the zygotes to a brief heat shock to suppress the first mitosis to thereby generate double haploids (DH)[4]. Male and female DH individuals (G0 founders) were bred and a pair that produced >20 mixed sex F1 individuals was selected for full genome sequencing. We mated the genetically identical F1s by inter breeding to generate a large number of F2 individuals.

1.1.2 Sequencing of AB and Tubingen haplotype genomes

The DH Tübingen female and DH AB male G0 founders were euthanised, frozen in liquid nitrogen and powdered with a mortar and pestle, DNA was prepared by Proteinase K digestion, phenol-chloroform extraction and isopropanol precipitation before Picogreen quantification[6]. Illumina libraries of 200, 300, 400 and 500 bp insert sizes were constructed without PCR amplification[7]. Additionally, 200 bp (Tübingen) and 3,000 bp mate pair (AB) libraries were generated according to standard Illumina instructions. Test lanes of 37bp paired end sequencing were performed for each library, reads were mapped to the Zv8 assembly using maq[8] and a library report generated for complexity, chimaerism or GC bias. High quality libraries were sequenced with 54bp or 76bp Illumina GAIIx technology to over 40X base coverage for each G0 (see Supplementary Table1). All sequence data has been submitted to the ENA (SRA Study: ERP000232 : The Sequence of the Two Most Common Zebrafish Laboratory Strains: AB and Tuebingen).

Individual	Library name	Library type	Insert (bp)	Read length	Mapped (Mbp)	Sample ID
DHAB	DHAB1bR-500NOPCR1	PE – No PCR	430	37	214	***ERS010792***
DHAB	DHAB1bR-500NOPCR2	PE – No PCR	485	37, 54, 76	15,547	***ERS010792***
DHAB	DHAB1bR-200NOPCR1	PE – No PCR	168	37, 54, 76	45,507	***ERS010792***
DHAB	DHAB1bR-3kb	MP – standard	3,000	37	3,550*	***ERS010792***
				Total	61,268 (**43x**)	
DHTu2	DHTu2-NOPCR2	PE – No PCR	129	54	11,794	***ERS010793***
DHTu2	DHTu2v2_350-400NOPCR	PE – No PCR	311	37, 76, 100	3,666	***ERS010793***
DHTu2	DHTu2v2_400-450NOPCRdblsel	PE – No PCR	369	37, 76, 100	6,485	***ERS010793***
DHTu2	DHTu2v2_450-500NOPCRdblsel	PE standard	426	74, 100	3,441	***ERS010793***
DHTu2	DHTu2v2_500-550HCdblsel	PE – No standard	425	37, 76, 100	38,737	***ERS010793***
				Total	60,682 (**45x**)	

Supplementary Table 2. G0 founder sequence data summary

Most data is paired end, as mate pair libraries require larger amounts of limited DNA. All libraries had one test run performed, aligned using maq[8] and analysed using mapcheck[8] and custom scripts that count read pair FLAGS and for levels of mapping, depth, duplicates, chimaerism and possible GC bias. The same package was used here to calculate mapped reads. Within ENA Study ERP000232 there are two Sample IDs that can be used to access data for each strain.

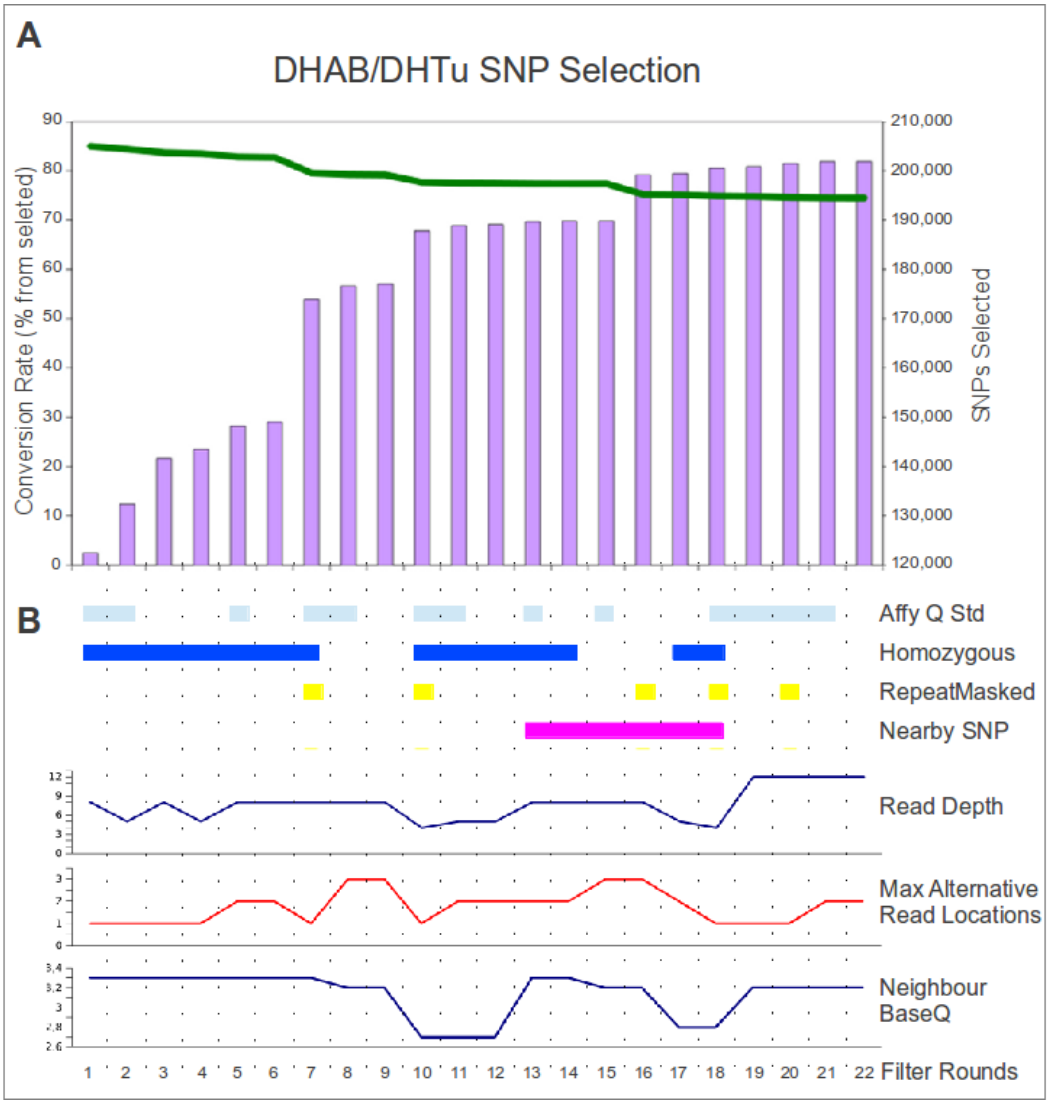
1.1.3 Marker SNP selection

For logistic reasons SNPs were selected when we reached a sequence depth of 25x for the DHAB founder, and 4.5X for the DHTü founder. After mapping reads to the Tübingen based Zv8 reference genome with maq[8] and removing duplicates we called all SNPs with cns2snp and insertions-deletions (indels)

with the indelpe and indelsoa programs. We minimally filtered SNPs using SNPfilter (within maq.pl) with default settings except we supplied indelpe and indelsoa files and used '–w =15' to suppress calling SNPs within 15bp of an indel. SNPs with all the quality-associated data were then entered into a MySQL database and MySQL scripts were used to select SNPs.

Our aim was to genetically anchor as much of the genome as possible and to provide a platform for mapping complex traits. We broke the Zv8 genome assembly (including small non-attached scaffolds) into small 3.5kb windows and choose the "best" SNP for each. This strategy risks a poorer rate of converting candidate SNPs into useful genotyped markers but ensures more even genome coverage.

The "best" SNP was chosen for each window starting with the highest quality candidate SNPs, and steadily decreasing the quality settings, that is, parameters of read depth, mapping quality, SNP quality, neighbouring quality, unique sequence and polymorphism likelihood. For polymorphism likelihood, we compared DHAB and the Tübingen-based Zv8 assembly with or without confirming early, low pass DHTü data. Ultimately, a total of 22 iterative steps were used (Supplementary Figure 1). Aware that some distinct but similar genomic regions might have been incorrectly collapsed as haplotypes in the Zv8 assembly we allowed some SNPs that appeared to be heterozygous in the DH fish if they were otherwise the best SNP in that window.



Supplementary Figure 1. SNP selection

SNP selection was performed using iterative rounds each decreasing or permuting a given SNP quality associated parameter. A. The steadily growing cumulative numbers of selected SNP at each iteration (mauve bars), and the decreasing conversion rate (in %) of the cumulative selected candidate SNPs into markers in the final SATmap. B. Filters changed at each stage: "Affy Q std" predicted to be present in an NspI fragment ≤1000bp in size plus 30-70%GC and less than 7 homopolymers in a 33bp window centred on the SNP, "Homozygous" homozygous in the AB and Tü G0 fish and polymorphic between them, "RepeatMasked" falls into a region softmasked by Repeatmasker[9], "Nearby SNP" a nearby SNP is allowed if might be due to small overlap with multi-copy sequence, "Read Depth" minimum read depth needed at that position, "Max. Alternative Read Locations" the maximum locations reads mapping in this area could also have also been mapped elsewhere in the genome (a measure of either real duplication, or different haplotypes of that region included in the assembly), "Neighbour BaseQ" minimum quality needed for the base calls either side of the SNP.

According to Affymetrix guidelines, SNPs were required to fall within NspI fragments of less than 1000bp in size as predicted from the genome sequence, have a GC ratio of 30%-70%, no other neighbouring polymorphism and with no homopolymers of 7bp or longer in the 25 bases centred on the SNP. A SNP list of 223,723 was passed to the manufacturer for design and a custom array of 201,917 SNP generated. This custom array is available from Affymetrix (Item # 520747, Array Name ZFSNP200m520747F, Array Format: 49-7875).

1.1.4 Microarray Genotyping

The custom SNP chip was used according to Affymetrix instructions for an NspI Whole Genome Sampling Array (WGSA) to genotype 21 DH samples, 13 F1s and 459 F2 fish. The BRLMM-P clustering algorithm in the Affymetrix Power Tools software (APT) was used to determine genotypes from the resulting CEL files, as we knew the assumed genotypes for each SNP in the cross, that is, F0=homozygous (DH), F1=heterozygous and F2=unknown, we were able to use a 'hints' file to supply assumed genotypes according to the manufacturers recommendations (www.affymetrix.com/support/developer/powertools/changelog/VIGNETTE-WGSA-clustering-without-priors.html).

1.1.5 Post Genotyping Filtering

We excluded 29 F2 samples that had "undetermined" genotype call rates more than 2 standard deviations (SD) outside of the mean, leaving 430 F2s. We defined a subset of 149,879 SNPs that were informative using filters based on the genetics of this cross. We required that SNPs must either be homozygous in DHAB and DHTü and also heterozygous in F1s, or segregate in a Mendelian fashion (AB/AB 25%, Tü/Tü 25%, 50% AB/Tü) in the much larger F2 dataset. SNPs that passed either of these filters were used to make the *de novo* genetic map using MSTmap[5]. This SNP dataset is the largest experimentally verified set of SNPs for zebrafish and has been deposited in

dbSNP with full G0, F1 and F2 genotype calls using BATCH_ID “SATMap-Markers” and Submitter_Handle “SANGER_STEMPLE”.

1.1.6 The de novo genetic map strategy

Generating a *de novo* genetic map from 64,447,970 genotypes (149,879 markers on 430 individuals) exceeds that of any published map. We empirically determined from test runs that the MSTMap was the most likely computer program to handle the data load. Despite the possibility that MSTMap could compute the map in one attempt, we estimated that it would require around 3 months and 800Gb of RAM to complete a run with just one set of parameters. As MSTMap itself was not designed to take advantage of a compute cluster by running in a highly parallel manner we needed to implement a parallel strategy. Test runs revealed that the MSTmap runtime and RAM requirements scaled linearly with the number of individuals, but exponentially with respect to the number of markers. We split the SNPs in random sets of 10,000 SNPs on 430 individuals and ran MSTMap on each set independently. Interestingly, each set possessed more markers and more individuals than any existing zebrafish meiotic map. Together with optimisation using the Intel compiler, instead of the default gcc, we were able to reduce the timing and memory requirements for a single MSTmap instance to 45 min and 800Mb of RAM, which is easily manageable on a standard server blade or powerful desktop PC. MSTMap was run with the following settings.

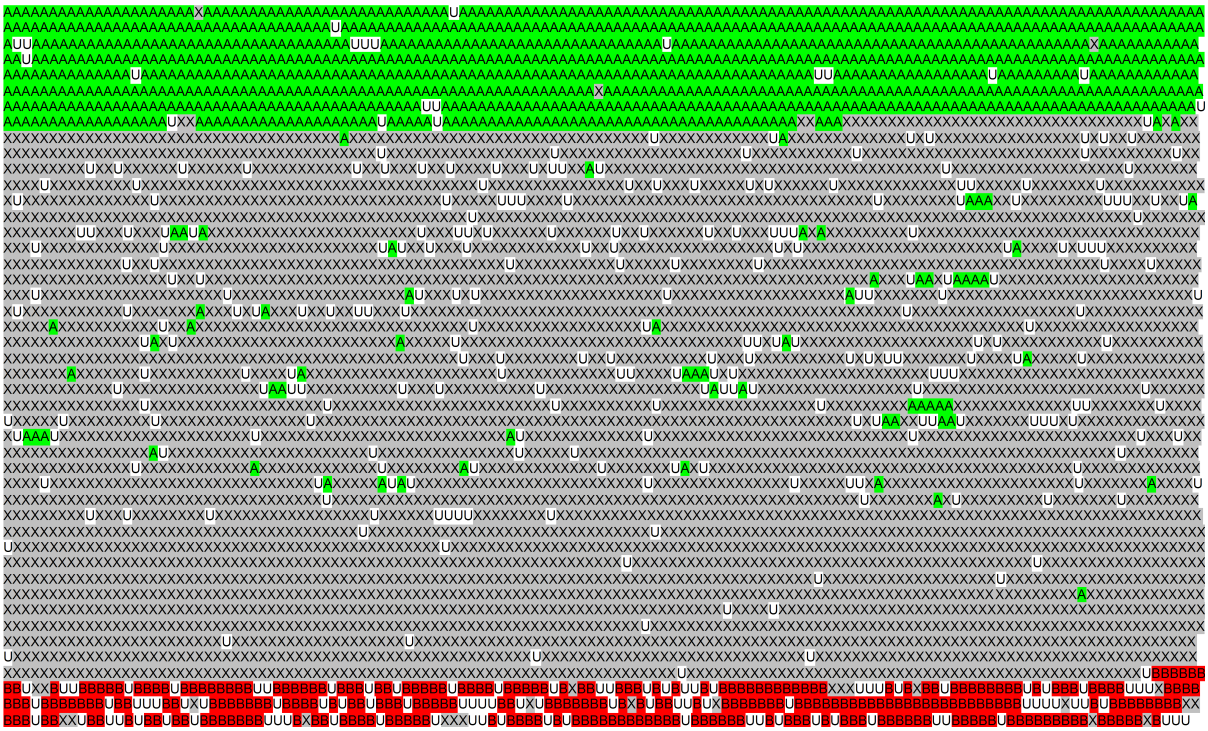
```
population_type: RIL2
population_name:TuAB
distance_function: kosambi
cut_off_p_value: 1e-20
no_map_dist: 2
no_map_sizs: 2
missing_threshold: 0.13
estimation_before_clustering: no
detect_bad_data: yes
objective_function: ML
```

To fuse the 15 new maps, the different linkage groups of markers were assigned to the most probable chromosome using a voting strategy based on their Zv8 assembly positions, which is anchored on existing genetic maps. All markers were then regrouped according to their chromosome into 25 datasets plus the unplaced markers. We then reran 25 instances of MSTmap, one for each chromosome dataset, and included all the unplaced markers from the 15 first round runs included in each new dataset. We found that none of the previously unplaced markers were assigned to more than one chromosome and we were left with 141,675 SNPs in our *de novo* meiotic map plus 8,204 remaining unplaced SNPs.

1.1.7 Genotyping error correction and the final SATmap

The first version of SATmap had genetic sizes of individual chromosomes at over 1000 cM whereas this should be ~100 cM. Manual examination of the genotype calls for individuals revealed large numbers of apparent close double crossovers (Supplementary Figure 2), which are extremely unlikely, particularly as SATmap markers are so dense, and are characteristic of genotyping errors. MSTmap like most other *de novo* map programs uses a built-in error correction algorithm to detect such errors. Given the large size of our dataset, however, it is possible that MSTMap was unable to correct such a large dataset, or that a small percentage were missed. The effect of genotyping errors in inflating the apparent genetic distances particularly as the number of genotype calls increases has been discussed elsewhere[10]. We undertook a detailed comparison to existing genetic and physical maps and found that the order of the markers was correct, but not the genetic distances. To solve this problem we implemented our own genotyping error correction to remove genotype errors, which appeared in the map to be double crossovers. Sorting markers by their SATmap positions we were able to identify SNP genotype calls within each F2 individual that produced apparent double crossovers, that is up to five genotype calls that disagreed with the ten genotyping calls on either side, and changed the putative erroneous genotype calls using a custom script to an “undetermined” genotype. We then reran the MSTMap process described above, and corrected once more. After two

rounds of MSTMap processing and removal of genotype errors the chromosome sizes were reduced to their expected sizes (Supplementary Table 3), with marker orders in agreement with the extant genetic and physical maps. A further 9,573 markers were either rejected by low genotyping rate, that is greater than 13% “undetermined” calls after replacing apparent errors with “undetermined” or were unplaced. Additionally 29 F2 individuals were removed due to poor call rates.



Supplementary Figure 2. Example of the uncorrected genotypes

Shown is an example of the uncorrected genotypes of a complete linkage group for an F2 individual. All genotypes are the sum of the two chromosomes in the diploid F2 individual, each chromosome having undergone one recombination during meiosis, here one each on the *p* arm and *q* arms. The “A” genotype calls are for homozygous Tübingen alleles (coloured in green), “B” alleles for the homozygous AB strain alleles (coloured in red), heterozygous calls are “X” (coloured in grey). Undetermined calls are “U” (coloured in white). Numerous apparent genotyping errors calls are visible as single or small groups of calls that disagree with a region of consistent calls. Note all apparent genotype errors are Het<=>Homo changes which are more common miscalls for the genotyping platform, and never the unlikely HomoTü<=>HomoAB.

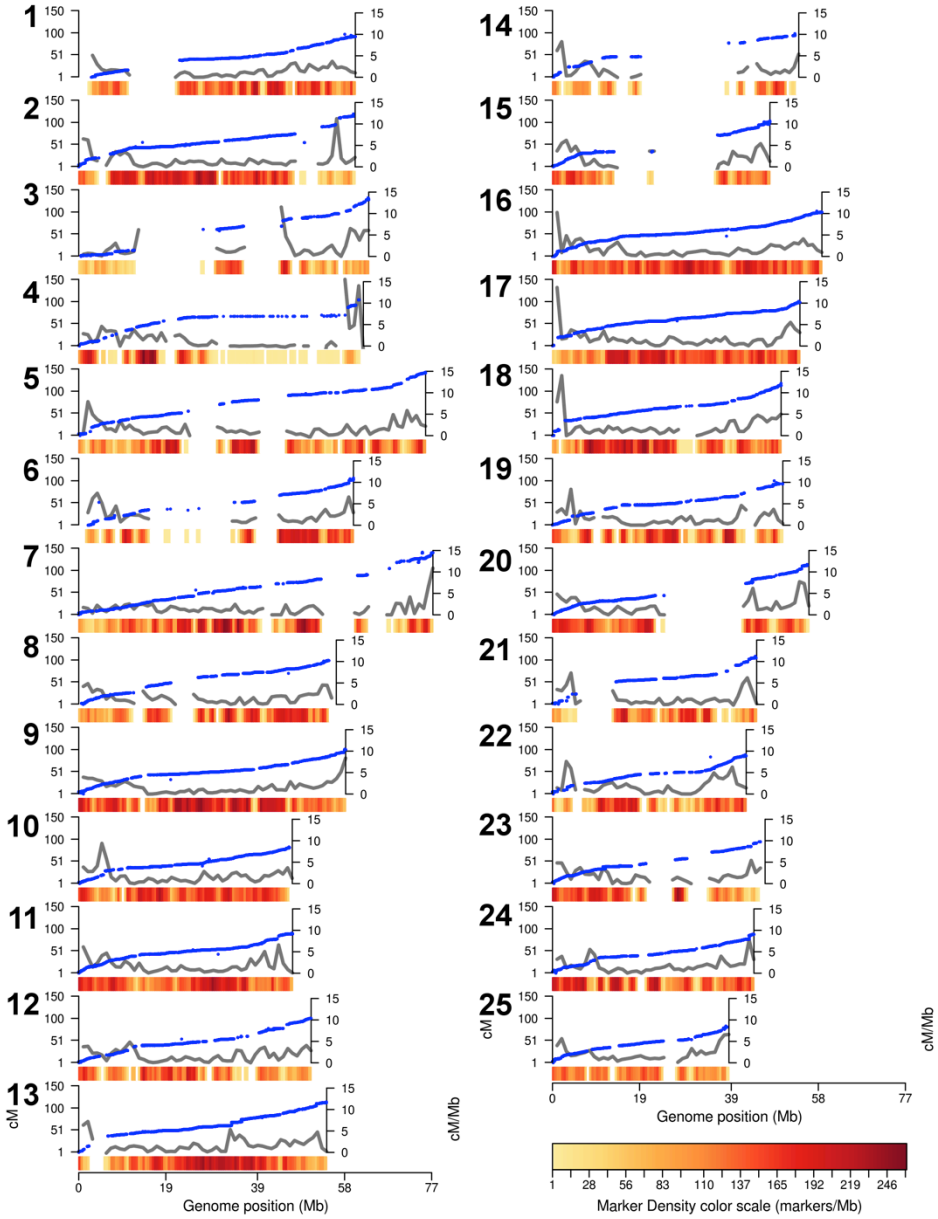
The final version of SATmap (Supplementary Figure 3) contains 140,306 markers on 430 F2 individuals with a total genetic map size of 2,627cM

comparable to the less dense but similarly sex averaged MGH panel (AB X IN) of 2,177cM[11], the two AB X NA families of 2,346cM and 2,333 cM[12], and as expected is lower than the female only recombination genetic map size of 3,192 cM from the Heat Shock (HS) panel[13].

Chr	Start Size	Start #Markers	Iteration1 Size (cM)	Iteration1 #Markers	Iteration2 Size (cM)	Iteration2 #Markers
1	1,286	6,029	167	5,999	98	5,975
2	1,585	7,866	219	7,817	119	7,790
3	830	3,010	174	2,978	132	2,965
4	778	3,247	161	3,225	104	3,207
5	1,477	7,078	222	7,034	143	7,002
6	963	4,429	165	4,410	106	4,392
7	1,635	7,814	274	7,795	141	7,749
8	1,279	5,882	182	5,839	98	5,813
9	1,803	9,048	201	8,991	100	8,958
10	1,356	6,628	171	6,578	83	6,555
11	1,529	6,965	178	6,930	90	6,903
12	1,227	5,111	194	5,084	103	5,054
13	1,564	7,718	194	7,660	114	7,632
14	559	2,219	133	2,205	98	2,191
15	720	3,131	172	3,132	103	3,112
16	1,763	8,866	230	8,830	102	8,790
17	1,249	7,831	157	7,810	101	7,789
18	1,311	6,590	210	6,547	118	6,518
19	1,016	5,166	160	5,144	100	5,125
20	1,146	5,107	183	5,077	113	5,057
21	695	3,543	159	3,534	108	3,518
22	1,001	4,352	165	4,331	88	4,309
23	963	4,410	164	4,383	94	4,361
24	1,223	5,887	160	5,853	87	5,837
25	989	3,748	152	3,721	83	3,704
	29,949	141,675	4,548	140,907	**2,627**	**140,306**

Supplementary Table 3. Genotyping error correction

Error correction of genotyping by removal of double crossovers, using two iterations of correction and re-calculation of the *de novo* map. The final size of the complete map is 2,627cM and 140,306 markers.



Supplementary Figure 3. SATmap overview

The position of each SATmap marker is shown in blue plotting its physical position on the X-axis (in Mbp), and on the left Y-axis its genetic map position (in cM). The sigmoidal line is characteristic of increased recombination at the telomeres and suppressed recombination near the centromere. The marker density across each chromosome is shown as a heat map, and the recombination rate region is plotted in black against the right Y-axis (cM per Mb). Varying recombination rates across chromosomes are apparent indicating hot and cold spots. Finally gaps in the SATmap that cover less than 10% of the genome, are visible as marker free regions, in these areas other less dense genetic maps were used to guide assembly. Chromosome graphs displaying additional information are seen in Supplementary Figures A1-A25.

1.2 FPC Map

At the beginning of the genome sequencing project a physical map was generated using the FPC method[14,15]. Two bacterial artificial clone (BAC) libraries, CHORI-211 (Pieter de Jong) and the DanioKey library (Ronald Plasterk and Keygene NV) were used to generate the map. Each BAC clone was first digested with *Hind*III then fragments were separated on agarose gels along with known standards. These data were then assembled into contigs on the basis of shared fingerprint bands using an FPC analysis program[14,15]. Additionally, each of the BAC clones was sequenced using primers set minimally into the cloning vector to establish a pair of end sequences for each BAC. During library construction sets of fosmid and plasmid libraries were also generated and individual clones were end sequenced. The end sequences are available as a public resource (www.ncbi.nlm.nih.gov/Traces/trace.cgi?cmd=retrieve&val=species_code%3D%22DANIO+RERIO%22).

Each contig, comprising one or more minimally overlapping BACs, was localized to chromosomes using a combination of RH mapping of BAC end sequences (Geisler Lab, Max-Planck Institute, Tübingen and L. Zon Lab, Children's Hospital, Boston) and alignment of markers to sequenced clones by ePCR to produce the T51 map[16]. Fluorescence *in situ* hybridization (FISH) experiments were also used to assist with localization of any unplaced contigs and to define contig order[17]. These experiments were also used to ascertain sizes of some of the gaps along the chromosome. This initial phase produced a map in a large number of contigs. The clone library resources are listed in Supplementary Table 4.

Library Name	Genomic DNA	Clone Type	No. Clones	No. Fingerprints	Sequence in Zv9 (bp)
CHORI-211	6 Tü Testes	BAC	105907	91718	454,220,798
Daniokey	6 Tü Testes	BAC	104064	85289	454,901,002
Daniokey pilot		BAC	11808	10247	72,217,431
RPCI-71	7,000 Tü embryos	BAC	33408	22439	16,975,566
CH73	1 dh Tü	BAC	297528	7548	138,215,717
CH1073	1 dh Tü	Fosmid	183168	2317	31,988,090
ZFISHFOS	1 dh Tü	Fosmid	269280	2967	7,138,096

Supplementary Table 4. Large insert libraries

Details of the libraries used for building the FPC map. Note that the libraries created once the main sequencing project was underway (CH73, CH1073 and ZFISHFOS) rely heavily on electronic mapping of end sequences rather than fingerprint for placement. Further information on these libraries and the source DNA used can be found at http://www.sanger.ac.uk/Projects/D_rerio/library_details.shtml.

Once much of the genome had been sequenced it became apparent that generation of a physical map and subsequent genomic sequence using libraries derived from multiple fish would cause problems in the latter stages of contiguation of the map due to high polymorphism rates. In many cases true overlaps confirmed by FISH data had a high level of discordance based on sequence[17]. Due to high polymorphism rates the map was subsequently augmented with fingerprint analysis and BAC end sequences from a double haploid (DH) fish prepared by J. Postlethwait (University of Oregon) using the 'heat-shock' technique[18,19]. All loci tested on this fish were found to be homozygous. DNA from this fish was used to create a BAC and fosmid library, CHORI-73 and CHORI-1073 respectively (Pieter de Jong). In addition to using these libraries to augment and improve the FPC map, double haploid DNA was used to create a plasmid whole genome shotgun library. An assembly of these data was then compared to the emerging clone based sequence from the original libraries as a way to detect whether or not a clone was from the same haplotype as the homozygous fish.

1.3 Genome assemblies

1.3.1 Zv7 and Zv8 Assemblies

Prior to the assembly of Zv9 there were only a few genetic maps available with sufficient marker density or resolution on which to anchor the physical maps. Specifically we used the T51 radiation hybrid map, which had ~10,000 markers, but the order and orientation of markers was only useful for short ranges[16,20]. Two meiotic maps, MGH[21,22] with ~5000 markers and Heat Shock (HS)[6,23] also with nearly 5000 markers, were also available and although these maps provided a long-range genetic position accuracy in comparison with the T51 map, none of the available maps were sufficient to obtain an accurate assembly of the genome.

After the generation of assembly Zv7 we undertook an analysis of the correlation between Zv7 and the three meiotic maps (HS, MGH and T51) that highlighted a particular problem in the Zv7 assembly. While the T51 map was used to anchor FPCs in Zv7 there was nevertheless poor correlation between the assembly and the map. Indeed, the average correlation between the three maps and the assembly was only around 0.7. To overcome this problem for Zv8 we reorganized the FPCs, prioritising the meiotic maps HS and MGH for long-range order and chromosome assignment, then used the T51 radiation hybrid map mostly to resolve local order and orientation. In this process we considered an FPC as an indivisible unit. An FPC provides a link between those markers mapped on that FPC and therefore makes an association between the three different maps. We used this association to integrate map position data derived from the three different maps by sequence alignment between genetic markers and the sequence of each FPC. By using all three maps we increased the coverage and resolution to the available maximum.

We weighted each marker based on the overall quality of each map and the mapping quality of the marker on the genome. In principle, the two meiotic

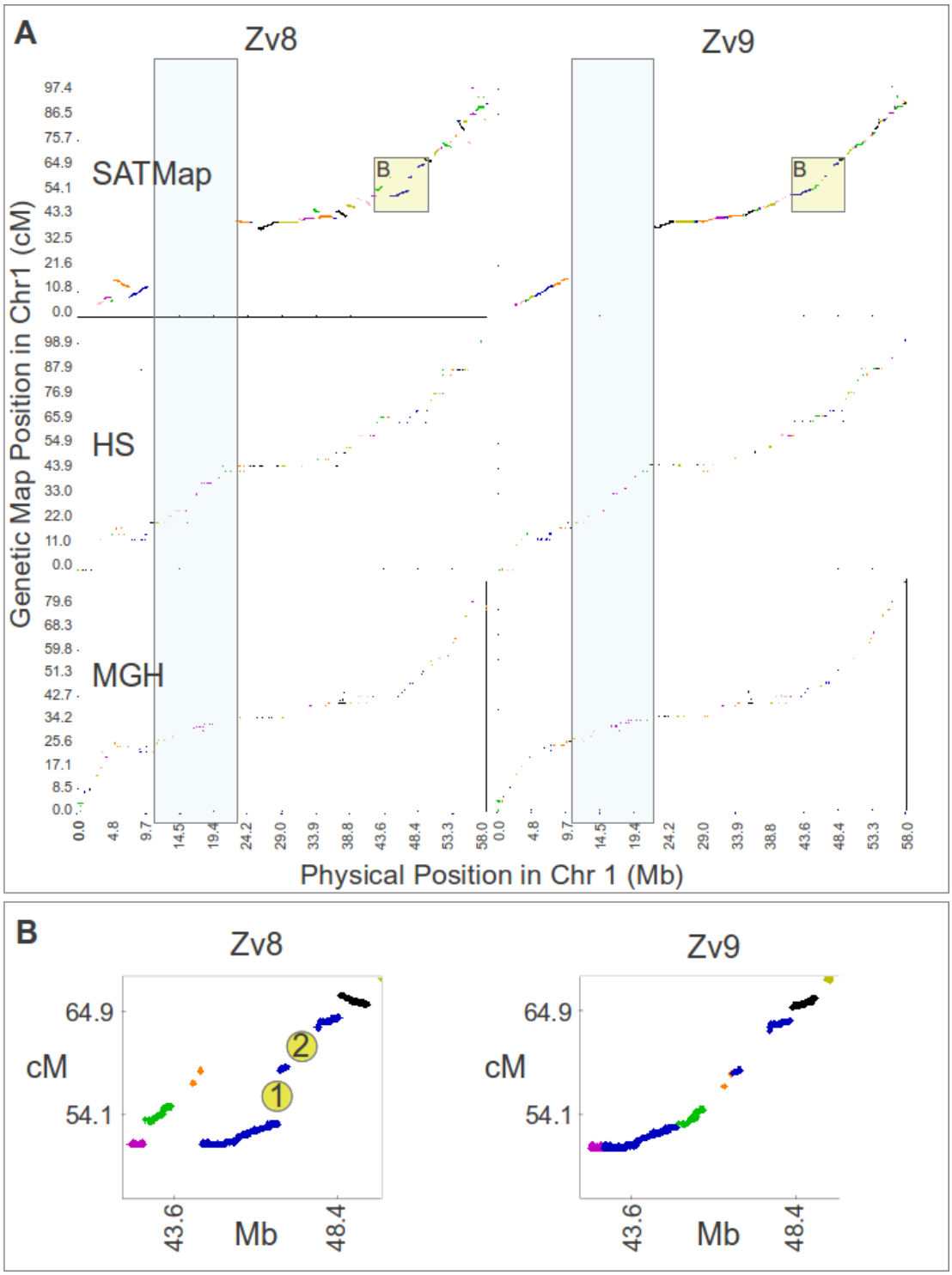
maps more accurately represent the genome structure, but are lower resolution than the T51 map. Between the two meiotic maps, the HS map is more reliable because allele scoring is more accurate than for the MGH map. With the initial marker mapping information each FPC was first assigned to a chromosome and an FPC position was evaluated by a weighted mean of the positions of the markers. Given the FPC chromosome assignment and position, the FPCs were then sorted based on their map positions giving precedence to HS data, then MGH and finally T51. We saw a striking improvement in the correlation between the Zv8 assembly and each of the genetic maps, which was at an average of 0.96 for each chromosome. In addition, we found that chromosome 4 had a significantly increased size closer to the size expected by flow cytometry. This approach allowed us to generate the Zv8 assembly, but we recognised some difficulties in the placement of FPCs and cDNA sequences.

1.3.2 The Zv9 genome assembly

Zv9 was assembled using the SATmap as the anchoring map. Similar to Zv8, we used a weighted voting of markers for the positioning of fingerprint contigs (FPC) and whole genome shotgun contigs using the genetic maps in the order SATmap>HS>MGH>T51 (Supplementary Figure 4). Given the much larger number of votes of SATmap markers other maps would only have a significant role in filling marker gaps in the SATmap. HS was weighted a little higher than MGH because it is a DH only cross and has fewer genotyping errors as they are called from homozygous alleles. T51 is still informative, especially in regions where recombination is rare, for example near centromeres.

Unfortunately, we found about 10% of the positions in the cross to be monomorphic, probably because the grandparent DH AB and DH Tü fish share a recent ancestor (Supplementary Figure 4). In these regions of monomorphism, we were forced to use the genetic map assembly strategy we had employed for Zv8 and integrated information from the HS, MGH and T51

maps to cover the monomorphic gaps in the SATmap. The monomorphic gaps are now being rectified using another cross and genotyping strategy.



Supplementary Figure 4. Chromosome 1, Zv8 versus Zv9

A. SATmap, HS and MGH map markers over Zv8 and Zv9 assemblies for chromosome 1. Colours represent different fingerprint contigs (FPC), which produce the main proportion of the zebrafish genome assembly. The blue frame highlights a gap in SATmap, which is covered at lower resolution by MGH or HS maps. The yellow frame indicates the region magnified in panel B. The density of SATMap markers in Zv8 highlights the Zv8 assembly errors. Agreement between genetic and physical maps leads to a sigmoidal graph, which deviates from the diagonal due to a higher recombination at telomeres, and a lower one at the centromere. SATmap markers make it apparent that many FPCs are misplaced on the physical map, and even inversion of FPCs are apparent. Previous maps did not possess sufficient density or resolution to unambiguously place as many regions as SATMap. B. SATMap can detect errors in FPC contigs. The panel represents one FPC contig (blue) with two gaps. Circle 2 shows a gap in the genetic map, which correlates with a gap in the physical assembly. This shows the absence of markers that cover the region. In circle 1, however, the gap in the genetic map is not correlated with a gap in the physical assembly, suggesting the absence of a sequence region in that position, which could result either because the region has not actually been sequenced or because the assembled FPC was incorrect. In this case the blue FPC is incorrect and the two adjacent FPCs (orange and green) contain the sequence region missing in the blue FPC. For Zv9 this FPC was split into 2 FPCs and the other two placed in the middle according to the genetic SATMap information. In Zv8 the black FPC contig is the correct place but in the wrong orientation according to SATmap data, thus it was inverted for Zv9.

1.4 Clone Sequencing

Clones were selected for sequencing based on the minimally overlapping tile path from the FPC and genetic maps. Tile path clones were then picked from the library plates and fingerprinted again with *Hind*III to check the quality and integrity prior to a shotgun library production. Clone DNA once prepped was sheared and then ligated into pUC vector for sequencing. The subclone insert sizes used were 2000-4000bp and 4000-6000bp. The number of sequencing reads attempted for each clone was calculated using the predicted insert size ascertained from *Hind*III analysis. For example for each 50kb of clone insert size a set of 384 subclones would be prepped and sequenced from either end. This formula meant that each clone had a minimum coverage of 6-8x attempted, which was then assembled using Phrap[24].

1.4.1 Clone finishing

The quality of the genome is central to its utility and clones were advanced to the best standard possible in a process known as 'finishing'. For the zebrafish genome more than 80% of the known genome length is covered by finished large-insert clone sequence. This level of finished sequence coverage in zebrafish is unique among the published fish genomes, as the other sequenced genomes are either mostly or entirely comprised of assembled WGS sequence. The high quality of the zebrafish reference allows much more detailed and accurate annotation and experimental analysis.

A finished clone is identifiable in the EMBL/NCBI sequence databases as Phase 3 (complete sequence). The finishing process is a structured one that produces a substrate that can be used with confidence and which supersedes the scaffolded supercontigs available from the WGS. A finished, or Phase 3 entry, is defined as a contiguous sequence that has been confirmed by double stranding or use of alternate chemistry where necessary to Phred score quality of >=30[25]. All sequencing anomalies were resolved and the sequence was compared to restriction digest data from more than one enzyme. The finishing process takes the sequence data produced in the shotgun phase and seeks to close gaps and raise the quality of the sequence to less than 1 error in 10,000 bases. This approach makes the zebrafish genome sequence consistent with the human and mouse genome-sequencing projects. The finishing process was carried out in two key stages: an initial automated round followed by manual finishing.

To produce an automated round of improvement the shotgun sequence was compared to the most up to date WGS assembly using SSAHA[26] and concurring consensus pieces were then integrated into the assembly. Once the clone based shotgun and WGS data were combined, new sequencing primers were automatically selected using a script to drive the Primer3 oligonucleotide selection program[27]. These custom primer oligonucleotides were used on suitable subclone templates to obtain sequence covering gap regions. Once the new extending sequence data were available, the clone

data were reassembled using Phrap before electronic assessment for contig number reduction and delivery into the manual finishing phase.

The manual finishing process used techniques, developed during the human and mouse genome projects, to improve existing sequence data and close all sequence gaps[28]. The zebrafish sequences presented some unique challenges in the finishing process, centring mainly on the repeat structure within a clone. As part of the assembly process, the sequence was screened against the known zebrafish repeat database and known repeats were tagged for reference. With the repeats correctly tagged we were able to apply standard approaches to them. Some examples of these repeats and the details of the approaches taken are listed in Supplementary Table 5.

Given the cost and effort required to produce finished clone sequence it was essential to control quality. Quality was assessed in three stages. Firstly, each 'finisher' would assess the overall quality of each clone sequence using a checklist and a suite of software. Secondly, an independent team assessed the quality. Finally, the sequence was checked against cDNA sequence and gene model predictions during the manual gene annotation stage.

Once finished and quality control checked the clones were then ready for integration into the WGS. It was possible at the point of integration to identify any issues arising either with placement or overlap with previously finished sequence. When two adjoining clones were seen to be from the same haplotype 2000bp of overlapping sequence was finished. If the haplotype was found to be different, the full overlapping insert was finished to facilitate investigation after integration into the WGS.

Impact on Finishing Process	Strategy
Small Unit Repeats (di-nucleotide typically TA)	
Variation amongst subclones (possible deletion of units during replication)	Where two contigs ended in di-nucleotide repeat and spanning subclone information was present, the join was forced and sizing information for missing di-nucleotides was taken from restriction digest information to accurately represent the original clone. Alternative chemistry was used to sequence through some of these in the early stages of the project to establish a maximum length for a TA di-nucleotide run. These data showed that gaps >100bp required a transposon library to capture any unique data within the repeat [29].
Local Structural Problems (Mono Runs and Hairpin repeats typically Dr284)	
Prohibitive to accurate chemistry use (sequencing reads misread bases and/or enzyme 'falls off' template)	Alternative chemistry was used. In many cases these sequencing artefacts were directional. Primers were designed such that the alternate strand of a subclone could be used as the sequencing template. Use of short insert libraries was also employed[30].
Zebrafish Large Tandem Repeats (repeat units of varying length but multiple copies in tandem array covering =>10kb of sequence)	
Total repeat array is in excess of insert size so subclones central to the repeat cannot be placed correctly.	The clone sequence was first subject to analysis to detect any whole or partial gene objects missing in the tandem array. If no gene objects were present then the following strategy was applied. Repeat units from subclones anchored in unique sequence were finished to represent the unique 'break point' to represent the repeat unit itself. Restriction digest data were then used to size the repeat region, this was done with a custom selection of enzymes. Accurately finished sequence contains the predicted length of the repeat, which was ascertained from restriction digest.

Supplementary Table 5. Repeats and finishing strategies

1.4.2 Further technical development for large insert clone sequencing

During the course of the genome project the emergence of new sequencing technologies demanded that methodologies for generation of data for clones were challenged to ensure the most efficient method was employed. High-throughput sequencing instruments increased efficiency of time and money such that the overall sequence output was high quality and was also flexible for post data generation improvement using traditional finishing techniques. A method for combining non-indexed pooled Illumina data for genomic clones with capillary-based WGS data for the Zebrafish Genome Project was evaluated. Complexities with the Zebrafish Genome Project meant that the

method would need to be robust to produce data that could be used as part of the effort to produce the final reference genome. To date with the inclusion of the validation data, 5.6Mb of clone-based data have been produced using this method.

We decided to pool the clones without use of indexing adapters to reduce the numbers of sequencing libraries that needed to be created and to avoid PCR in the library construction stage, as this was known to introduce sequence bias[7]. Pooling clone DNAs without using indexing adapters to help de-convolute the data presented a bioinformatics challenge, as we required that data could be broken into the original clone parts for finishing.

All DNA was quality controlled and normalised to ensure even representation across the generated data set. Libraries were created for the clone pools and each was run on an Illumina Genome Analyser II instrument. All runs were done using a 76-cycle length on a paired end module. The resulting data sets gave us approximately 1000x fold coverage of ~1.2Mb total clone length for each pool.

To achieve the best assembly for all the available datasets, we used a method to combine pooled Illumina reads with previously obtained double haploid capillary WGS reads. For each pool, we first used Fuzzypath (ftp://ftp.sanger.ac.uk/pub/zn1/fuzzypath/), an Euler path based algorithm to extend short reads into much longer sequences with a length of 2-3kb. A traditional capillary assembler Phusion[3] was then used to assemble the extended sequences. After initial assembly, the resulting contigs were compared with the scaffolds obtained from pure double haploid capillary WGS reads using SSAHA2 (www.sanger.ac.uk/resources/software/ssaha2/)[26]. Parameters for this SSAHA2 alignment were carefully selected to ensure that consensus from genomic duplications were not used. WGS reads that contributed to the matched scaffolds were extracted. In the last step, extended sequences from Illumina data and extracted WGS reads were assembled using Phusion again to obtain a final assembly for finishing.

The presence of clone data within a pool was confirmed using available clone end sequences. These were aligned against the final assembly using cross_match (www.phrap.org/phredphrapconsed.html). Matches with clone end sequences confirmed the representation of a clone in the final data set and provided initial contig identification and orientation. A further enhancement to this process using transcript sequences is currently being developed and will allow additional contig ordering and orientation to provide a set of target regions for improvement in the first parse of manual intervention with the data sets.

Presently these Illumina clone pools are in active manual finishing, are present in the public databases and will be incorporated into the new builds for the genome. Finishing to a high standard using Illumina sequencing has been possible by adaptation of some of the software tools and traditional approaches used on solely capillary data[31]. We have proved that this is an efficient method of producing clone data without compromising quality, an issue of major concern to the sequencing community[32]. We are now working to use larger pools of clones to increase the efficiency of cost and time for clone data generation and have shifted to generating indexed libraries for each clone.

1.5 Whole genome shotgun (WGS) assembly

From the beginning of the project, to complement the clone-based physical map and to create public access to the genomic sequence for the community WGS assemblies were produced from the original DNA sources. Phusion[3] assemblies of these data were made at intervals and released to the public via Ensembl[2]. Existing FPC contigs could be aligned to these assemblies to verify them and also to suggest further contiguation that fell outside the parameters set for contig building.

Although several previous WGS assemblies were used during earlier stages of the genome project, they were generated from sequences of many individual fish, which led to ambiguous assembled contigs due to haplotypic

variation, as well as artefactual sequence duplication. For the Zv9 assembly, a new WGS assembly, WGS31, was created using Illumina sequencing reads from a single female double-haploid Tübingen fish, which was the grandmother of the SATmap cross. From this fish, 289 million reads providing approximately 30-fold coverage were combined with capillary sequencing reads from a second related double-haploid Tübingen fish (from which the CHOR-1173 and CHORI-73 libraries were made), which contributed 12.2 million reads providing approximately 6-fold coverage from a mix of sequenced plasmid clones as well as BAC and fosmid end sequences. This use of data from double-haploid Tübingen fish resulted in less artificial haplotypic duplication than was found in previous WGS assemblies. A de Bruijn graph based algorithm called Fuzzypath[33] was used to assemble the Illumina reads into short sequence contigs; these contigs were then combined with the capillary reads using the Phusion assembler. This resulted in 119,136 contigs with an N50 size of 25 kb. Contigs were joined in supercontigs based on read pair information where the sizes of gaps were estimated using insert sizes of different lengths. There are 32,044 supercontigs in the WGS31 assembly with an N50 size of 614 kb.

1.6 Assembly integration Process

To provide Zv9, the ordered and oriented clone path derived from the mapping and sequencing/finishing approach was combined with WGS31. The WGS31 contigs, which were chained together into supercontigs based on clone end-pair information, were aligned to the clone path. Starting with an initial seed alignment of more than 95% identity, 65% coverage and less than 90% repeat content within a contig of at least 5kb size, the alignment was extended in both directions. This resulted in alignment chains with each chain consisting of a continuous block of WGS contigs. Contigs within a supercontig that aligned to a clone in their entirety were discarded, which thereby divided the supercontig into multiple chains that extended into clone path gaps.

The chains were used to fill those clone path gaps if neither cDNA nor genetic map marker data contradicted the placement. In cases where multiple alignments could be found for a gap, the aligned chains were processed in order of length and added if satisfactory to the criteria described above. In a subsequent step, chains without alignments to clone sequence were added into gaps according to information from cDNA and BAC/fosmid end sequence alignments and marker placement by giving genetic map markers the following priority ranking: SATmap, then HS, MGH and lastly T51 markers.

1.7 Future Maintenance and Improvement

With the release of Zv9, the Genome Reference Consortium[34] took on responsibility for the maintenance and improvement of the zebrafish genome sequence. Current work aims to replace WGS sequence with high quality finished clone sequence, to close remaining gaps in the genetic maps and to place yet un-localised sequence onto chromosomes, with further improved assemblies to follow as significant improvements are realised. User input is encouraged and can be provided at genomereference.org.

1.8 Other applications of the SATmap

1.8.1 Zebrafish sex and lethality linkage analyses

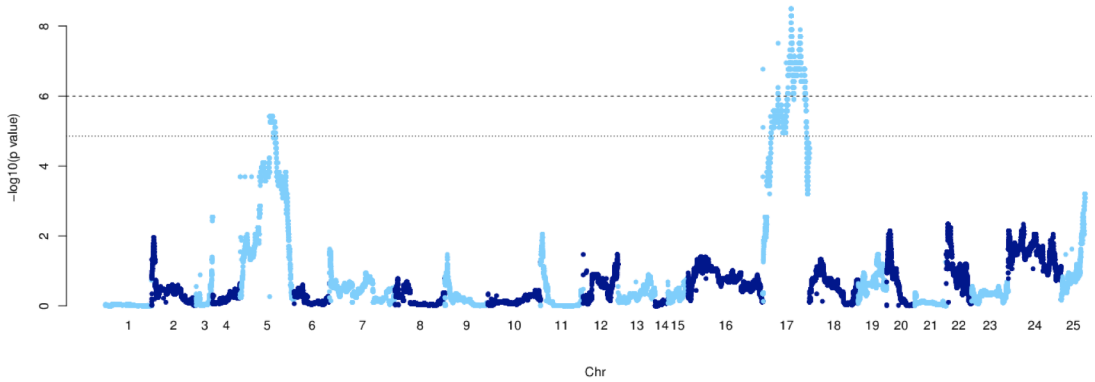
A total of 332 fish of confidently defined sex (235 males, 97 females) genotyped on 125,811 QC+ markers (after removing 12,628 due to > 5% missing genotypes) were taken forward for further analysis.

We performed two main analyses. Firstly, we used a two degrees of freedom (df) chi-square test of the three possible genotypes, Tü/Tü, Tü/AB and AB/AB, between males and females to identify loci influencing sex determination. The 2 df test is maximally powered when the underlying model, that is, dominant, additive or recessive is unknown. The genome-wide distribution of these statistics is not null distributed, unsurprising as the finite recombination in the dataset yields strong correlation among nearby markers, but is also likely influenced by differential rates of viability between sexes or other factors. To

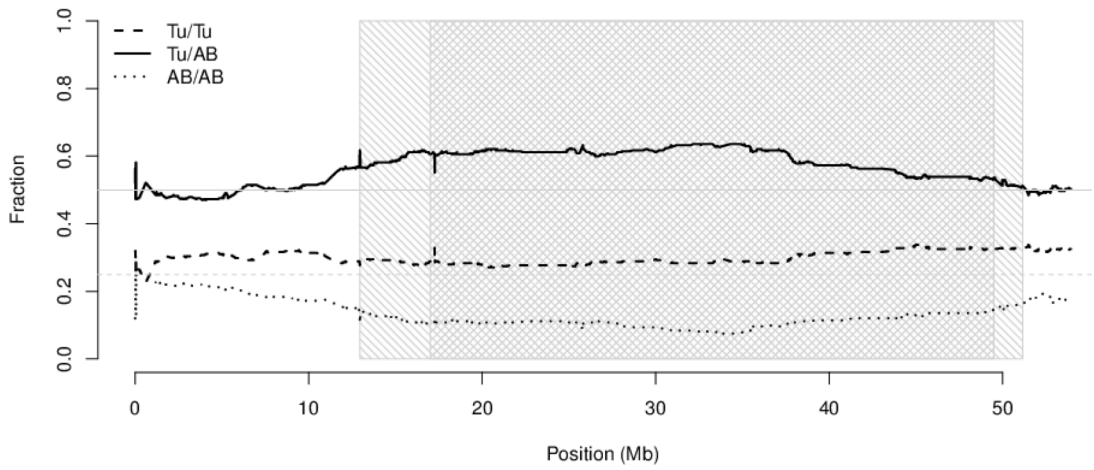
assess the significance of these tests we conducted 1000 permutations of sex, preserving the number of males and females, and carried out genome scans. This analysis established experiment-wide empirical p value thresholds. A p-value of 1.4×10^{-5} corresponds to an observation seen only once in 100 genome scans ($p = 0.01$) and 1×10^{-6} corresponds to an observation seen only once in 1000 genome scans ($p = 0.001$). Thus the maximal peak seen on chromosome 16 has a value of $p = 9.1 \times 10^{-7}$, permutation $p < 0.001$ (Figure 4).

When considering the three possible genotypes in all fish irrespective of sex, the vast majority of the genome conforms to the Mendelian transmission expectation of 25% Tü/Tü, 50% Tü/AB, 25% AB/AB. Excluding chromosomes 5 and 17 the median p value for a binomial test away from the expectation of 50% heterozygotes is 0.48 and is distributed as expected under the null (Supplementary Figure 5). By contrast, chromosomes 5 and 17 show marked departure from this ratio with maximal signal on chromosome 17 corresponding to a binomial p value = 3.2×10^{-9}. In each case AB/AB homozygotes are depleted compared to the expectation.

A



B



Supplementary Figure 5. Viability signal on chromosome 17
(A) Genome-wide p-values for binomial test of heterozygote frequency (i.e. excess lethality of either homozygote). Dotted and dashed lines correspond to empirical p-values of 0.01 and 0.001 as in Figure 3. (B) Genotype frequencies of all fish (male and female) on chromosome 17. The wide grey box corresponds to region with empirical p-value < 0.01, narrow cross-hatched box corresponds to region with p < 0.001.

The reference genome sequence and the SATmap will help to resolve an outstanding issue of the control of SD in zebrafish. Published observations of developing zebrafish show that all larvae initially develop oocytes and widespread apoptosis is seen in gonads of fish that ultimately become male[35,36]. Additionally, homozygous mutants of several DNA damage repair genes such as *brca2* mutants are 100% male and sterile[37]. Combining the

homozygous *brca2* mutation with a homozygous *tp53* mutation can restore a normal sex ratio, but not fertility[37]. Taken together, these previous results suggest a bias toward female differentiation in the germ line and a bias toward male differentiation in somatic tissue. Therefore, it is possible that one aspect of the zebrafish SD mechanism is cell-cell communication between the germ line and somatic cells. Indeed, in the region of the SD signal peak on chromosome 16 are a group of MHC genes and the HoxAb cluster, which may provide some candidate genes for control of SD events[38,39]. The significant peak of the signal we observe on chromosome 16 covers the same interval as a previously reported SD locus[11]. In contrast to previous reports, we did not obtain any evidence for SD signals on chromosome 5 or chromosome 4[11,12]. As described below, chromosome 16 bears a unique relationship in its high degree of conserved synteny with chromosome 19. One possible evolutionary pressure for this conserved synteny could be a whole chromosome inactivation mechanism, such as mammalian or *C. elegans* X chromosome inactivation[40]. In this case the SATmap data may suggest that SD in zebrafish is controlled by a mechanism that is not a simple genetic one.

1.8.2 Zebrafish strain variation

We sequenced each DH founder to over 40X depth, to establish the haplotypes for these individual AB and Tübingen strain zebrafish. As the SAT strain was generated from these G0 individuals these haplotypes represent the only two variants that can exist in an SAT strain individual. To identify the SNPs and indels in these two haplotypes we used a modified version of the 1000 genomes pipeline[41]. Reads were aligned to the Zv9 reference assembly using BWA[42] and SNPs were called by SAMtools mpileup[43], QCALL[44] and the GATK Unified Genotyper[45]. SNPs not called by all three callers were removed from the analysis, along with any SNP that did not pass a caller's standard filters. Additional SNPs were removed where the genotype quality was lower than 100 for GATK and lower than 50 for QCALL and SAMtools mpileup. Finally, SNPs within 10 bp of an indel (called by both SAMtools mpileup and Dindel) were removed.

By these parameters we find 6,995,534 SNPs between the two founders, about 50 SNPs per 10kb. The two founders share 104,600,000 bp (7.4% of genome).

Additional analyses excluding those shared regions revealed that Tü intra-strain variation is high at 29 +/- 44 SNPs per 10kb of genomic sequence (mean +/- standard deviation), while the inter-strain variation between the two individual SATMap founder fish was considerably higher at 84 +/- 37 SNPs per 10kb (Supplementary Table 6). Interestingly 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 1092 human diploid genomes[46], highlighting the high polymorphism rate in zebrafish.

Comparison		SNPs	Genome Length	Golden Path length	Density (%)	SNPs per 10kb
Intra strain	Tü/Ref	3,924,070	1,505,581,940	1,412,464,843	0.2778	29 +/- 44
Inter strain	Tü/AB*	6,995,534	1,505,581,940	1,307,864,843	0.5349	84 +/- 37
	AB/Ref	7,755,823	1,505,581,940	1,412,464,843	0.5491	

Supplementary Table 6. Variation

Intra strain (Tübingen) and Inter strain (Tübingen versus AB) SNP density comparisons. We removed 104,600,000 bp (7.4% of genome) where there were fewer than 5 SNPs due to monomorphic genomic regions shared between the individual AB and Tü double haploid fish used for this analysis.

1.8.3 Haploid genome assemblies of the AB and Tü founders

We assembled the Illumina paired end sequence data for the AB and Tübingen founders using Phusion2[47]. As we did not generate a series of mate pair libraries with which to scaffold this generated an assembly with hundreds of thousands of short contigs with an N50 of about 5kb (Supplementary Table 7). These assemblies should be of use for zebrafish researchers, especially in conjunction with the SAT line, the assemblies are available for BLAST search using the databases: 'AB (DHAB) Illumina de

novo assembly' and 'Tuebingen strain (DHTu2) Illumina de novo assembly' (www.sanger.ac.uk/cgi-bin/blast/submitblast/d_rerio).

Strain	Total size (bp)	% of Zv9	Contigs	N50
Tü	1,481,080,193	105	809,867	4,942
AB	1,325,654,336	94	699,243	5,010

Supplementary Table 7. Single Haplotype De Novo Assemblies

Results of de novo genome assembly using the same Illumina sequence data used to call the SNPs. For the genome size calculation the Zv9 genome size of 1,412,464,843bp was used.

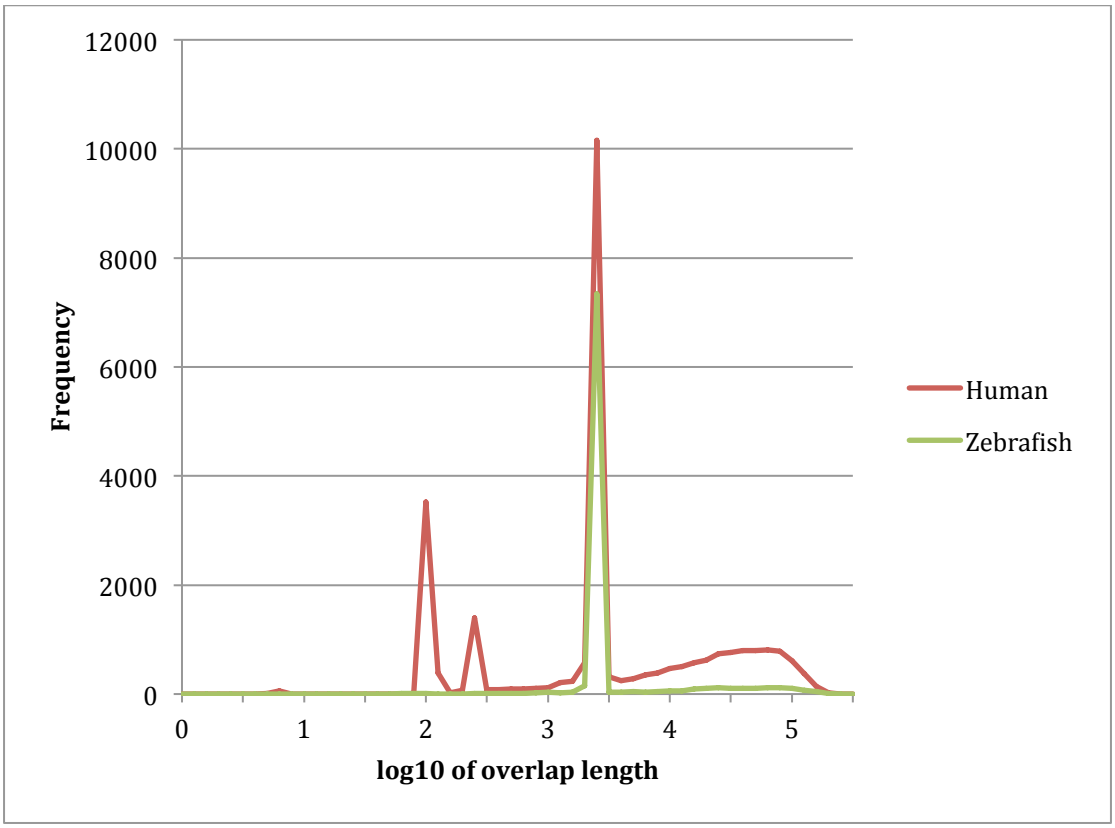
2 Assembly Characteristics

Zv9 was generated using the clone path ordered and oriented using SATmap, which allowed to successfully place and orient previously unattached or mis-placed FPCs, and make the zebrafish reference genome one of the most accurate and complete available. The resolution and marker density of the SATmap allowed FPCs and even individual clones to be ordered and oriented, providing a reliable genetic scaffold to stably anchor the Zv9 reference genome sequence. Through careful inspection of annotated gene order and application of the SATmap, many of the originally sequenced clones were found to represent haplotypic variants and were set aside..

Finally, to provide the best possible representation of the genome sequence, the resulting clone path was improved by insertions from a whole genome shotgun assembly (WGS) of a single DHTü individual, representing a single haploid genome (WGS31, CABZ00000000). The Zv9 assembly is therefore a hybrid of high-quality finished clone sequence (83%) and WGS sequence (17%) with a total size of 1.412 Gb. A previous study, using chromosome flow measurements in comparison with human chromosomes, estimated the size of the zebrafish genome to be 1.454 Gb[17]. The Zv9 assembly thus accounts for at least 97.1% of the genome and resides fully within the error margin of the chromosome flow estimate.

2.1 Clone Overlaps

The clone path is the backbone of the Zv9 assembly and is constantly being improved by the Genome Reference Consortium by assessing whether neighbouring clones do belong next to each other and feature a valid overlap. When assessing the quality of an assembly, measuring the sequence identity between overlapping clones provides key information. We compared clone overlaps in a recent zebrafish clone path (November 2011) to the clone path of the human GRCh37.p6 assembly (January 2012). This enabled analysis of the quality of overlapping clone sequence for 6.7 % of the current zebrafish and 13.3 % of the current human assemblies (Supplementary Figure 6, Supplementary Table 8). Note that in zebrafish, clone overlaps have not been cut back to improve overlap quality, but finished to either 2kb overlap or full insert length. The sequence represented in the clone overlaps averages at 99.29% sequence similarity in zebrafish and 99.93% in human.



Supplementary Figure 6. Zebrafish versus human overlap lengths

Shown is the distribution of overlap lengths in the zebrafish and human clone paths from November 2011 and January 2012, respectively. The overlaps peak at 2 kb for both species due to the commitment to finish overlaps to at least this length. There's an additional peak for 100 bp overlaps in human derived from a previous commitment in the early stages of the human genome project.

Species	Zebrafish	Human
Total Path Length	1,095,433,316 bp	2,853,072,938 bp
Number Overlaps	9,067	26,787
Total Overlap Length	73,834,763 bp (6.7%)	380,700,684 bp (13.3%)
Mean length	8,143 bp	14,212 bp
Median length	2,000 bp	2,000 bp
Total Length Overlap >2kb	59,305,373 bp	362,255,878 bp
Sequence Similarity	99.30% (99.98%*)	99.93%
Total Indel** Length	7,792,605 bp	1,481,315 bp
Indel Covered by Repeat Sequence	79.3%	31.9%

Supplementary Table 8. Zebrafish versus human overlap similarity

Analysis of sequence similarity in clone overlaps as an assembly quality measurement in zebrafish and human. For both species, the overlaps were designed to have a length of 2 kb if neighbouring clone sequence was available at the time of finishing, with a legacy of 100 bp overlaps from the early stages of the human genome project. However, 59.3 Mb or 80.3% of the zebrafish overlap sequence is found in overlaps longer than 2 kb. For human this figure is even higher: 362.3 Mb (95.2 %) of all overlaps exceed 2 kb.

*Sequence similarity for overlaps between CHORI-73 and/or CHORI-1073 clones only (derived from single DH fish).

**Insertion / Deletion (indel)

Due to the high level of sequence variation between different individual zebrafish, the sequence similarity was also measured for clone overlaps where both clones were derived from the DH fish (CHORI-73 and CHORI-1073 libraries). Clones derived from the single haploid individual should show no variation and their overlaps are therefore more suited to assess clone path and sequencing than those from the mixed libraries. Here, near-total identity (99.98%) in aligned sequence between overlaps could be found as expected from the boundary set by the overall clone sequencing and finishing quality (99.99%, see above 1.4.).

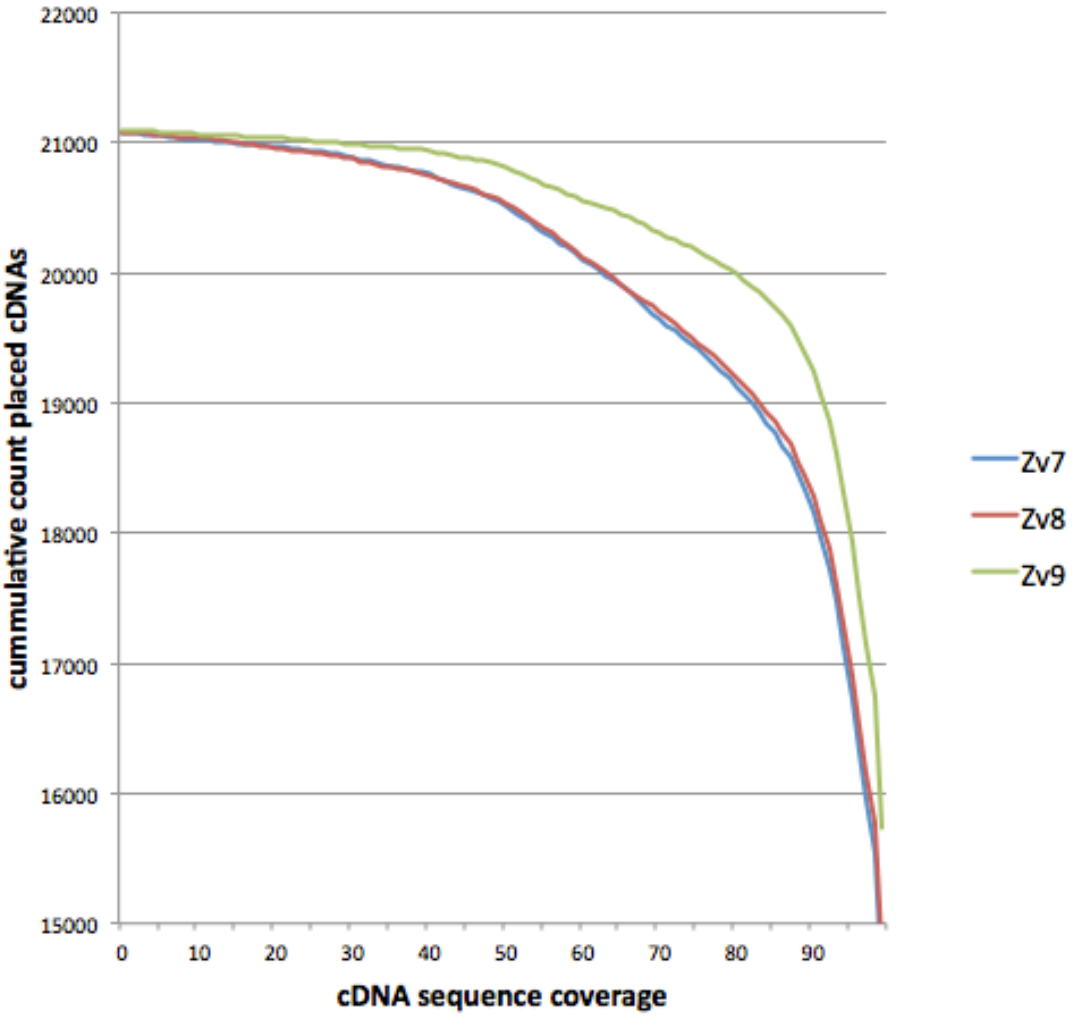
Apart from the aligned sequence, 7,792,605 bp of insertion / deletion (indel) sequence was identified in the zebrafish clone overlaps. About 80% of this sequence is covered in repeats, making it possible that overlaps differ due to

sequencing errors in tandem or simple repeat regions or recent transposition events, rather than erroneous clone order. It remains to be determined whether the overlaps showing high sequence variation are true overlaps. This task has been remitted to the Genome Reference Consortium as part of standard overlap checking and certification.

2.2 Placement of cDNA Sequence

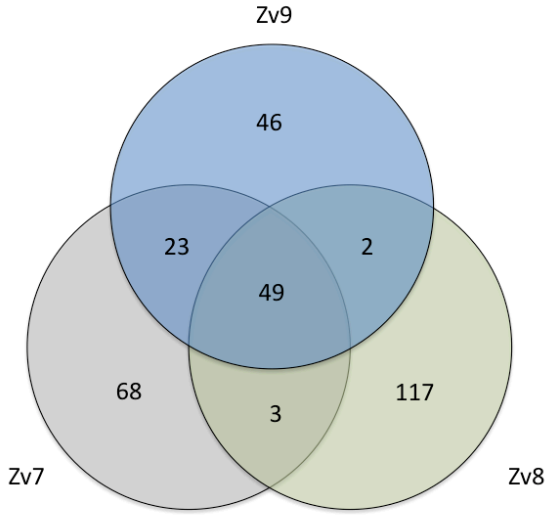
An aspiration of a complete genome assembly is the accurate localisation and annotation of every gene. To assess the alignment of existing cDNAs to the Zv9 assembly, we downloaded all available cDNA sequences from EMBL/Genbank (28,544 sequences on 13th January 2012). Of these, 77 chimeric cDNAs (as documented in the accessions) were removed, polyA tails were clipped and 7 cDNAs with less than 100bp length were excluded leaving a total of 28,460 cDNAs, which were clustered using CD-hit [48]. The requirements for forming clusters were set to a minimum of 97% sequence identity, with shorter sequence permitted to have up to 5 unaligned bases and coverage of long and short sequences of at least 10%. This resulted in 21,471 clusters. The longest cDNA of each cluster was taken as a representative for the following analyses. The cDNA cluster set was aligned to Zv7, Zv8 and Zv9 using Blat [49]. The cumulative number of cDNAs and their contiguous coverage in the three assemblies at an alignment sequence identity of at least 97% in Supplementary Figure 7. Whereas Zv7 and Zv8 are comparable, Zv9 clearly shows more cDNAs covered at a higher rate than previous assemblies. Assuming a contiguous coverage of at least 90% of any given cDNA with at least 97% sequence identity, 85% of all representative cDNAs can be placed in Zv7 and Zv8 and 90% in Zv9. The cDNA clusters with less than 10% coverage with at least 90% identity were regarded as not placed and are summarized Supplementary Figure 8. A total of 49 cDNAs are missing from all three assemblies (Supplementary Table 9). An additional 15 unplaced cDNAs are not of zebrafish origin but apparently mislabelled cDNAs from other species (Supplementary Table 10). Overall, the Zv9 assembly (120 unplaced cDNAs) covers 20 more presumed genes than the Zv7 assembly (143 unplaced cDNAs) and 51 more than the Zv8 assembly (171

unplaced cDNAs). This clearly demonstrates the improved genome coverage provided by Zv9. It also shows that very few (0.6%) presumed genes remain to be incorporated into the genome assembly. The Genome Reference Consortium is currently working on representing these missing regions by identifying and sequencing appropriate clones and integrating them into future assemblies.



Supplementary Figure 7. cDNA sequence coverage

Shown is a histogram of cumulative cDNA numbers and their coverage in the Zv7, Zv8 and Zv9 assemblies with a sequence identity of at least 97%.



Supplementary Figure 8. Missing cDNAs

Venn diagram listing the number of cDNAs missing (less than 10% of at least 90% using Blat) from the Zv7, Zv8 and Zv9 assemblies, respectively. Zv7 was built from a 2007 clone path integrated with the DH-only WGS29 assembly, Zv8 was build from a 2008 clone path integrated with WGS28, a mixed library capillary reads assembly, Zv9 was build from an April 2010 clone path integrated with WGS31, a DH capillary and NGS reads assembly.

Accession	Description
AB331779.1	Danio rerio mRNA for chemokine CCL-CUi, complete cds.
AF137534.1	Danio rerio MHC class I protein mRNA, complete cds.
AF155580.1	Danio rerio proteasome subunit beta 9B (PSMB9B) mRNA, complete cds.
AF520426.1	Danio rerio secreted frizzled-related protein mRNA, complete cds.
AY193830.1	Danio rerio clone YF-755 unknown mRNA.
AY882988.1	Danio rerio hypothetical protein mRNA, complete cds.
AY899291.1	Danio rerio rtn4 mRNA, complete cds.
BC056726.1	Danio rerio major histocompatibility complex class I UXA2 gene, mRNA (cDNA clone MGC:65799 IMAGE:6791792), complete cds.
BC066488.1	Danio rerio cDNA clone IMAGE:6525119, partial cds.
BC092777.1	Danio rerio zgc:110181, mRNA (cDNA clone MGC:110181 IMAGE:7292143), complete cds.
BC092889.1	Danio rerio zgc:110346, mRNA (cDNA clone MGC:110346 IMAGE:7403304), complete cds.
BC092892.1	Danio rerio zgc:110349, mRNA (cDNA clone MGC:110349 IMAGE:7403733),complete cds.
BC092908.1	Danio rerio cDNA clone IMAGE:7410112.
BC095216.1	Danio rerio cDNA clone IMAGE:7400184.
BC095829.1	Danio rerio hypothetical protein LOC553487, mRNA (cDNA clone IMAGE:7408970), partial cds.
BC097165.1	Danio rerio cDNA clone IMAGE:7448942.
BC097184.1	Danio rerio zgc:114126, mRNA (cDNA clone MGC:114126 IMAGE:7451061), complete cds.
BC097227.1	Danio rerio zgc:136302, mRNA (cDNA clone IMAGE:7430981), partial cds.
BC105745.1	Danio rerio cDNA clone IMAGE:7264666.
BC108067.1	Danio rerio zgc:123276, mRNA (cDNA clone MGC:123276 IMAGE:7899032), complete cds.
BC109437.1	Danio rerio zgc:123301, mRNA (cDNA clone MGC:123301 IMAGE:7903359), complete cds.
BC109438.1	Danio rerio zgc:123294, mRNA (cDNA clone MGC:123294 IMAGE:790229a2), complete cds.
BC110119.1	Danio rerio zgc:123292, mRNA (cDNA clone MGC:123292 IMAGE:7902107), complete cds.
BC110120.1	Danio rerio zgc:123290, mRNA (cDNA clone MGC:123290 IMAGE:7901502), complete cds.
BC116489.1	Danio rerio zgc:136302, mRNA (cDNA clone MGC:136302 IMAGE:7398190), complete cds.
BC122235.1	Danio rerio zgc:153316, mRNA (cDNA clone MGC:153316 IMAGE:7924101), complete cds.
BC122240.1	Danio rerio zgc:153322, mRNA (cDNA clone MGC:153322 IMAGE:7924986), complete cds.
BC122243.1	Danio rerio zgc:153325, mRNA (cDNA clone MGC:153325 IMAGE:7925735), complete cds.
BC122372.1	Danio rerio zgc:153668, mRNA (cDNA clone MGC:153668 IMAGE:7925164), complete cds.
BC122373.1	Danio rerio cDNA clone IMAGE:7925381, partial cds.
BC122398.1	Danio rerio zgc:153724, mRNA (cDNA clone MGC:153724 IMAGE:8145987), complete cds.
BC124257.1	Danio rerio zgc:153138, mRNA (cDNA clone MGC:153138 IMAGE:7401710), complete cds.

BC128862.1	Danio rerio major histocompatibility complex class I UDA gene, mRNA (cDNA clone MGC:158407 IMAGE:7068091), complete cds.
BC129341.1	Danio rerio cDNA clone IMAGE:7213476.
BC129505.1	Danio rerio zgc:158870, mRNA (cDNA clone MGC:158870 IMAGE:8148956), complete cds.
BC134215.1	Danio rerio zgc:163069, mRNA (cDNA clone MGC:163069 IMAGE:7898840), complete cds.
BC139526.1	Danio rerio cDNA clone IMAGE:7257886.
BC163951.1	Danio rerio zgc:110181, mRNA (cDNA clone MGC:191126 IMAGE:100059435), complete cds.
BC164293.1	Danio rerio zgc:153668, mRNA (cDNA clone MGC:191468 IMAGE:100059777), complete cds.
BC164331.1	Danio rerio zgc:110349, mRNA (cDNA clone MGC:191506 IMAGE:100059815), complete cds.
BC165267.1	Danio rerio zgc:153322, mRNA (cDNA clone MGC:192355 IMAGE:100060771), complete cds.
BC165541.1	Danio rerio zgc:110346, mRNA (cDNA clone MGC:192629 IMAGE:100061086), complete cds.
CU638748.1	Danio rerio cDNA, clone cssl:d0261
EF060285.1	Danio rerio liver patristacin (PASTN) mRNA, partial cds.
FJ643620.1	Danio rerio leukolectin mRNA, complete cds.
FN428721.1	Danio rerio partial mRNA for protein-tyrosine phosphatase zeta-b (ptprzb gene)
FN428740.1	Danio rerio partial mRNA for receptor-type tyrosine-protein phosphatase F Precursor a (ptprfa gene)
FN428742.1	Danio rerio partial mRNA for receptor-type tyrosine-protein phosphatase sigma b (ptprsb gene)
FN658836.1	Danio rerio partial mRNA for protein tyrosine phosphatase TCPTPb (ptpn2b gene), strain TL

Supplementary Table 9. Missing cDNAs

A list of the 49 Zebrafish cDNAs missing from the assemblies Zv7, Zv8 and Zv9 (i.e. less than 10% coverage at more than 90% sequence identity)

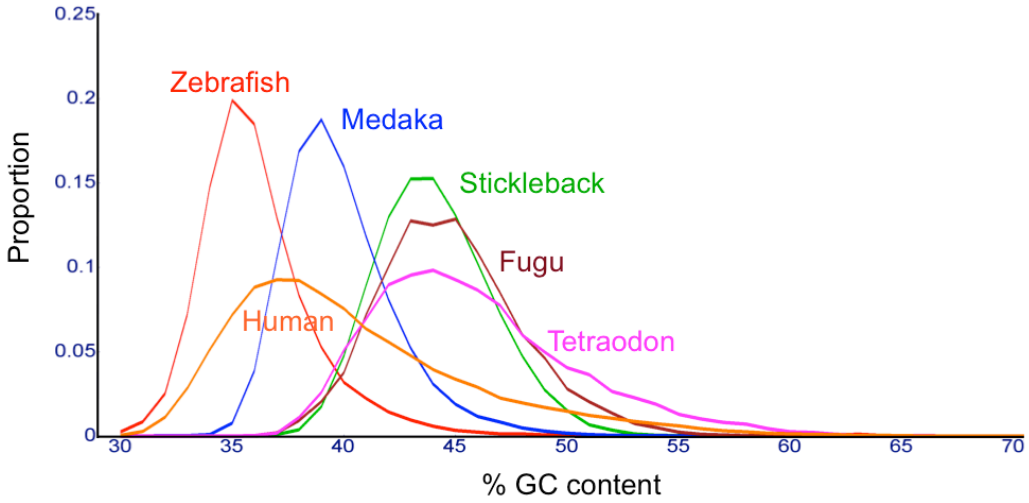
Accession	Current description	Presumed species
AF020527.1	Danio rerio brain-derived neurotrophic factor (BDNF) mRNA, partial cds.	Rattus norvegicus
AF151352.1	Danio rerio cripto mRNA, partial cds.	Homo sapiens
AY391448.1	Danio rerio phosphoglycerate mutase 1 (brain) (PGAM1) mRNA, complete cds.	Bos Taurus
AY423006.1	Danio rerio glutamate-ammonia ligase (GLUL) mRNA, complete cds.	Bos Taurus
AY423030.1	Danio rerio hypothetical protein FLJ20811 (FLJ20811) mRNA, complete cds.	Bos Taurus
AY960873.1	Danio rerio F-box and WD-40 domain protein FBXW14 (fbxw14) mRNA, complete cds.	Mus muscuslus or Artemia franciscana
BC086702.1	Danio rerio zgc:101551, mRNA (cDNA clone MGC:101551 IMAGE:7215732) , complete cds.	Artemia franciscana
BC107847.1	Danio rerio zgc:123289, mRNA (cDNA clone MGC:123289 IMAGE:7901443) , complete cds.	Artemia franciscana
BC110116.1	Danio rerio zgc:123298, mRNA (cDNA clone MGC:123298 IMAGE:7902776) , complete cds.	Artemia franciscana
BC124292.1	Danio rerio zgc:153264, mRNA (cDNA clone MGC:153264 IMAGE:7899298) , complete cds.	Artemia franciscana
BC124411.1	Danio rerio cDNA clone IMAGE:7902704	Artemia franciscana
BC134217.1	Danio rerio zgc:163074, mRNA (cDNA clone MGC:163074 IMAGE:7906924) , complete cds.	Artemia franciscana
BC164534.1	Danio rerio zgc:153264, mRNA (cDNA clone MGC:191709 IMAGE:100060018) , complete cds.	Artemia franciscana
CU458931.1	Danio rerio cDNA, clone cssl:d0169	Mus musculus
FJ984487.1	Danio rerio lectin-associated matrix protein (lamp-1) mRNA, partial cds.	Gallus gallus

Supplementary Table 10. Foreign cDNAs

A list of 15 cDNAs submitted to ENA/Genbank as being of zebrafish origin, but presumably derived from different species (Blast analysis).

2.3 GC Content

The zebrafish genome shows a uniform GC content, not majorly influenced by gene density or repeat content as reported in mammals (Supplementary Figure 9). This is in concordance with previous observations of a less marked GC compositional heterogeneity in poikilothermic compared to homoeothermic animals[50]. Fish genomes show an inverse correlation between genome size and average GC content[51] which is supported by our analyses. The zebrafish GC content is with 36.7% remarkably lower and more uniformly distributed than that of the teleost fish genomes present in Ensembl at the time of investigation (Stickleback, Medaka, *Tetraodon* and *Takifugu*) but similar to the closest relative of zebrafish sequenced so far, carp, *Cyprinus carpio*, 36.8% with a genome size comparable to zebrafish[52]. The distribution of the GC content over the individual linkage groups is illustrated in the supplementary chromosome graphs.



Supplementary Figure 9. GC content
Average GC content calculated over 20kb windows, data from Ensembl version 59. The histogram shows distribution of GC content in 20 kb windows. Windows were discarded if more than 25% of them consisted of sequence gaps, or if they were smaller than 20 kb due to falling at the end of a top-level region.

The GC content showed no influence by repeat distribution. Sequence both unmasked and masked by RepeatMasker[9] totals at 36.7% GC (Type I

transposons total 37.4%, Type II transposons total 35.5%, Satellites total 36.9%). Similar to the human genome, zebrafish transcribed sequence has an elevated GC content (46.3%), which may be caused by selection to maintain nucleosome-positioning sequences[53].

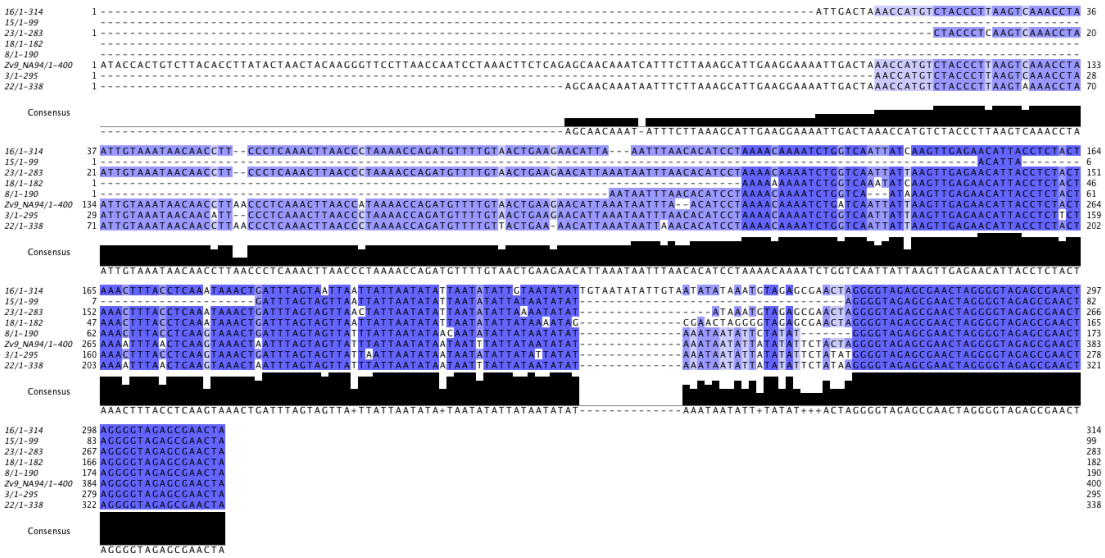
Creating WGS31 and subsequently Zv9 led to an accumulation of single sequence reads (478 Mb not included in WGS31) and sequence contigs (77 Mb from WGS31 not included in Zv9) that were rejected on the basis of poor quality. These sequences were included in the genome assessments such as GC content and repeat distribution to account for possibly missing features. The GC content of these sequence collections differs from Zv9. The 478 Mb non-WGS31 sequence has a GC content of 41% and the 77 Mb non-Zv9 sequence a GC content of 45%.

2.4 Telomeres, Centromeres and Satellite Repeats

To localise the position of centromeres and telomeres reported markers from 6 investigations[17,21,22,54-56] were collated and identified in the Zv9 assembly (Supplementary chromosome graphs).

In 16 chromosomes, near-telomeric markers were found associated with both ends of the physical sequence. The remaining 9 chromosomes contained near-telomeric markers on one end of the sequence. Zebrafish telomeres have the sequence TTAGGG and are localised exclusively at the end of chromosomes[57]. In Zv9, the telomere sequence was found at or nearby the end of chromosomes 3, 15, 17 and 22. Interestingly, the telomeric repeats were preceded by a 100% conserved and yet un-described tandem repeat (monomer sequence GTAGAGCGAACTAGGG) of varying length that shows similarity to the Merlin-1_Aplcal transposable element (TE). TEs of the Merlin class are yet unreported in zebrafish[58]. This novel subtelomeric repeat could be identified in 6 further locations (chromosomes 8, 16, 18, 21 and 23, and Scaffold Zv9_NA94) at the end of sequence components suggesting these might be incorrectly placed away from the telomeres. The subtelomeric repeat is immediately preceded by a highly conserved AT-rich sequence,

again not yet reported in Repbase (Supplementary Figure 10). So far it has been reported that subtelomeric repeats vary greatly in vertebrates even between different chromosomes of the same species and are highly polymorphic. This may prevent recombination between non-homologuous chromosomes and increase their stability[59].



Supplementary Figure 10. Alignment of zebrafish subtelomeric regions
The alignment shows a conserved AT-rich region yet unknown to Repbase, followed by a 100% conserved region with similarity to the Merlin-1_Aplcal element from Aplysia. This region is tandemly repeated between 14 and 64 times and immediately followed by the telomeric TTAGGG repeat. The alignment above was arbitrarily restricted to show 3 units of this subtelomeric repeat.

The analysis of the published near-centromeric markers is limited by the low precision of feature location through metaphase chromosome hybridisation. Nevertheless, potential centromeric regions could be identified in all 25 chromosomes, ranging in width up to 11 Mb as determined by the distance of the marker placements.

Centromeres generally consist of satellite repeats hence an analysis of the distribution of known satellite repeats in zebrafish might help in localising these chromosome features. It has been reported that the sequence

submitted as AF175725 localised to all zebrafish centromeres in FISH experiments[60], yet it is not readily identifiable in the Zv9 assembly, other than being evenly distributed over the chromosomes at low level (0.006% coverage of chromosomes, 0.016% coverage of unlocalised scaffolds). Since the AF172725 is highly similar to a part of the LOOPERN2_DR type II DNA transposons this probably reflects the distribution of LOOPERN2_DR in the genome, while suggesting that the real centromeres have so far escaped sequencing.

Also reported to be associated with centromeres are AT-rich satellite repeats[60]. These repeats have been included in the repbase repeat libraries as BRSATI and DRSATII. In Zv9, BRSATI can be found arranged in one to three pronounced clusters per chromosome, with each cluster being located close to the presumed centromere position. Exceptions are chromosomes 1 and 5, where no pronounced cluster is present, and chromosomes 2, 14 and 24 where one of the two clusters is located towards the distal end of the long arm. Discounting those three distal clusters, all other clusters can be found within less than an average of 4 Mb distance from centromeric marker. BRSATI covers 0.10% of the sequence allocated in chromosomes but is tenfold enriched in unplaced scaffolds (1.04%), possibly aiding the future placement of these scaffolds. DRSATII, on the other hand, is more or less evenly distributed over the chromosome sequence, without any association with the presumed centromere positions.

The analysis of distribution of other known satellite repeats using RepeatMasker with the RepBase *Danio* library from October 2011 revealed several repeats were being found exclusively towards the telomeres (MSAT-2, MSAT-3, MSAT-5, SAT-3), SAT-4 being identified towards the telomeres with additional near centromeric appearances on 8 chromosomes (and being rather evenly distributed over chromosome 22) and MOSAT, the satellite with the highest coverage in Zv9 (0.59% of genome sequence), being evenly distributed over all chromosomes.

MOSAT-2 shows a unique pattern found nearly exclusively on the long arm of chromosome 4, and also in un-placed scaffolds, suggesting possible future placement. SAT-2 complements the pattern of MOSAT-2 in that it is uniformly distributed over all chromosomes, but missing from the long arm of chromosome 4.

2.5 Gene Structure

In contrast to mammalian genomes, the sequenced teleost fish genomes, Stickleback, Medaka, *Takifugu* and *Tetraodon* possess compact gene structures with short introns. To determine how these structures differ in zebrafish, a member of the otocephala (or otomorph) taxon, we collected and analysed the longest protein-coding transcript of each gene in all these species. The analysis was mainly restricted to the coding part of the representative transcripts only, to avoid bias caused by different level of quality in untranslated region (UTR) annotation between species. Note both the UTR and general annotation differs significantly between the sequenced fish genomes. This is due to the fact that differing amounts of supporting evidence are available for the automated annotation in Ensembl. Where there is no or only little own-species cDNA evidence available, the Ensembl annotation is based on comparisons with coding sequence from other species, thereby omitting UTRs. For the zebrafish genome, as with the mouse and human genomes, the Havana group at the Sanger Institute carries out manual annotation. In manual annotation special attention is paid to splice variant and UTR annotation. The resulting genes are then merged with the Ensembl genes, improving the respective gene sets significantly.

The median transcript size without UTRs in all species investigated is very similar (1184 +/- 73 bp), as are the median coding exon sizes (121 +/- 2 bp) (Supplementary Table 11). The differing range in gene span between species is caused by differences in intron sizes, ranging from about 150 bp median length in the teleosts to nearly 1000 bp in zebrafish and 1500 bp in human. It has already been observed in human that first introns are significantly longer than other introns[61]. In the mouse, human and teleost genomes, we could

also show that first introns of the representative transcripts were at least twice as long as last introns. Zebrafish, however, generally has first and last introns of similar size, with the median last intron size in zebrafish exceeding the median length of the mammalian last introns. This increase in size is not caused by a significant increase in repeat content compared to the first introns. It is possible that the increased length of first introns facilitates increased regulatory complexity[61].

	zebrafish	fugu	tetraodon	medaka	stickle	mouse	human
protein-coding genes							
count	26039	18523	19602	19686	20787	22667	20031
median length bp	12342	4116	3229	6137	4726	15169	24824
genome coverage %	51.02	37.25	34.59	29.77	41.66	37.45	42.19
representative transcripts incl UTRs							
median length bp	1741	1311	1239	1242	1371	2298	2687
genome coverage %	4.15	8.07	8.65	3.66	7.78	2.35	2.11
representative transcripts CDS only							
median length bp	1185	1287	1143	1122	1134	1098	1209
genome coverage %	2.91	7.97	8.28	3.43	6.93	1.28	1.06
total CDS coverage Mb	41.1	31.3	29.7	29.8	32	34.8	32.9
exons in representative transcripts							
median count	7	8	8	7	7	6	7
median length in bp	124	122	119	119	120	126	124
median length first exon	105	114	111	106	106	105	103
median length last exon	144	134	131	131	131	144	144
introns in representative transcripts							
median length in bp	980	142	118	246	215	1275	1501
median length first intron	1594	356	264	530.5	479.5	2692	3335
repeat coverage %	52.32	6.99	7.09	4.93	10.73	35.58	45.72
median length last intron	1424	140	111	260	223	1148	1367
repeat coverage %	57.68	8.34	9.25	5.43	12.53	34.86	44.28

Supplementary Table 11. Gene structure statistics

These statistics are based on the gene annotation found in Ensembl 63. The longest protein-coding transcript of each protein-coding gene was chosen as a representative transcript. Note that UTR data is dependent on the availability of supporting evidence and the nature of the annotation process. First and last exons/introns (without UTRs) were calculated for representative transcripts with at least 3 introns. The intron repeat coverage was calculated as total repeat coverage. Genome coverage is given relative to the golden path length of the respective assembly.

2.6 Repeats

The Zv9 sequence has been repeat masked using 'tandem repeats finder'[62], DUST[63] and RepeatMasker[9] using RepBase[64] *Danio* library from March 2010.

The three single most abundant TEs in zebrafish are the SINE HE1_DR1 (109,922 instances, 1.88 % genome coverage), the non-autonomous Type II DNA transposon TDR18 (72,927 / 1.58 %), and the non-autonomous unclassified transposon TE-X-5_DR (106,567 / 1.38 %) (Supplementary Table 12).

These findings are in disagreement with the earlier reports of the tRNA-derived SINE element DANA apparently covering 10% of the genome[65]. In Zv9, RepeatMasker revealed only 0.14% of the genome sequence to be covered by this element. When extending our searches for DANA to all zebrafish sequences that were rejected for building either the combined Illumina and whole genomes shotgun assembly WGS31 or Zv9 (see above), we did not encounter the potentially missing sequence. In fact, the DANA content of rejected sequences was lower than that of Zv9.

We further investigated the positioning of repeats in intergenic and intronic regions to identify a possible bias (Supplementary Table 13). Simple repeats are slightly over-proportional in intronic regions, whereas Satellite repeats and Type I LTR and non-LTR TEs are underrepresented in intronic regions. Looking in further detail, it turns out that Tx1 transposons are tenfold less frequent in introns. Conversely, SINE2/tRNA repeats and Kolobok repeats are less frequent in intergenic regions. We could not detect any bias in direction for intronic TEs when looking at the different TE classes. When looking at family level, the only notable exception is the Type I LINE element I, which is located on the same strand ten times more often than on the opposite strand.

One major sort of repetitive element are the set of ribosomal RNA (rRNA) genes. Indeed rRNA clusters have been reported elsewhere [57] but we have found them difficult to accurately place in the assembled genome.

Repeat type	Class/superfamily	Occurrence	Coverage bp	Coverage %	Total coverage %
Simple repeats		2072975	90907462	6.436	6.44
Tandem repeats		1179751	150883666	10.682	10.68
Satellite repeats	SAT	57288	10760259	0.76	0.90
	MSAT	6942	1923376	0.136	
Type I Transposons	LINE/CR1	90707	28363335	1.985	10.61
	LINE/L1	10740	4546141	0.312	
	LINE/RTE	8778	2735632	0.191	
	LINE/I	4437	1404969	0.098	
	LTR/Gypsy	56138	21346971	1.353	
	LTR/ERV1	60878	13200650	0.915	
	LTR/LTR	17170	7548002	0.527	
	LTR/Copia	10293	3974894	0.273	
	LTR/BEL	4585	2845601	0.167	
	LTR/ERV2	1220	494447	0.035	
	LTR/Endogenous	1365	438649	0.029	
	LTR/BHIKHARI_I	71	182403	0.012	
	LTR/ERV3	243	47776	0.003	
	non-LTR/DIRS	46806	29254263	2.061	
	non-LTR	15984	3725066	0.263	
	non-LTR/Tx1	9197	2092679	0.144	
	non-LTR/Nimb	599	281628	0.019	
	non-LTR/Hero	308	128465	0.008	
	non-LTR/R2	28	13746	0	
	SINE	126287	30373852	2.149	
	SINE2/tRNA	6091	956290	0.067	
Type II Transposons	DNA	1075315	274101019	19.312	38.51
	hAT	551400	124222405	8.738	
	EnSpm	190549	31126103	2.193	
	Mariner/Tc1	96965	31019366	2.186	
	Harbinger	100115	29414042	2.072	
	Kolobok	107371	25602277	1.804	
	Helitron	54463	12075801	0.852	
	piggyback	30190	9224709	0.650	
	Polinton	34401	5048013	0.356	
	MuDR	4750	2220915	0.157	
	DNA/TcMar-Tc1	6742	1924542	0.135	
	P	1805	477278	0.033	
	ISL2EU	2408	283589	0.019	
	DNA/Chapaev	9	687	0	
Transposons		167078	28908390	2.043	2.04
Unclassified	LRS_DR	679	677694	0.047	0.05

Supplementary Table 12. Repeat elements

Overview of repeat elements found in the zebrafish genome assembly Zv9. Note that the coverage is not additive, i.e. repeats can overlap each other.

Repeat Type	Intergenic			Intronic			Rate ratio	Coverage ratio
	count	rate per Mb	coverage %	count	rate per Mb	coverage %		
Simple repeats	1,127,830	1,466	0.07	918,079	1,563	0.07	1.07	1.03
Satellite repeats	42,106	55	0.01	221,28	38	0.01	0.69	0.67
Tandem repeats	664,428	864	0.12	504,807	859	0.09	0.99	0.77
Type I/LINE	65,174	85	0.03	493,62	84	0.02	0.99	0.82
Type I /LTR	95,031	124	0.04	569,01	97	0.03	0.78	0.59
Type I /SINE	77,451	101	0.02	542,33	92	0.02	0.92	0.92
Type I /non-LTR	49,534	64	0.03	231,18	39	0.02	0.61	0.6
Type II /DNA	1,220,333	1,586	0.38	1,021,333	1,739	0.42	1.1	1.09
Transposons	92,580	120	0.02	734,46	125	0.02	1.04	0.99
unclassified	628	1	0	53	0	0	0.11	0.14

Supplementary Table 13. Repeat location

Location of repeats segregated by repeat type and intergenic or intronic location.

2.6.1 DNA Transposons

All of the previously described Type II superfamilies are extensively represented in zebrafish, with the exception of Merlin, which we could only find in the novel subtelomeric repeat described above.

Zebrafish DNA transposons are divided into 14 superfamilies with 401 repeat families in total, covering 38% of the zebrafish genome. The DNA and hAT superfamilies are the most abundant and diverse in the zebrafish genome. The DNA superfamily has 194 members with over 1 million occurrences, the hAT superfamily 107 members with nearly half a million occurrences, together covering 28% of the sequenced genome.

The abundance of DNA transposons is of particular interest since they are reported to cause chromosome rearrangements through alternative transposition and recombination[66], which is consistent with earlier reports of zebrafish showing a lack of long-range synteny with human genes compared to Medaka, *Tetraodon* and *Takifugu*[67,68]. With assembly Zv9, however, we find a comparable degree of synteny conservation in zebrafish and the other sequenced teleost fish. The zebrafish genome is no more rearranged than other fish genomes when considering various measures of intra-chromosomal gene-order or gene-linkage conservation (Supplementary Figures 14 and 15). This contradicts previous studies, where the apparent rearrangement was likely due to the poor quality of previous versions of the assembly[68]. In contrast, the distribution of ohnologs among chromosomes shows that inter-chromosomal rearrangements are more frequent in the zebrafish lineage (Supplementary Figure 16).

2.6.2 Retrotransposons

It has been suggested that the deleterious effect of homologous recombination provides strong negative selection for restraining retrotransposon copy number in fish and possibly favoured their divergence[69]. The repeat structure of the zebrafish genome is consistent with this notion. Despite retrotransposons comprising 10% of the zebrafish genome, less than one third of the DNA transposon coverage, there are a large variety of retrotransposable elements. The most diverse type I group are the LTR TEs, represented by nearly 500 different families from 9 superfamilies (BEL, Endogenous, Copia, ERV1, ERV2, ERV3, Gypsy, LTR, BHIKHARI_I), with the Gypsy superfamily being remarkably extended (more than 160 members covering 1.4 % of the genome sequence). The most abundant type I superfamilies are DIRS (non-LTR class, 2.1% genome coverage, 22 families, 47,000 occurrences) and CR1 (LINE class, 2 % genome coverage, 43 families, 91,000 occurrences).

Type I transposons in zebrafish comprise recently or still assumed active families that have had a great impact on the zebrafish genetic landscape.

Indeed, there are more than 600 genes that that likely originate from retrotransposition, 440 of which are active genes [70].

2.6.3 Pseudogenes

The zebrafish reference genome contains comparatively few pseudogenes, a total of 154 manually annotated pseudogenes compared to 13,340 pseudogenes in the human genome (Supplementary Table 14). This lack of zebrafish pseudogenes may be related to the balance of Type II TEs relative to Type I retrotransposable elements. Nearly 40% of the human genome comprises LINE and SINE elements, which derive from retrotransposons. The majority of processed pseudogenes, i.e. those with no apparent intronic sequence, are thought to arise from retrotransposition. Consistent with this notion, in the human reference genome 75% of all pseudogenes are processed and 22% are unprocessed. By contrast, in zebrafish 14% of pseudogenes are processed and 75% are not. Several zebrafish processed-pseudogenes are flanked by Type I, retrotransposon elements, indicating that retrotransposon activity has modified the zebrafish genome, but has not led to the expansions seen in mammalian genomes.

	Human Pseudogenes		Zebrafish Pseudogenes	
Biotype	**Count**	**%**	**Count**	**%**
IG_pseudogene	161	1.2	9	5.8
Polymorphic_pseudogene	27	0.2	4	0.3
Processed_pseudogene	9992	74.9	21	13.6
Pseudogene	7	0.1	5	3.3
TR_pseudogene	44	0.3		
Unitary_pseudogene	159	1.2		
Unprocessed_pseudogene	2950	22.1	115	74.7
Total	13,340		154	

Supplementary Table 14. Human and zebrafish pseudogenes

Human versus Zebrafish manually annotated pseudogene comparison based on zebrafish Vega version 47. The biotype classification is described at (vega.sanger.ac.uk/info/about/gene_and_transcript_types.html).

2.6.4 Recent Transposition

Transposable elements are of special interest because they are involved in genomic reorganisation and speciation events helping to explain teleost evolution [71], but also because they have become a powerful vector for genetic modification, as is evident by the widespread use of Tol2 for zebrafish transgenesis [72]. Despite several attempts, no currently active TEs have been identified yet, so the current genetic techniques rely on the modification and reactivation of ancient elements [73,74].

2.6.5 Clone overlaps

As DNA transposition occurs via a cut-and-paste mechanism one possibility is that recent transposition would produce rearrangements in the genome that would be evident as indels. To examine this we considered all clone overlaps within a current clone path searching for indels possibly caused by the insertion or excision of transposable elements. Since the clones are derived from different libraries, the overlaps investigated were separated into two groups. We examined clones taken from the CHORI-73 and CHORI-1073 libraries, which were made from a single Tübingen double haploid individual, and we examined clones originating from any other library.

Clones derived from the single haploid individual should show no variation, however their overlaps contained 27kb (4.3%) of indel sequence. The vast majority, 91% of this indel sequence, is covered by repeat sequences, with more than 80% of these indels caused by simple and tandem repeats, which are likely to have arisen from cloning or sequencing errors rather than natural variation. None of the indels observed between clones from the double haploid libraries satisfied our criteria for possible recent transposition.

Examining the non-double haploid clones, 80% of the observed indels were covered in repeats and the distribution of the repeats was found to be similar to the genome wide repeat distribution. To identify possible recent transposition events, the indels were searched for the presence of repeats of at least 95% of their expected length, with 95% of the repeat falling into the

indel. This resulted in the identification of 854 putative recent transposition events, caused by 211 different TEs (Supplementary Table 15).

	zebrafish DH only		zebrafish mixed libraries		human all libraries	
total length of overlaps	6,269,414 bp		67,712,683 bp		380,700,684 bp	
Indels total	27,006 bp (0.43 %)		7,765,599 (11.47 %)		1,481,315 bp (0.39 %)	
putative transposition events *	**no.**	**total indel seq. cover**	**no.**	**total indel seq. cover**	**no.**	**total indel seq. cover**
Type I/LINE	0	0 %	10	0.60 %	0	0 %
Type I/SINE	0	0 %	12	0.07 %	98	2.04 %
Type I /LTR	0	0 %	98	0.13 %	0	0 %
Type I/non-LTR	0	0 %	62	3.85 %	na	na
Type II	0	0 %	672	10.58 %	0	0 %
other	0	0 %	0	0 %	2	0.29 %
indel sequence without repeats	2,496 bp	9.24%	1,606,745 bp	20.69 %	472,326 bp	31.89 %

Supplementary Table 15. Clone overlap variation

Variation in overlaps between clones from the double haploid Tübingen libraries or between clones from at least one other library and their repeat content and human GRCh37.p6 for comparison. Total lengths given at the top relate to the sequence length of the golden path part of the overlap (i.e. one clone only, percentages given for indels total are in relation to total overlap length). The table lists the number of occurrences plus the sequence coverage of the respective TEs compared to the total indel length. Any given portion of indel sequence is only classified as having one repeat type in the order tandem repeats, then simple repeats and finally TEs, preferring the longest TE present. The total length of clone sequence in Zv9 comprises 1.095 Mb.
*(full size TEs in indels)

An identical repeat analysis in human clone overlap indels identifies AluSx, AluY, SVA_E and SVA_F, previously reported to still be active in human [75], supporting the suitability of this method for detection of recent transposition events.

2.6.6 Pairwise distance

Another approach to assess recent transposition in zebrafish is to investigate the pairwise distance of repeat instances. An average pairwise distance of

less than 2% is reported as a hint for recent transposon activity in *Xenopus tropicalis* [76]. To this end, all transposons within 10% length difference compared to their repbase library consensus sequence were extracted from the genome and aligned to their family members using the MUSCLE multiple sequence aligner [77]. The selection was limited to only those repeats found on finished clones to avoid noise arising from possibly mis-assembled WGS sequence. The pairwise distance was calculated using DNADIST with Jukes-Cantor and Kimura parameters [78] (Supplementary Table 16).

Class	All	Dist-Br	Dist-Ol	Br-Ol	Dist	Br	Ol	Total
Transposable		2						2
DNA	26	10	8	59	4	8	27	142
LINE	3	1	3	1	20	1		29
LTR	1	3	2	8	189	3	18	224
SINE				2			1	3
non-LTR	5	2		1	2	5	4	19
LRS_DR						1		1

Supplementary Table 16. Possible recent transposition

Overall comparison of possible recent transposition events from pairwise distance (Dist), Breakdancer (Br) and overlap (Ol) analysis. All = Repeat Elements identified as possibly recently transposed in all three analyses, Dist-Br= elements identified as such in the Dist and Br analyses, Dist-Ol = same for Dist and Ol analyses, Br-Ol = same for Br and Ol, rest identified in single analysis only.

2.6.7 Alignment to NGS reads

In addition we assessed possible recent transposition by comparing Zv9 to the sequences obtained from Illumina runs on a separate double-haploid Tübingen fish using Breakdancer [79].

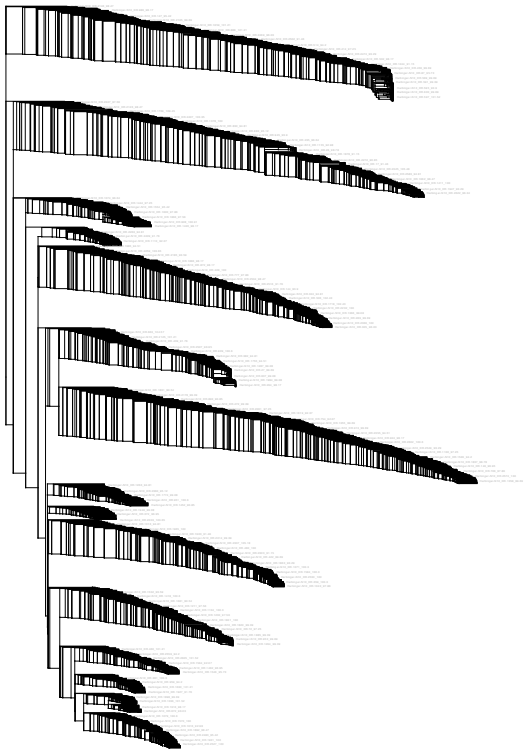
After aligning the reads to Zv9, Breakdancer analysis was performed to detect insertions in the reference relative to the reads. The corresponding reference regions were searched for TEs, requiring 75% of the TE present within the

reference region, and in return 75% of the reference region covered by the TE. We found 2041 instances matching these requirements.

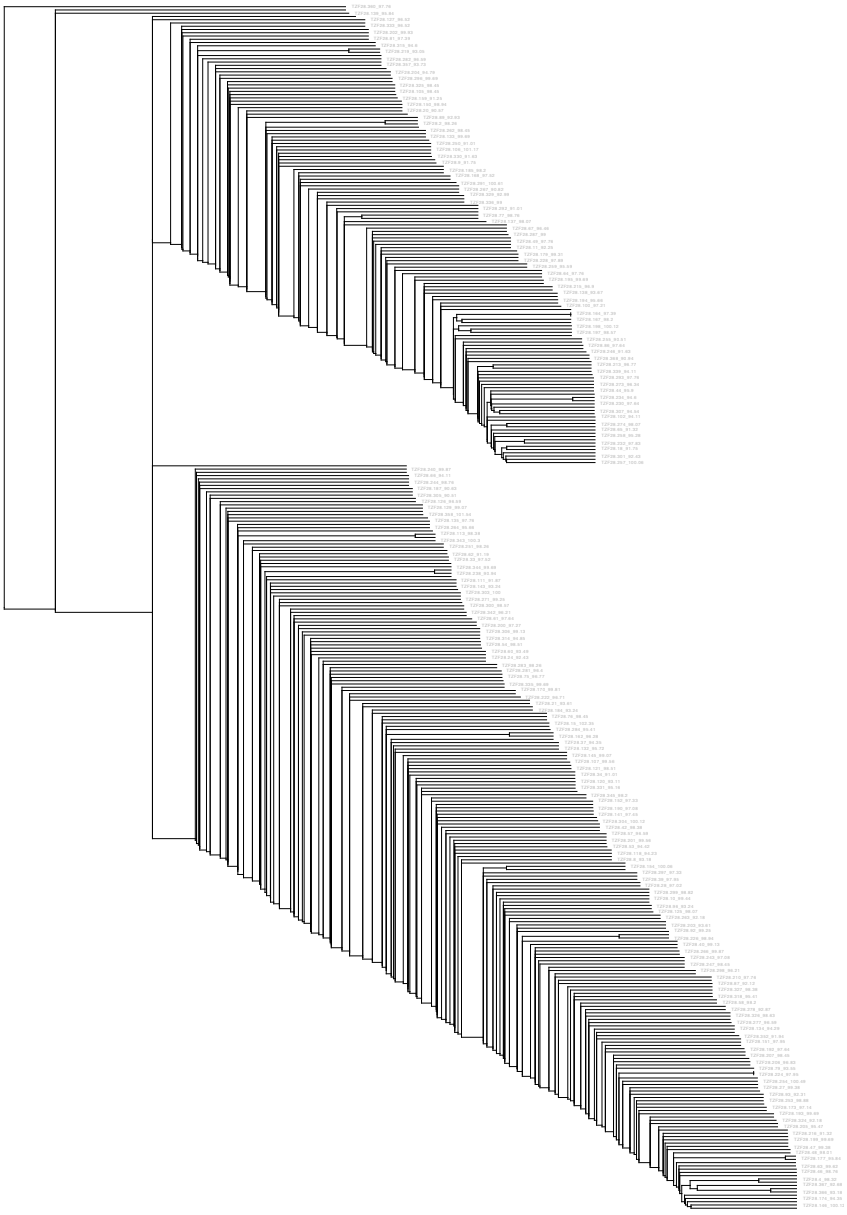
Our analyses identified TE instances and their pairwise distances and suggests that recently active elements may be present for all classes of transposable elements, roughly corresponding to their overall distribution in the genome, apart from an overrepresentation of LTR elements (Supplementary Table 16). These analyses provide an excellent starting point for further investigations into recently or currently active TEs in zebrafish.

2.6.8 Bursts in TE spread

Using the data generated from pairwise distance comparison (see above) to generate trees with MUSCLE (2 iterations), we observed patterns that can be interpreted as initial bursts of transposon activity after a new acquisition until all copies are inactivated (Supplementary Figures 11 and 12).



Supplementary Figure 11. Phylogenetic tree of the Harbinger-N10_DR Harbinger-N10_DR is a Type II Harbinger element and shown is a pairwise distance tree. The distribution suggests 15 bursts of activity before fossilisation.



Supplementary Figure 12. Phylogenetic tree of the TZF28

TZF28 is a Type II Mariner/Tc1 element and shown is a pairwise distance tree. This repeat has been reported as active [80] yet we could not detect anything that would support this notion. The distribution suggests two bursts of activity before fossilisation.

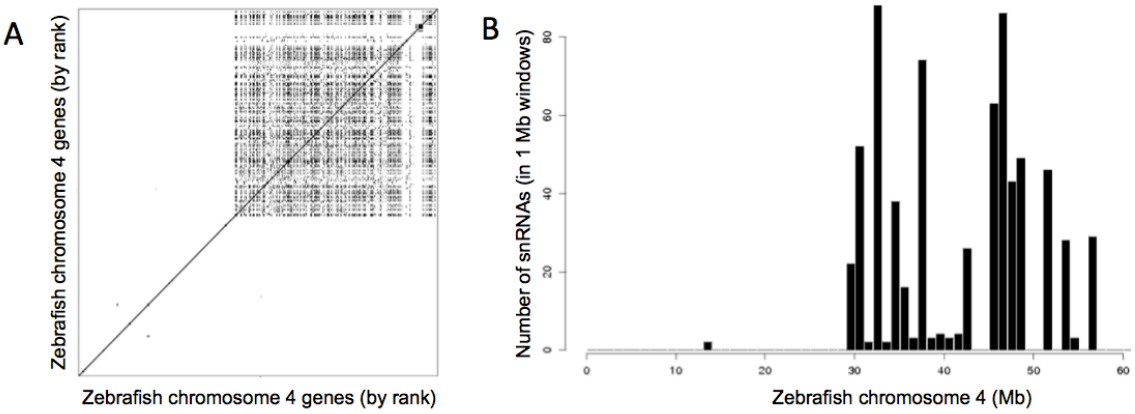
2.7 Chromosome Landscapes

To visualise the distribution of selected features, we generated chromosome landscape graphs (see Chromosome Graphs). These graphs feature exon

and repeat distribution, sequence composition, marker placement, GC content, gap density and reported near-centromeric markers[6,13,14,38-40]. The analyses have been performed over 100 kb windows.

2.8 Chromosome 4

Chromosome 4 or LG4 (consistent with a community agreement reached at the European Zebrafish Meeting in Paris, 2003, we have translated linkage group numbers directly into chromosome numbers, e.g. linkage group 1 = chromosome 1) was initially identified as chromosome 3. As described in the main section and visible in the chromosomal landscape graphs chromosome 4 has a special structure regarding repeat and gene content, as well as gene orthology and synteny (Supplementary Figure 13).



Supplementary Figure 13. Chromosome 4 gene duplications
A: The distribution of duplicated genes on zebrafish chromosome 4 is shown in a classical dot-plot representation. A black dot is indicated each time a gene on the X-axis (ranked by order along the chromosome, from telomere to telomere) faces either itself or a duplicate copy on the Y-axis. The distribution shows a striking compartmentalisation, with the second half (approximately from 30 to 60 Mb) containing a high density of locally duplicated genes (NLR-like proteins and zinc-finger proteins). **B**: The same region from approximately 30 Mb to 60 Mb shows a surprisingly high density of snRNAs (absolute numbers, Y-axis, in 1 Mb windows, X-axis).

The long arm of chromosome 4 represents 2.7% of the total genome length, but contains 15.5% of all non-coding zebrafish genes annotated in Ensembl. There is no increase in protein-coding genes (3.2% of total gene count), however the genes located here show a strong bias towards certain functions.

More than 30% of all Znf_C2H2-like, Znf_C2H2, Znf_BED_prd and Znf_C2H2_jaz domains present in zebrafish genes can be found on the long arm (Supplementary Table 17), reflected by an enrichment of genes with zinc ion-binding and nucleic properties in the GO terms (Supplementary Table 18). The overabundance of leucin-rich repeat, NACHT and B30.2 domains are caused by the large number of NLR genes[81] on chromosome 4. The NLR gene family is involved in the inflammation and innate immune response, a system that has been shown to be very likely to undergo extensive lineage-specific expansion accompanied with a high rate of diversification[81]. In zebrafish, this expansion seems to have experienced evolutionary pressure to stay restricted to very few areas of the genome, of which the long arm of chr4 is the most remarkable one.

Interpro accession	count	percentage of total genome count	Interpro short description	Interpro description
IPR015880	431	34.26	Znf_C2H2-like	Zinc finger, C2H2-like
IPR007087	431	34.31	Znf_C2H2	Zinc finger, C2H2
IPR003590	160	41.34	Leu-rich_rpt_RNase_inh_sub-typ	Leucine-rich repeat, ribonuclease inhibitor subtype
IPR008985	157	22.3	ConA-like_lec_gl	Concanavalin A-like lectin/glucanase
IPR001870	155	31.76	B30.2/SPRY	B30.2/SPRY domain
IPR003877	155	32.9	SPRY_rcpt	SPla/RYanodine receptor SPRY
IPR007111	154	42.89	NACHT_NTPase	NACHT nucleoside triphosphatase
IPR006574	154	34.92	PRY	SPRY-associated
IPR003879	153	34.38	Butyrophylin	Butyrophylin-like
IPR018355	148	34.82	SPla/RYanodine_receptor_subgr	SPla/RYanodine receptor subgroup
IPR001611	128	26.06	Leu-rich_rpt	Leucine-rich repeat
IPR003656	28	45.9	Znf_BED_prd	Zinc finger, BED-type predicted
IPR011029	25	19.23	DEATH-like	DEATH-like
IPR022755	22	32.35	Znf_C2H2_jaz	Zinc finger, double-stranded RNA binding
IPR004020	22	50	DAPIN	DAPIN domain
IPR007237	16	43.24	CD20-like	CD20-like
IPR012337	15	14.7	RNaseH-like_dom	Ribonuclease H-like domain
IPR006912	10	40	HARBI1_nuclease_put	Putative harbinger transposase-derived nuclease
IPR009057	7	1.65	Homeodomain-like	Homeodomain-like
IPR014940	5	55.55	BAAT_C	BAAT/Acyl-CoA thioester hydrolase C-terminal

Supplementary Table 17. Chromosome 4 Long arm Interpro Domains

A list of the 20 most common Interpro domains for genes on the long arm of chromosome 4 (downstream of 24 Mb). The domains where gathered using BioMart on Ensembl version 69.

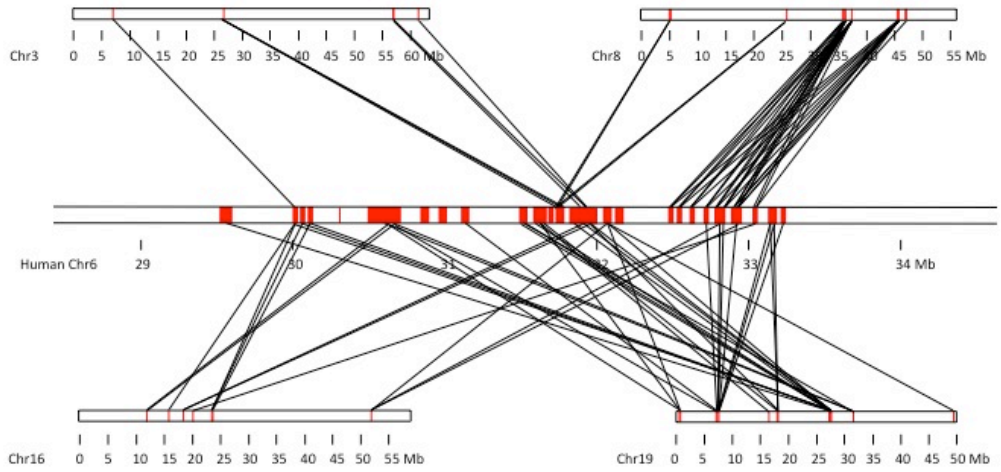
GO accession	count	percentage of total genome count	GO domain	GO description
GO:0008270	445	16.21	molecular_function	zinc ion binding
GO:0005622	443	15.27	cellular_component	intracellular
GO:0003676	256	18.18	molecular_function	nucleic acid binding
GO:0005515	197	3.18	molecular_function	protein binding
GO:0008150	49	2.32	biological_process	biological_process
GO:0003677	39	2.72	molecular_function	DNA binding
GO:0016021	27	0.87	cellular_component	integral to membrane
GO:0005524	15	0.8	molecular_function	ATP binding
GO:0016020	13	0.4	cellular_component	membrane
GO:0005575	12	0.42	cellular_component	cellular_component
GO:0000166	11	0.61	molecular_function	nucleotide binding
GO:0046872	10	0.78	molecular_function	metal ion binding
GO:0016788	10	15.87	molecular_function	hydrolase activity, acting on ester bonds
GO:0016740	10	1.36	molecular_function	transferase activity
GO:0005634	9	0.45	cellular_component	nucleus
GO:0016787	8	0.94	molecular_function	hydrolase activity
GO:0003674	7	0.37	molecular_function	molecular_function
GO:0008152	6	0.42	biological_process	metabolic process
GO:0016772	5	0.47	molecular_function	transferase activity, transferring phosphorus-containing groups
GO:0006278	5	21.73	biological_process	RNA-dependent DNA replication

Supplementary Table 18. Chromosome 4 Long arm GO terms

A list of the 20 most common GO terms for genes on the long arm of chromosome 4 (downstream of 24 Mb). The GO terms where gathered using BioMart on Ensembl version 69.

2.9 Major Histocompatibility Complex

One human genomic region of particular disease interest is the major histocompatibility complex (MHC) or HLA (human leukocyte antigen) system. In humans, the MHC comprises a super-locus that encodes, among other proteins, a large number involved in immune system function. This super-locus extends over a large region of human chromosome 6 and, despite a high degree of polymorphism[82], its content and synteny are conserved among mammals, including a comparable genomic super-locus on mouse chromosome 17[83]. Using the Ensembl Compara data to identify zebrafish orthologues of the human HLA genes, we find the great majority of genes are situated mainly on four chromosomes (3, 8, 16 and 19), while several singleton HLA orthologues are scattered throughout the zebrafish genome (Supplementary Figure 14).

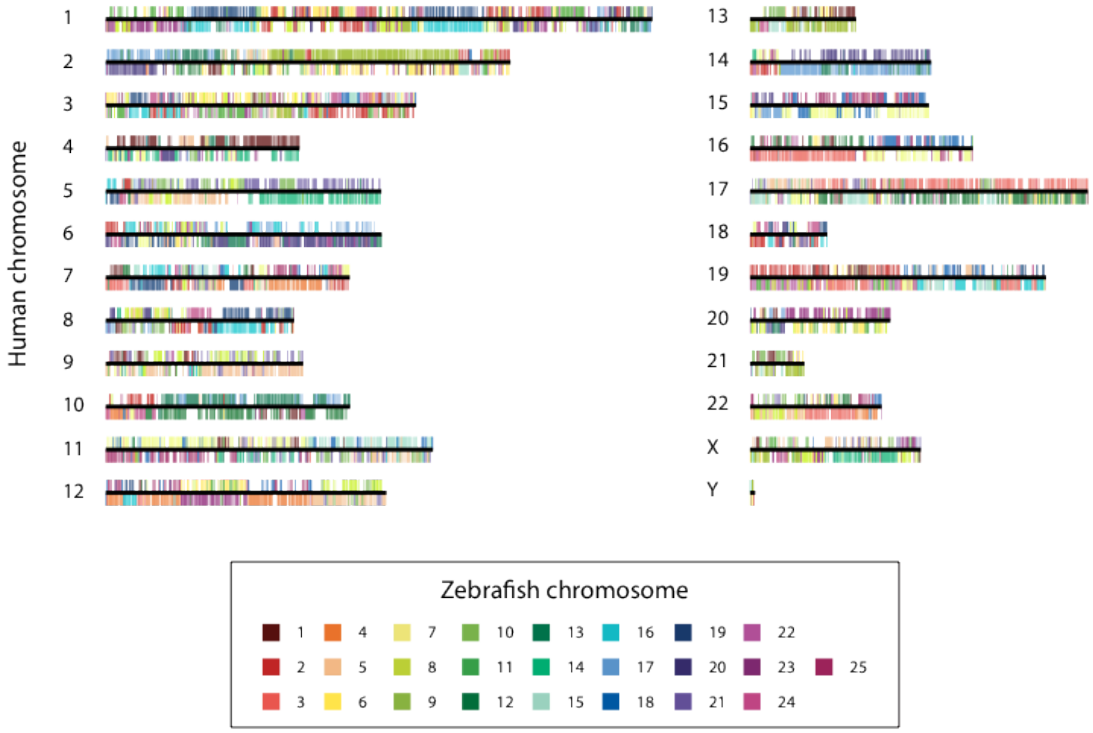


Supplementary Figure 14. MHC orthology
Orthologous relationship between human MHC genes clustered on chromosome 6, and the zebrafish chromosomes 3, 8 16 and 19, carrying the majority of orthologous MHC genes (Ensembl Compara 67).

3 Evolution

3.1 Double Conserved Synteny between Zebrafish and Human

Double conserved synteny (DCS) blocks are defined as runs of genes in the non-duplicated species that are found on two other chromosomes in the species that underwent a WGD[84], although the genes may not be adjacent in the duplicated species[85]. The DCS between human and zebrafish are represented on either side of each human chromosome (Supplementary Figure 15) and can be used to deduce chromosomal ancestry in zebrafish. For example, chromosomes 16 and 19 in zebrafish are syntenic with the same regions of human chromosomes 1, 7, 8 and 19 and likely correspond to the same ancestral chromosome.



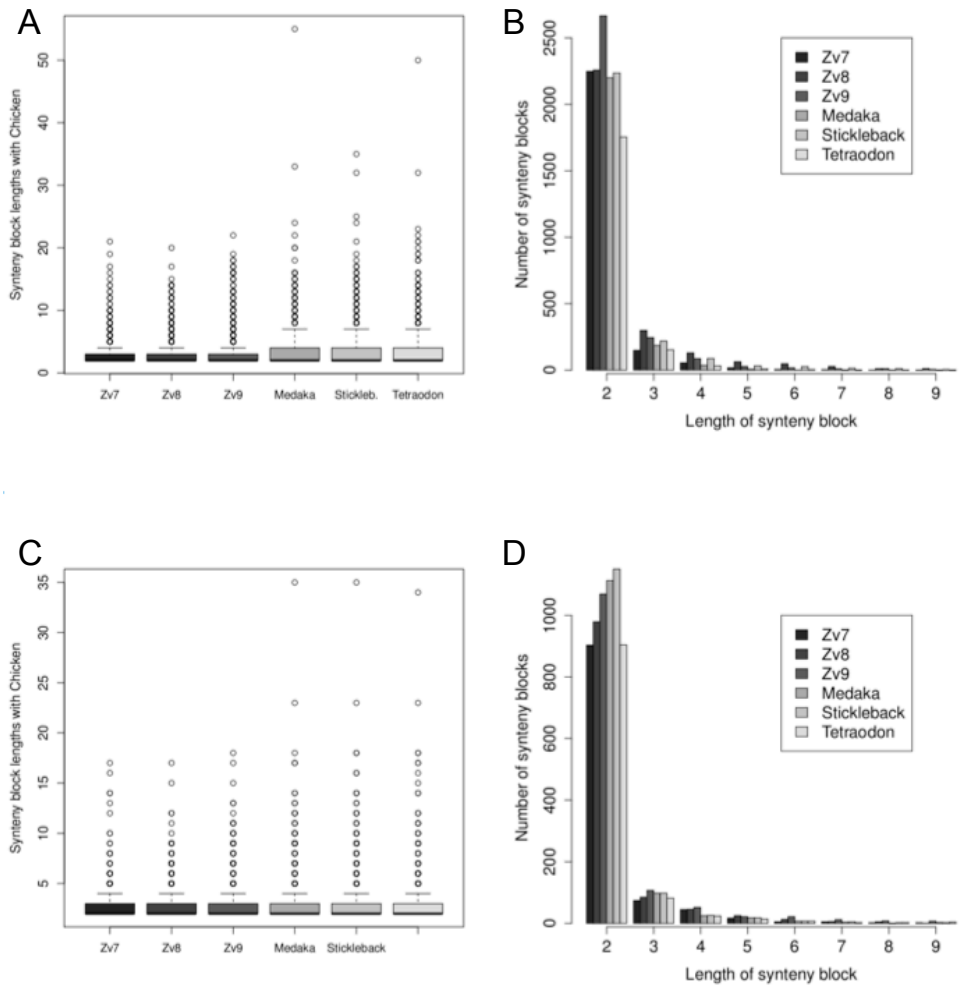
Supplementary Figure 15. Double-conserved Synteny

Each human chromosome is represented as a horizontal black line. The chromosomal locations of the orthologues of human genes in the zebrafish genome are represented along human chromosomes by ticks of colour (see legend). When two paralogous zebrafish genes are orthologous to one human gene, they are represented as ticks on either sides of the human chromosome (above and below). Human genes with only one zebrafish orthologue display a tick on one side only. As often as possible, zebrafish genes found on the same chromosome are represented on the same side of the human chromosome to highlight regions of conserved linkage between zebrafish and human (which appear as coloured blocks along human chromosomes). "Gap" regions corresponding to human genes with no orthologue in zebrafish are not represented, to facilitate visualisation. Double-conserved synteny is observed whenever two distinct chromosomes in zebrafish are orthologous to the same region(s) in the human genome as a result of TSD (two coloured blocks on either side of a human chromosome).

3.2 Conservation of Gene Linkage with the Human Genome

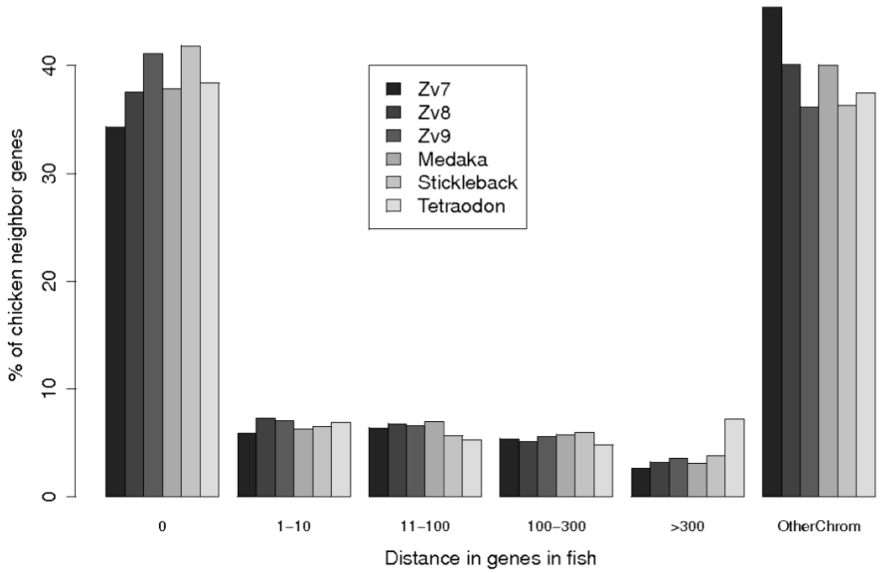
The zebrafish genome has previously been reported to be more highly rearranged than other fish genomes, with both higher inter- and intra-chromosomal rearrangement rates[68]. We tested whether this analysis still holds with the Zv9 genome assembly, by comparing the length of synteny blocks between chicken and different fish genomes. Here the chicken

genome was chosen as a reference species because, compared to mammals, its genome has been subject to fewer changes since the ancestral Euteleostomi genome. It therefore provides better sensitivity than the human or mouse genomes for comparisons with distant species. Direct comparison of the average length of synteny blocks between chicken and the four fish shows that synteny blocks in Zv9 are slightly shorter on average than for other fish (Supplementary Figure 16A), but in absolute numbers zebrafish has more synteny blocks (long and short) than the other fish (Supplementary Figure 16B). This is likely due to the fact that zebrafish has more annotated orthologues in chicken than the other fish. To avoid this bias, we built synteny blocks only with genes that exist in one-to-one copy in all four fish species. The difference in average block lengths between zebrafish and the other three fish then becomes non-significant after correction for multiple testing (Supplementary Figures 16C and D). Although the local conservation of gene adjacencies is equivalent between fish species, we examined the possibility that this may be different for long-range linkage between genes. We thus used a more global measure of synteny conservation similar to that used by Semon and Wolfe[68]. For a given pair of adjacent genes in chicken, we measured the number of genes that separate their orthologues in a fish genome (considering only 1-to-1 orthologues). With this measure, the profile of the Zv9 assembly is very similar to that of other fish (Supplementary Figure 17). For example, approximately 40% of chicken neighbouring genes are also direct neighbours in zebrafish, and this measure is similar in other fish genomes. This result confirms that zebrafish shows no evidence of being more rearranged than the other fish genomes.



Supplementary Figure 16. Conservation of synteny

Illustrated is the synteny conservation between chicken and fish genomes. **A**: Box-plot representation of the distribution of synteny block lengths between the chicken genome and different fish genomes, including three successive versions of the Zebrafish genome assembly and annotation. **B**. Absolute counts of synteny blocks of sizes comprised between 2 and 9 genes, between the chicken genome and the same fish genomes as in A. **C** and **D:** same distributions as in respectively A and B, except that only 1-to-1 orthologues between all species were considered.



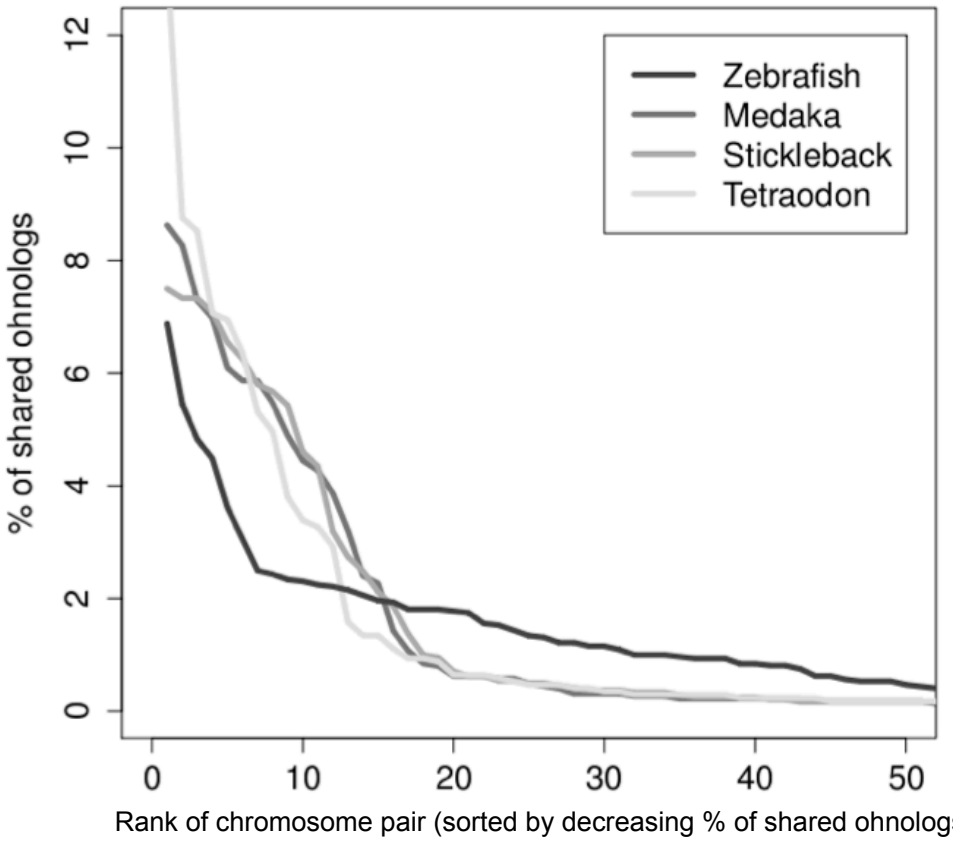
Supplementary Figure 17. Inter-gene distance conservation

Shown is the conservation of inter-gene distance between chicken and fish genomes. The Y-axis indicates the % of neighbouring chicken genes that are separated by a given number of genes in the fish genomes (X-axis).

3.3 Higher rate of interchromosomal rearrangements in zebrafish

The availability of ohnolog pairs defined on the basis of the common ancestor of five sequenced fish genomes allows us to compare chromosomal architectures between fish genomes at a scale that was not permitted using conventional measures of synteny with tetrapod genomes. Indeed, immediately after the teleost WGD, duplicated chromosome pairs exclusively share all their ohnologs, and this neat pattern will progressively be degraded by interchromosomal rearrangements, leading to situation where a given chromosome may share ohnologs with more than one other chromosome. Counting the number of ohnolog pairs shared between any two chromosomes in modern fish genomes thus provides insight into the degree of interchromosomal rearrangement perturbing the initial genome architecture. It has been estimated that the ancestral teleost genome was composed of 13 chromosomes prior to the WGD[86]. If no rearrangements (fusions, fissions, translocations) had taken place between chromosomes after the WGD, each modern genome should still contain 13 pairs of chromosomes with a significant number of ohnologs pairs. We find here that the genomes of Stickleback, Medaka and *Tetraodon*. All have between 15 and 20

chromosome pairs that share more than 1% of the total number of ohnologs (Supplementary Figure 18), suggesting that they have been affected by few interchromosomal rearrangements since the WGD. With the same measure, Zebrafish displays 31 pairings of chromosomes that account for more than 1% of the total number of ohnologs. This result can only be explained by a higher rate of interchromosomal rearrangements in zebrafish since the ancestral genome compared to the other fish lineages, leading to a higher rate of redistribution of ohnologs among the different chromosomes.



Supplementary Figure 18. Shared ohnologs

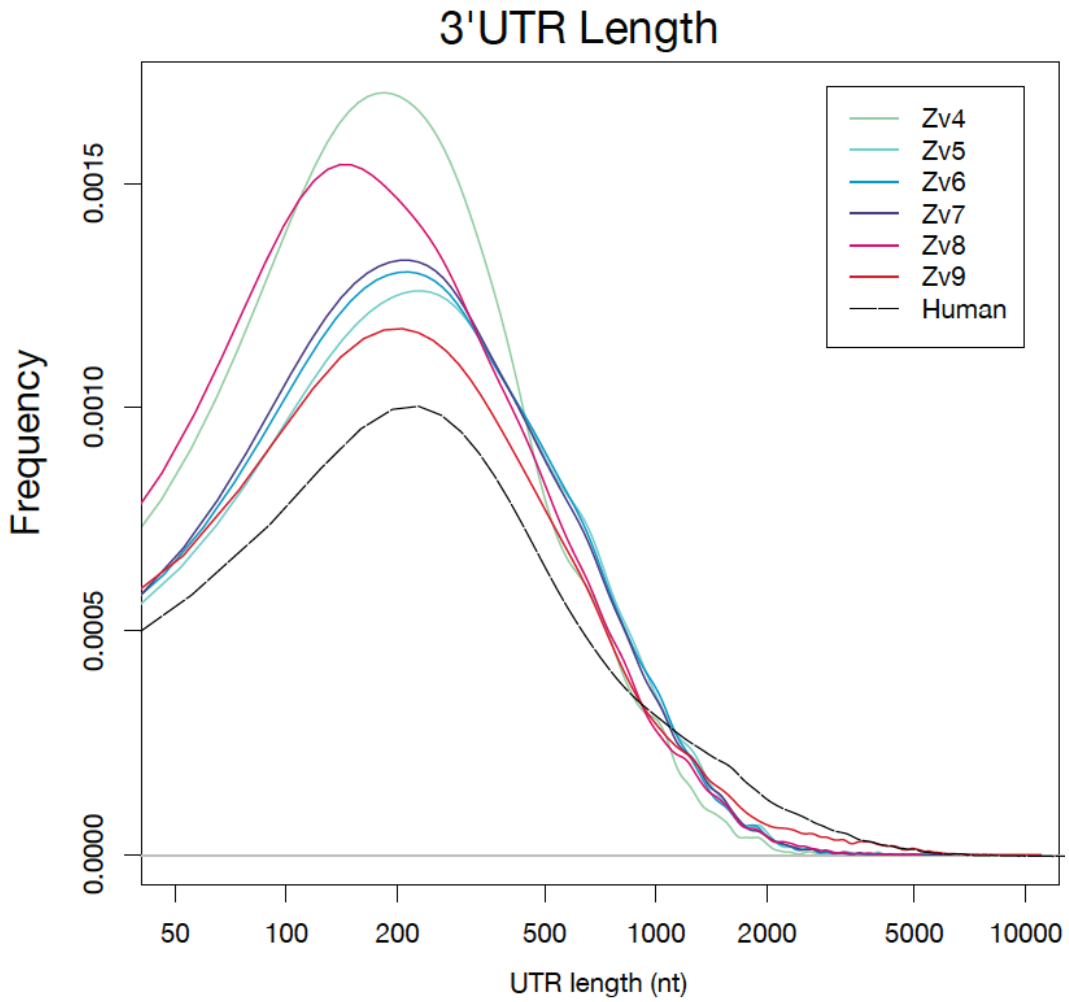
The degree of ohnolog sharing between chromosome pairs in fish genomes is diagrammed. The graph shows the % of shared ohnologs (out of all ohnologs present in the genome) between any two pairs of chromosomes. For each fish genome, pairs of chromosomes are ranked by decreasing % of shared ohnologs.

4 3'UTR Sequences and microRNA binding

The 3'UTRs of mRNAs are important for transcript stability, maintenance and regulation. Accurate delineation of transcriptional boundaries is extremely important for the prediction of regulatory events such as microRNA (miRNA) binding[87]. In particular it has been shown that inaccuracies in 3'UTR definitions are a major source of error for miRNA target prediction[87]. The Zv9 assembly presented here represents a significant effort to improve the quality of the genome sequence and transcript annotation. We sought to examine the extent of this improvement to the 3'UTRs of the current zebrafish assembly and to investigate how they may help to improve miRNA regulatory target prediction and shed light on mRNA regulation in zebrafish. All sequence information was extracted using the Ensembl API[2]. The results clearly show a marked improvement in 3'UTR quality and accuracy in the latest zebrafish genome assembly, which allows the identification of many more functional miRNA binding sites.

4.1 3'UTR Length

Early releases of the zebrafish genome had shorter mean 3'UTR lengths of 150-180nt with very few long 3'UTR sequences (Supplementary Figure 19). Subsequent releases until this latest build have seen 3' UTR lengths increasing gradually. The current assembly and genome build has average 3'UTR length approaching 200nt with an increasing proportion of long 3'UTRs (>1000nt). This is comparable to the situation observed for Human 3'UTRs (Ensembl v68, assembly version GRCh37) (Supplementary Figure 19). The presence of many short truncated 3'UTRs in earlier assemblies is likely due to 3'UTR delineation coming from cDNA and EST overlap data.



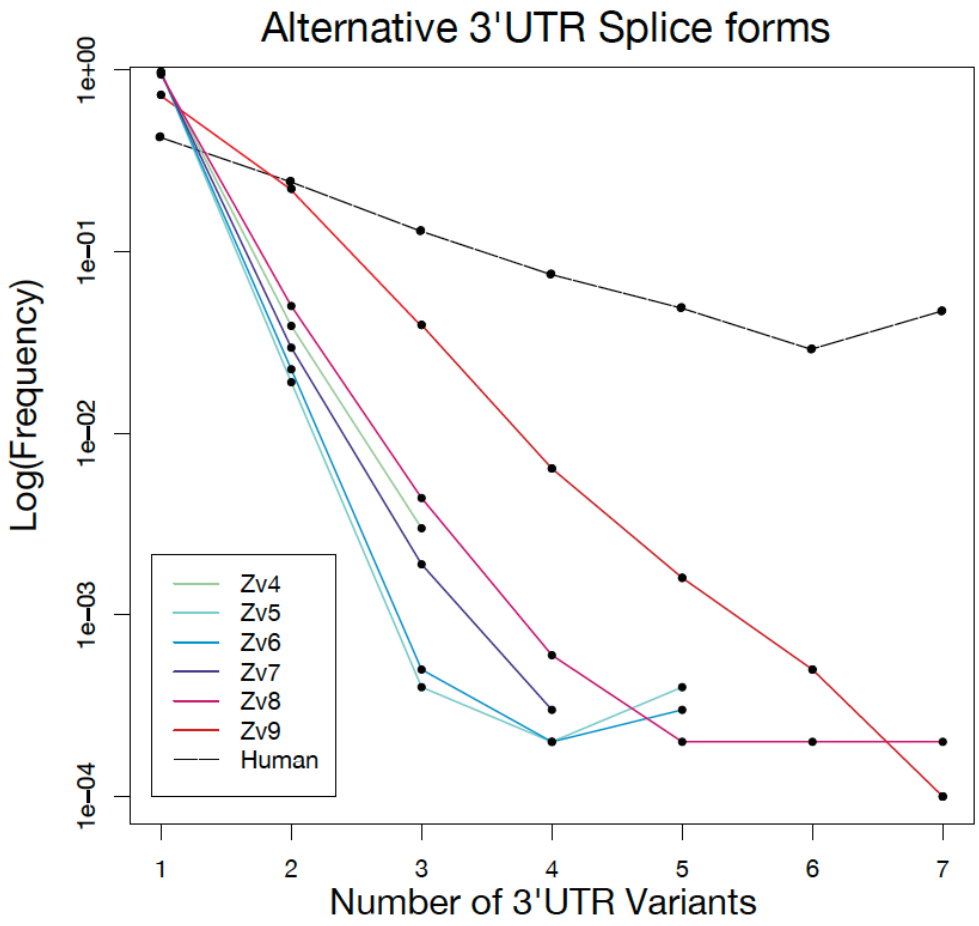
Supplementary Figure 19. 3'UTR lengths

The distribution of 3'UTR lengths from the zebrafish Zv4 to Zv9 and human GRCh37 genome builds is shown. The x-axis shows the 3'UTR length of each transcript and the y-axis shows the frequency of each length.

4.2 3'UTR Splicing

Of particular interest is the prevalence of splice-forms of a transcript that may have different 3' ends and hence be subject to different and or alternative modes of regulation. The current assembly, Zv9, shows an expanded repertoire of 3'UTR splice forms that were previously unknown (Supplementary Figure 20). Additionally, the number of 3'UTR isoforms was significantly lower in Zv4 compared to Zv9, with over 20% of transcripts having alternative 3'UTRs compared to 4% in Zv4, although less highly variable 3'UTRs than found in the Human genome (Supplementary Figure 20). This increased repertoire will hopefully allow more accurate target

prediction using tissue-specific splice forms to be performed in future experiments.



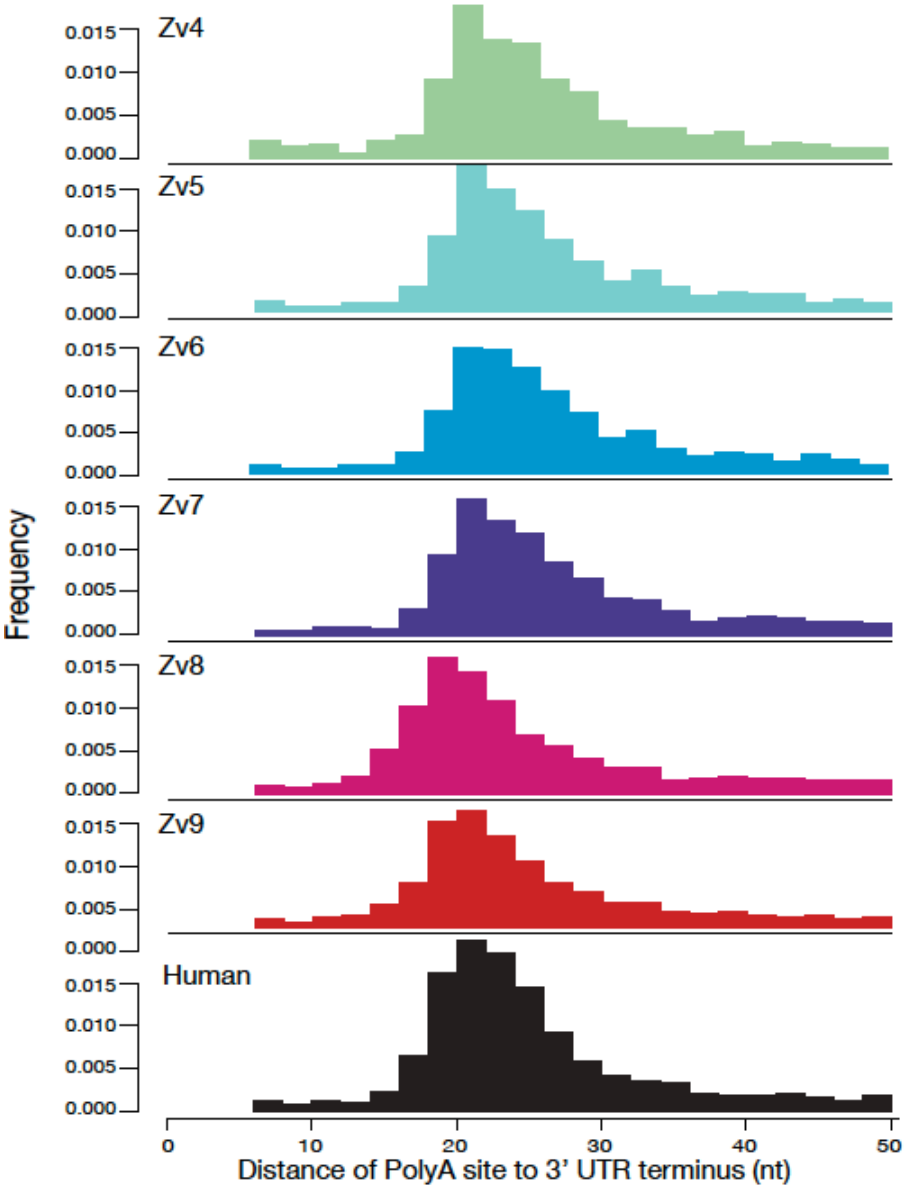
Supplementary Figure 20. Alternative 3'UTRs

Shown is a graph of the frequency of splice forms with alternative 3'UTRs. The x-axis shows the number of alternative 3'UTRs per gene. The y-axis shows the log frequency of each occurrence.

4.3 3' UTR Poly-Adenylation Signal Analysis

Assessing the accuracy of 3'UTR delineation is difficult, although one metric that may be employed is to examine the occurrence and distribution of poly-adenylation signals that fall near the end of the 3'UTR. Accurate 3'UTR sequences will usually show a strong peak of poly-A signals approximately 20nt from the end of the 3'UTR sequence before the poly-A tail. Examination of poly-A signal (`AATAAA`, `ATTAAA`, `AGTAAA`, `TATAAA`, `CATAAA`, `GATAAA`, `AATATA`, `AATACA`, `AATAGA`, `AAAAAG` and `ACTAAA`, most upstream match being recorded for the analysis) occurrence across multiple zebrafish assemblies

shows that the latest version has a sharp peak of poly-A signals at 19-20nt and echoes closely the situation observed for Human 3'UTR sequences (Supplementary Figure 21). Previous genome versions showed more variation and a wider spread of signals indicating that the quality of 3'UTR boundary delineation in the current genome version appears to have improved (Supplementary Figure 21).

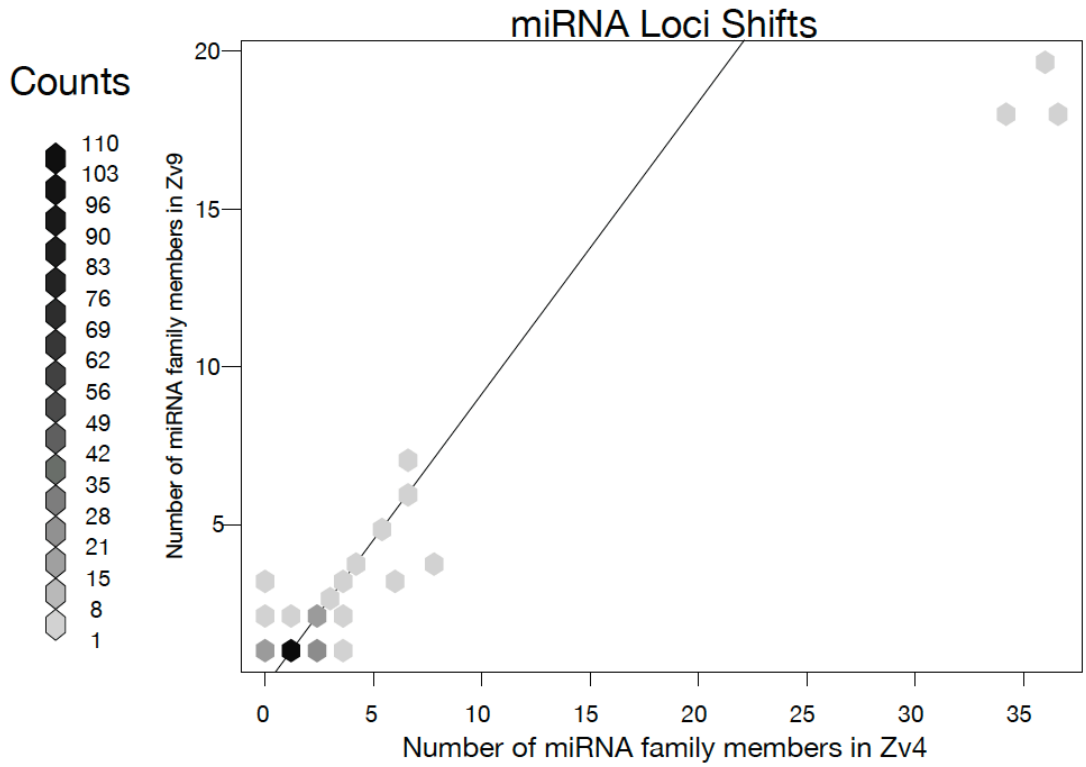


Supplementary Figure 21. Distribution of poly-A sites

Shown is the distribution of poly-A sites across the terminal 50nt of each transcript across multiple builds with Human for comparison.

4.4 Changes to microRNA loci

The current zebrafish assembly has allowed the copy number of many miRNAs to be more accurately recorded. In particular the high-copy number miRNAs from the miR-430 family occur in large repeat clusters that were previously difficult to assemble. Most miRNA loci are stable and are still present at a 1:1 ratio as compared to early assemblies (Zv4). A total of 36 miRNAs, which were previously reported at two loci, are now only present as single copies, while 30 miRNAs have been show to be expanded in the Zv9 assembly. The high-copy number miR-430 family[88] has stabilised at 18-20 copies per family member from 35 copies in earlier assemblies (Supplementary Figure 22).

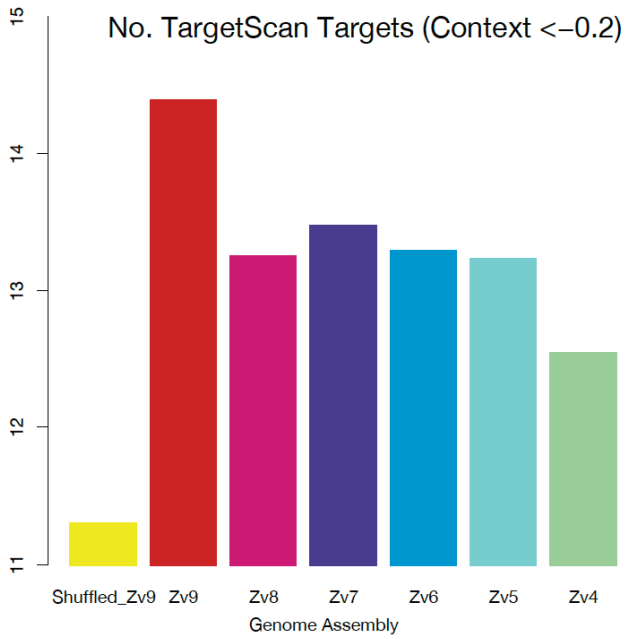


Supplementary Figure 22. Total number of miRNA loci

Shown is a comparison of the total number of miRNA loci reported in Zv4 vs. Zv9. The solid line shows where a 1:1 correspondence would be observed.

4.5 Global miRNA target analysis

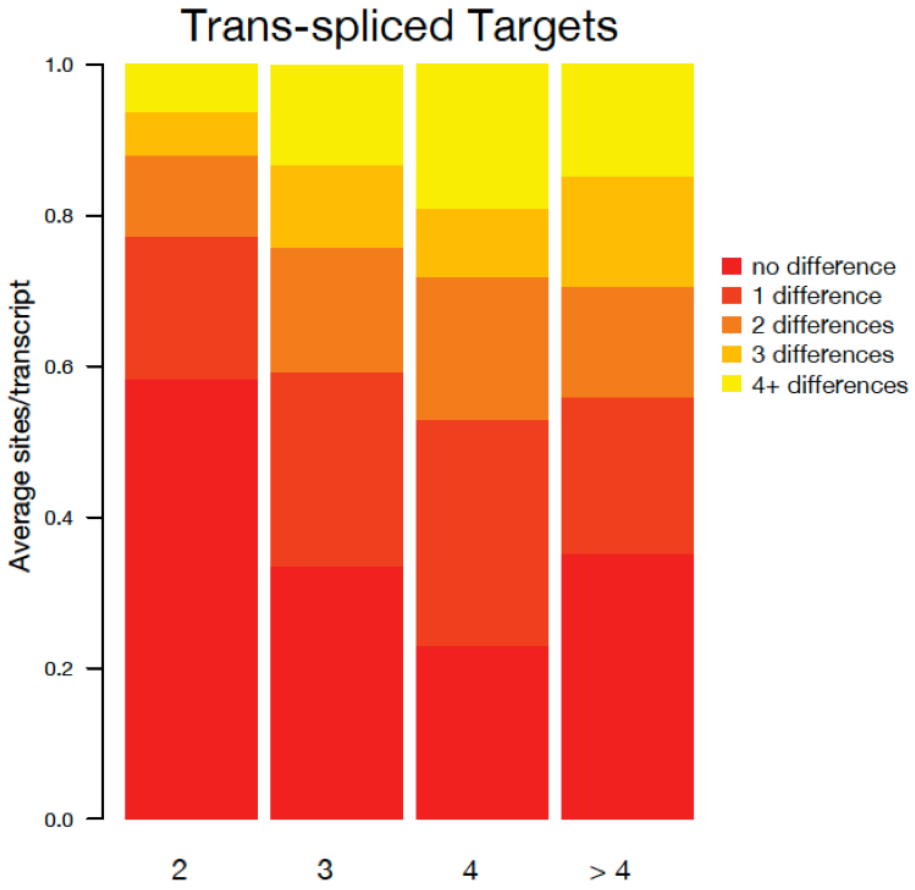
A more general method to explore the impact of higher quality and better delineated 3'UTR sequences is to perform genome-wide miRNA target prediction using TargetScan[89] (version v5.2 with context score filter >0.2, merged by seed sequence). We investigated this by performing TargetScan analysis on different zebrafish assemblies to assess the frequency of predicted targets compared to a shuffled background model. This analysis shows a dramatic increase in biologically relevant (context score ≥ 0.2) target predictions in *D. rerio* between Zv4 and Zv9 compared to shuffled negative controls from Zv9 (Supplementary Figure 23). The Zv9 assembly harbours significantly more predicted miRNA target sites per transcript as compared to previous builds and to a shuffled control set of sequences, created by di-nucleotide shuffling using *ushuffle*. This is likely due to increased 3'UTR length, quality and the presence of multiple splice forms with alternative 3'UTR regulatory modules.



Supplementary Figure 23. TargetScan targets
Displayed is a plot of the total number or TargetScan targets (context score > 0.2) predicted per transcript for Zv4-Zv9.

4.6 Analysis of trans-spliced miRNA targets

There are 4206 alternatively spliced 3'UTR sequences present in Zv9 and 1948 of these UTR isoforms show significant differences in predicted miRNA target sites as predicted by TargetScan[89] (Supplementary Figure 24). In 40% of cases where there are two alternative 3'UTR sequences for a gene we observe differences in targets. Genes with more 3'UTR splice variants show increasing amounts of differential target sites (Supplementary Figure 24). This result indicates that many alternatively spliced transcripts present different 3'UTR regulatory modules to the RNA regulatory machinery in the cell allowing for fine-grained regulation.



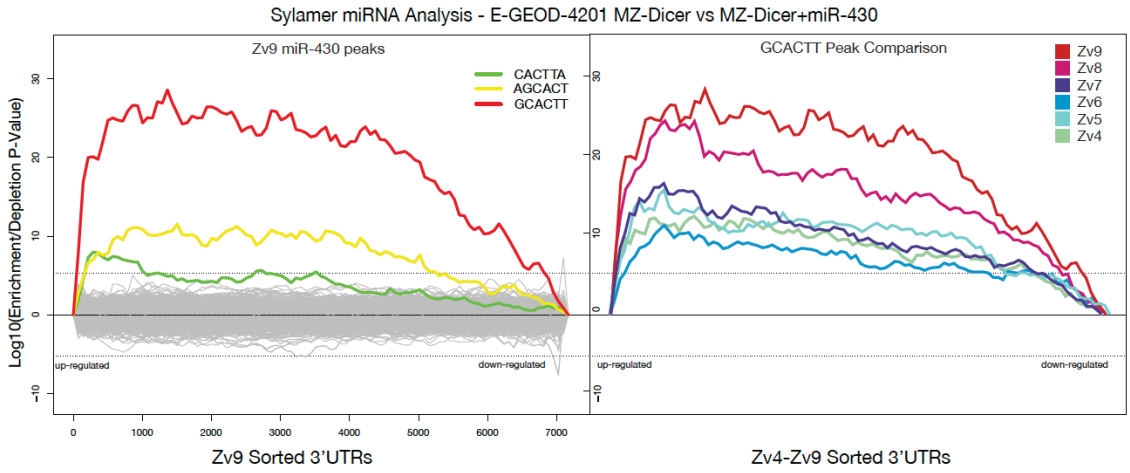
Supplementary Figure 24. Differential target frequency
Displayed is the frequency of differential targets detected in 3'UTRs with multiple splice forms.

4.7 Improved 3'UTRs allow better miRNA target detection

One of the earliest large-scale miRNA target analysis experiments in zebrafish established the importance of miRNA miR-430 in the maternal-zygotic transition of development. Dicer deficient embryos were compared to embryos micro-injected with miR-430 (a miRNA involved in early development). Injection of miR-430 clearly showed a large-scale effect on transcript abundance in microarray studies, specifically a significant down-regulation of thousands of transcripts[90]. Those transcripts with delineated 3'UTRs and containing a miR-430 seed-matching site, were found to be significantly enriched in the down-regulated genes upon miR-430 injection and highly enriched for maternally derived transcripts. In this case miR-430 acts at the maternal to zygotic switch.

The miR-430 experiment represents a gold-standard set of miRNA targets and expression data for testing the quality of 3'UTR sequences. We have reanalysed the original data from this experiment based on the Zv4 *D. rerio* assembly using the latest 3'UTR sequences from Zv9. Our goal was to assess the extent of improvement possible over the original result with high-quality 3'UTR sequences.

We reanalysed these data from the original MZ-Dicer microarray experiment using Sylamer[91] (Supplementary Figure 25). The analysis was performed on the original dataset from E-GEOD-4201 sorted from down to up-regulated according to log fold change. Affymetrix probes were remapped to transcripts for each genome assembly using the Ensembl API and 3'UTRs were assigned based on those present at the time of the assembly. Where multiple probes matched the same 3'UTR sequence the probe with the highest IQR was retained while the others were excluded. Shuffling was performed retaining di-nucleotide frequencies using *ushuffle*. MicroRNA family information for the loci analysis was obtained from miRBase[92] sequences using the MapMi[93] algorithm.



Supplementary Figure 25. Effect of miRNA binding site

Shown are Sylamer plots illustrating the effect of miRNA binding site on gene expression for the E-GEOD-4201 MZ-Dicer vs. MZ-Dicer miR430 microinjection. The left panel shows all seed sites detected in 3'UTRs from Zv9 with a peak P-Value of $1X10^{-30}$. The right panel shows the shape of the canonical miR430 seed from equivalent analyses using 3'UTRs from previous zebrafish genome assemblies. The original experiment (Zv4) peak P-value was $1x10^{-14}$, showing a significant improvement in miRNA targets from these gene expression data with high-quality 3'UTR sequences.

The original experiment showed a significant enrichment for miR-430 seeds with 380 3'UTRs showing enrichment at a P-value of $< 1x10^{-14}$. Remapping the array probes to the Zv9 3'UTRs captures a far stronger signal for many more 3'UTR sequences (1320 3'UTRs at a P-value $< 1x10^{-30}$). This improvement is due to the availability of many more 3'UTR sequences in Zv9 and an improvement in their boundary accuracy. The new set of 3'UTR sequences is far better at explaining the expression changes observed in the experiment and presents a broadened and more accurate set of miR-430 targets involved in maternal zygotic switching in the early zebrafish embryo.

5 References

1 Jekosch, K. The zebrafish genome project: sequence analysis and annotation. *Methods Cell Biol* **77**, 225-239 (2004).

2 Flicek, P. *et al.* Ensembl 2012. *Nucleic Acids Res* **40**, D84-90, doi:10.1093/nar/gkr991 (2012).

3 Mullikin, J. C. & Ning, Z. The phusion assembler. *Genome Res* **13**, 81-90, doi:10.1101/gr.731003 (2003).

4 Streisinger, G., Walker, C., Dower, N., Knauber, D. & Singer, F. Production of clones of homozygous diploid zebra fish (Brachydanio rerio). *Nature* **291**, 293-296 (1981).

5 Wu, Y., Bhat, P. R., Close, T. J. & Lonardi, S. Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. *PLoS genetics* **4**, e1000212, doi:10.1371/journal.pgen.1000212 (2008).

6 Clark, M. D. *et al.* Single nucleotide polymorphism (SNP) panels for rapid positional cloning in zebrafish. *Methods Cell Biol* **104**, 219-235, doi:10.1016/B978-0-12-374814-0.00013-6 (2011).

7 Kozarewa, I. *et al.* Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods* **6**, 291-295, doi:10.1038/nmeth.1311 (2009).

8 Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**, 1851-1858, doi:10.1101/gr.078212.108 (2008).

9 Smit, A. F. A., Hubley, R. & Green, P. *RepeatMasker Open-3.0*, <http://www.repeatmasker.org> (1996-2010).

10 Hackett, C. A. & Broadfoot, L. B. Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. *Heredity (Edinb)* **90**, 33-38, doi:10.1038/sj.hdy.6800173 (2003).

11 Bradley, K. M. *et al.* An SNP-Based Linkage Map for Zebrafish Reveals Sex Determination Loci. *G3 (Bethesda)* **1**, 3-9, doi:10.1534/g3.111.000190 (2011).

12 Anderson, J. L. *et al.* Multiple Sex-Associated Regions and a Putative Sex Chromosome in Zebrafish Revealed by RAD Mapping and Population Genomics. *PLoS ONE* **7**, e40701, doi:10.1371/journal.pone.0040701 (2012).

13 Woods, I. G. *et al.* The zebrafish gene map defines ancestral vertebrate chromosomes. *Genome Res* **15**, 1307-1314, doi:10.1101/gr.4134305 (2005).

14 Soderlund, C., Longden, I. & Mott, R. FPC: a system for building contigs from restriction fingerprinted clones. *Comput Appl Biosci* **13**, 523-535 (1997).

15 Soderlund, C., Humphray, S., Dunham, A. & French, L. Contigs built with fingerprints, markers, and FPC V4.7. *Genome Res* **10**, 1772-1787 (2000).

16 Geisler, R. *et al.* A radiation hybrid map of the zebrafish genome. *Nat Genet* **23**, 86-89, doi:10.1038/12692 (1999).

17 Freeman, J. L. *et al.* Definition of the zebrafish genome using flow cytometry and cytogenetic mapping. *BMC Genomics* **8**, 195, doi:10.1186/1471-2164-8-195 (2007).

18 Corley-Smith, G. E., Brandhorst, B. P., Walker, C. & Postlethwait, J. H. Production of haploid and diploid androgenetic zebrafish (including methodology for delayed in vitro fertilization). *Methods Cell Biol* **59**, 45-60 (1999).

19 Westerfield, M. *THE ZEBRAFISH BOOK; A guide for the laboratory use of zebrafish (Danio rerio)*. 5th edn, (University of Oregon Press, 2007).

20 Kwok, C. *et al.* Characterization of whole genome radiation hybrid mapping resources for non-mammalian vertebrates. *Nucleic Acids Res* **26**, 3562-3566 (1998).

21 Knapik, E. W. *et al.* A microsatellite genetic linkage map for zebrafish (Danio rerio). *Nat Genet* **18**, 338-343, doi:10.1038/ng0498-338 (1998).

22 Shimoda, N. *et al.* Zebrafish genetic map with 2000 microsatellite markers. *Genomics* **58**, 219-232, doi:10.1006/geno.1999.5824 (1999).

23 Kelly, P. D. *et al.* Genetic linkage mapping of zebrafish genes and ESTs. *Genome Res* **10**, 558-567 (2000).

24 Gordon, D., Desmarais, C. & Green, P. Automated finishing with autofinish. *Genome Res* **11**, 614-625, doi:10.1101/gr.171401 (2001).

25 Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**, 186-194 (1998).

26 Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Res* **11**, 1725-1729, doi:10.1101/gr.194201 (2001).

27 Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132**, 365-386 (2000).

28 Beasley, H., Grafham, D. & Willey, D. in *eLS* (John Wiley & Sons, Ltd, 2001).

29 Devine, S. E., Chissoe, S. L., Eby, Y., Wilson, R. K. & Boeke, J. D. A transposon-based strategy for sequencing repetitive DNA in eukaryotic genomes. *Genome Res* **7**, 551-563 (1997).

30 McMurray, A. A., Sulston, J. E. & Quail, M. A. Short-insert libraries as a method of problem solving in genome sequencing. *Genome Res* **8**, 562-566 (1998).

31 van Tonder, A. J. & Grafham, D. in *eLS* (John Wiley & Sons, Ltd, 2001).

32 Chain, P. S. *et al.* Genomics. Genome project standards in a new era of sequencing. *Science* **326**, 236-237, doi:10.1126/science.1180614 (2009).

33 Sudbery, I. *et al.* Deep short-read sequencing of chromosome 17 from the mouse strains A/J and CAST/Ei identifies significant germline variation and candidate genes that regulate liver triglyceride levels. *Genome Biology* **10**, R112 (2009).

34 Church, D. M. *et al.* Modernizing reference genome assemblies. *PLoS Biol* **9**, e1001091, doi:10.1371/journal.pbio.1001091 (2011).

35 Takahashi, H. Juvenile hermaphroditism in the zebrafish, Brachydanio rerio. *Bull Fac Fish Hokkaido Univ* **28**, 57–65 (1977).

36 Uchida, D., Yamashita, M., Kitano, T. & Iguchi, T. Oocyte apoptosis during the transition from ovary-like tissue to testes during sex differentiation of juvenile zebrafish. *J Exp Biol* **205**, 711-718 (2002).

37 Rodriguez-Mari, A. *et al.* Roles of brca2 (fancd1) in oocyte nuclear architecture, gametogenesis, gonad tumors, and genome stability in zebrafish. *PLoS genetics* **7**, e1001357, doi:10.1371/journal.pgen.1001357 (2011).

38 Amores, A. *et al.* Developmental roles of pufferfish Hox clusters and genome evolution in ray-fin fish. *Genome Res* **14**, 1-10, doi:10.1101/gr.1717804 (2004).

39 Amores, A. *et al.* Zebrafish hox clusters and vertebrate genome evolution. *Science* **282**, 1711-1714 (1998).

40 Straub, T. & Becker, P. B. Dosage compensation: the beginning and end of generalization. *Nat Rev Genet* **8**, 47-57, doi:10.1038/nrg2013 (2007).

41 Consortium, G. P. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:10.1038/nature09534 (2010).

42 Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595, doi:10.1093/bioinformatics/btp698 (2010).

43 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

44 Le, S. Q. & Durbin, R. SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Res* **21**, 952-960, doi:10.1101/gr.113084.110 (2011).

45 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

46 Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:10.1038/nature11632 (2012).

47 Murchison, E. P. *et al.* Genome sequencing and analysis of the Tasmanian devil and its transmissible cancer. *Cell* **148**, 780-791, doi:10.1016/j.cell.2011.11.065 (2012).

48 Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659, doi:10.1093/bioinformatics/btl158 (2006).

49 Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-664, doi:10.1101/gr.229202. Article published online before March 2002 (2002).

50 Bernardi, G. Compositional transitions in the nuclear genomes of cold-blooded vertebrates. *J Mol Evol* **31**, 282-293 (1990).

51 Costantini, M., Auletta, F. & Bernardi, G. Isochore patterns and gene distributions in fish genomes. *Genomics* **90**, 364-371, doi:10.1016/j.ygeno.2007.05.006 (2007).

52 Xu, P. *et al.* Generation of the first BAC-based physical map of the common carp genome. *BMC Genomics* **12**, 537, doi:10.1186/1471-2164-12-537 (2011).

53 Tilgner, H. *et al.* Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol* **16**, 996-1001, doi:10.1038/nsmb.1658 (2009).

54 Kauffman, E. J. *et al.* Microsatellite-centromere mapping in the zebrafish (Danio rerio). *Genomics* **30**, 337-341, doi:10.1006/geno.1995.9869 (1995).

55 Mohideen, M. A., Moore, J. L. & Cheng, K. C. Centromere-linked microsatellite markers for linkage groups 3, 4, 6, 7, 13, and 20 of zebrafish (Danio rerio). *Genomics* **67**, 102-106, doi:10.1006/geno.2000.6233 (2000).

56 Phillips, R. B., Amores, A., Morasch, M. R., Wilson, C. & Postlethwait, J. H. Assignment of zebrafish genetic linkage groups to chromosomes. *Cytogenet Genome Res* **114**, 155-162, doi:10.1159/000093332 (2006).

57 Sola, L. & Gornung, E. Classical and molecular cytogenetics of the zebrafish, Danio rerio (Cyprinidae, Cypriniformes): an overview. *Genetica* **111**, 397-412 (2001).

58 Feschotte, C. Merlin, a new superfamily of DNA transposons identified in diverse animal genomes and related to bacterial IS1016 insertion sequences. *Mol Biol Evol* **21**, 1769-1780, doi:10.1093/molbev/msh188 (2004).

59 Kapitonov, V. V., Pavlicek, A. & Jurka, J. in *Encyclopedia of Molecular Cell Biology and Molecular Medicine* (Wiley-VCH Verlag GmbH & Co. KGaA, 2006).

60 Phillips, R. B. & Reed, K. M. Localization of repetitive DNAs to zebrafish (Danio rerio) chromosomes by fluorescence in situ hybridization (FISH). *Chromosome Res* **8**, 27-35 (2000).

61 Kalari, K. R. *et al.* First exons and introns--a survey of GC content and gene structure in the human genome. *In Silico Biol* **6**, 237-242 (2006).

62 Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573-580 (1999).

63 Morgulis, A., Gertz, E. M., Schaffer, A. A. & Agarwala, R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J Comput Biol* **13**, 1028-1040, doi:10.1089/cmb.2006.13.1028 (2006).

64 Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110**, 462-467, doi:10.1159/000084979 (2005).

65 Izsvak, Z., Ivics, Z., Garcia-Estefania, D., Fahrenkrug, S. C. & Hackett, P. B. DANA elements: a family of composite, tRNA-derived short interspersed DNA elements associated with mutational activities in zebrafish (Danio rerio). *Proc Natl Acad Sci U S A* **93**, 1077-1081 (1996).

66 Feschotte, C. & Pritham, E. J. DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet* **41**, 331-368, doi:10.1146/annurev.genet.40.110405.090448 (2007).

67 Kasahara, M. *et al.* The medaka draft genome and insights into vertebrate genome evolution. *Nature* **447**, 714-719, doi:10.1038/nature05846 (2007).

68 Semon, M. & Wolfe, K. H. Rearrangement rate following the whole-genome duplication in teleosts. *Mol Biol Evol* **24**, 860-867, doi:10.1093/molbev/msm003 (2007).

69 Furano, A. V., Duvernell, D. D. & Boissinot, S. L1 (LINE-1) retrotransposon diversity differs dramatically between mammals and fish. *Trends Genet* **20**, 9-14 (2004).

70 Fu, B., Chen, M., Zou, M., Long, M. & He, S. The rapid generation of chimerical genes expanding protein diversity in zebrafish. *BMC Genomics* **11**, 657, doi:10.1186/1471-2164-11-657 (2010).

71 de Boer, J. G., Yazawa, R., Davidson, W. S. & Koop, B. F. Bursts and horizontal evolution of DNA transposons in the speciation of pseudotetraploid salmonids. *BMC Genomics* **8**, 422, doi:10.1186/1471-2164-8-422 (2007).

72 Kawakami, K. Tol2: a versatile gene transfer vector in vertebrates. *Genome Biology* **8 Suppl 1**, S7, doi:10.1186/gb-2007-8-s1-s7 (2007).

73 Ivics, Z., Hackett, P. B., Plasterk, R. H. & Izsvak, Z. Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. *Cell* **91**, 501-510 (1997).

74 Kawakami, K. Transposon tools and methods in zebrafish. *Dev Dyn* **234**, 244-254, doi:10.1002/dvdy.20516 (2005).

75 Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* **10**, 691-703, doi:10.1038/nrg2640 (2009).

76 Hellsten, U. *et al.* The genome of the Western clawed frog Xenopus tropicalis. *Science* **328**, 633-636, doi:10.1126/science.1183670 (2010).

77 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797 (2004).

78 Felsenstein, J. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* **5**, 164-166, doi:citeulike-article-id:2344765 (1989).

79 Hormozdiari, F. *et al.* Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics* **26**, i350-357, doi:10.1093/bioinformatics/btq216 (2010).

80 Lam, W. L., Lee, T. S. & Gilbert, W. Active transposition in zebrafish. *Proc Natl Acad Sci U S A* **93**, 10870-10875 (1996).

81 Stein, C., Caccamo, M., Laird, G. & Leptin, M. Conservation and divergence of gene families encoding components of innate immune response systems in zebrafish. *Genome Biology* **8**, R251, doi:10.1186/gb-2007-8-11-r251 (2007).

82 Robinson, J. *et al.* The IMGT/HLA database. *Nucleic Acids Res* **39**, D1171-1176, doi:10.1093/nar/gkq998 (2011).

83 Kulski, J. K., Shiina, T., Anzai, T., Kohara, S. & Inoko, H. Comparative genomic analysis of the MHC: the evolution of class I duplication blocks, diversity and complexity from shark to man. *Immunol Rev* **190**, 95-122 (2002).

84 Kellis, M., Birren, B. W. & Lander, E. S. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* **428**, 617-624, doi:10.1038/nature02424 (2004).

85 Jaillon, O. *et al.* Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature* **431**, 946-957, doi:10.1038/nature03025 (2004).

86 Nakatani, Y., Takeda, H., Kohara, Y. & Morishita, S. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res* **17**, 1254-1265, doi:10.1101/gr.6316407 (2007).

87 Maziere, P. & Enright, A. J. Prediction of microRNA targets. *Drug Discov Today* **12**, 452-458, doi:10.1016/j.drudis.2007.04.002 (2007).

88 Giraldez, A. J. *et al.* MicroRNAs regulate brain morphogenesis in zebrafish. *Science* **308**, 833-838, doi:10.1126/science.1109020 (2005).

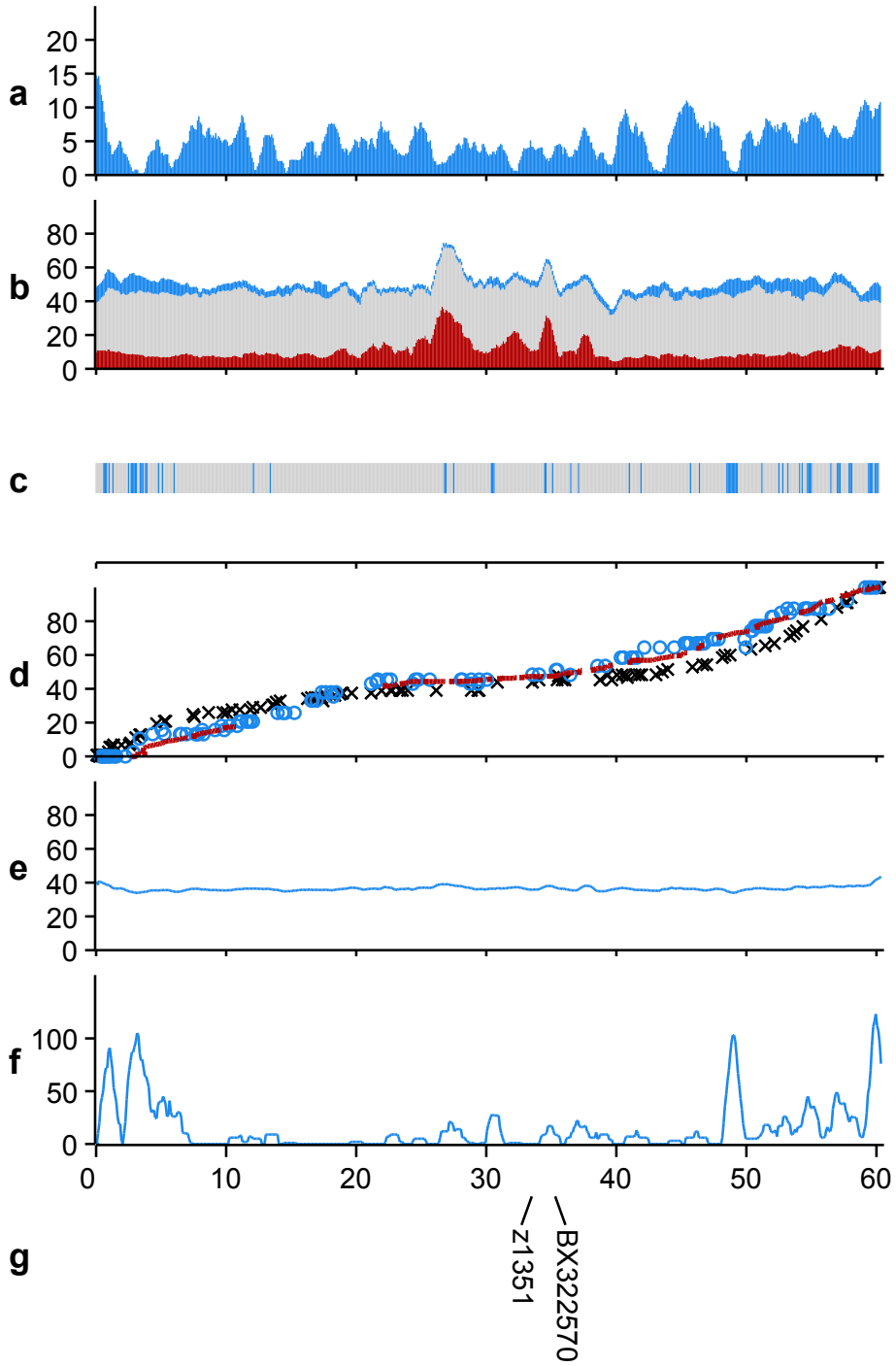
89 Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15-20, doi:10.1016/j.cell.2004.12.035 (2005).

90 Giraldez, A. J. *et al.* Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science* **312**, 75-79, doi:10.1126/science.1122689 (2006).

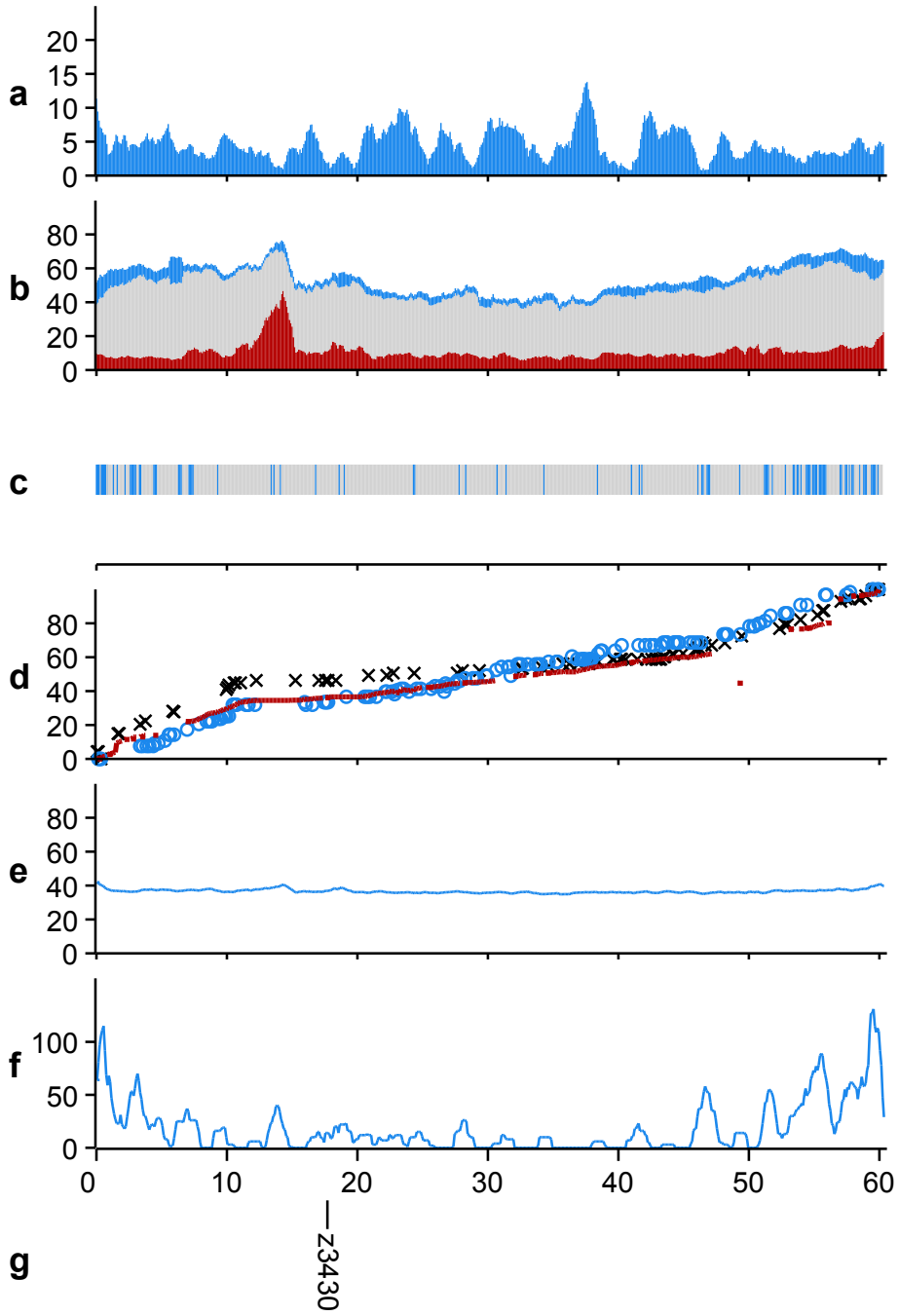
91 van Dongen, S., Abreu-Goodger, C. & Enright, A. J. Detecting microRNA binding and siRNA off-target effects from expression data. *Nat Methods* **5**, 1023-1025, doi:10.1038/nmeth.1267 (2008).

92 Kozomara, A. & Griffiths-Jones, S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* **39**, D152-157, doi:10.1093/nar/gkq1027 (2011).

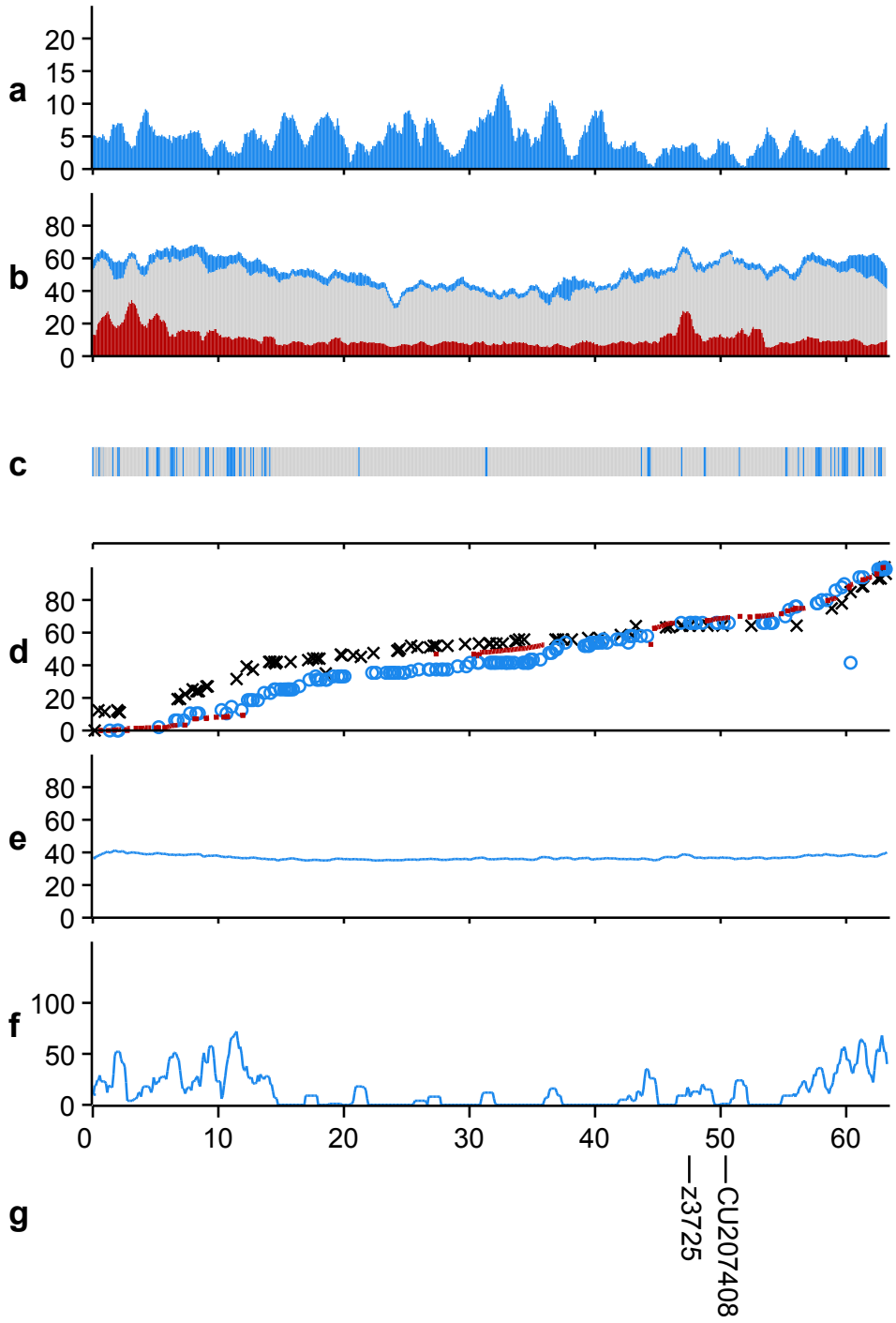
93 Guerra-Assuncao, J. A. & Enright, A. J. MapMi: automated mapping of microRNA loci. *BMC Bioinformatics* **11**, 133, doi:10.1186/1471-2105-11-133 (2010).



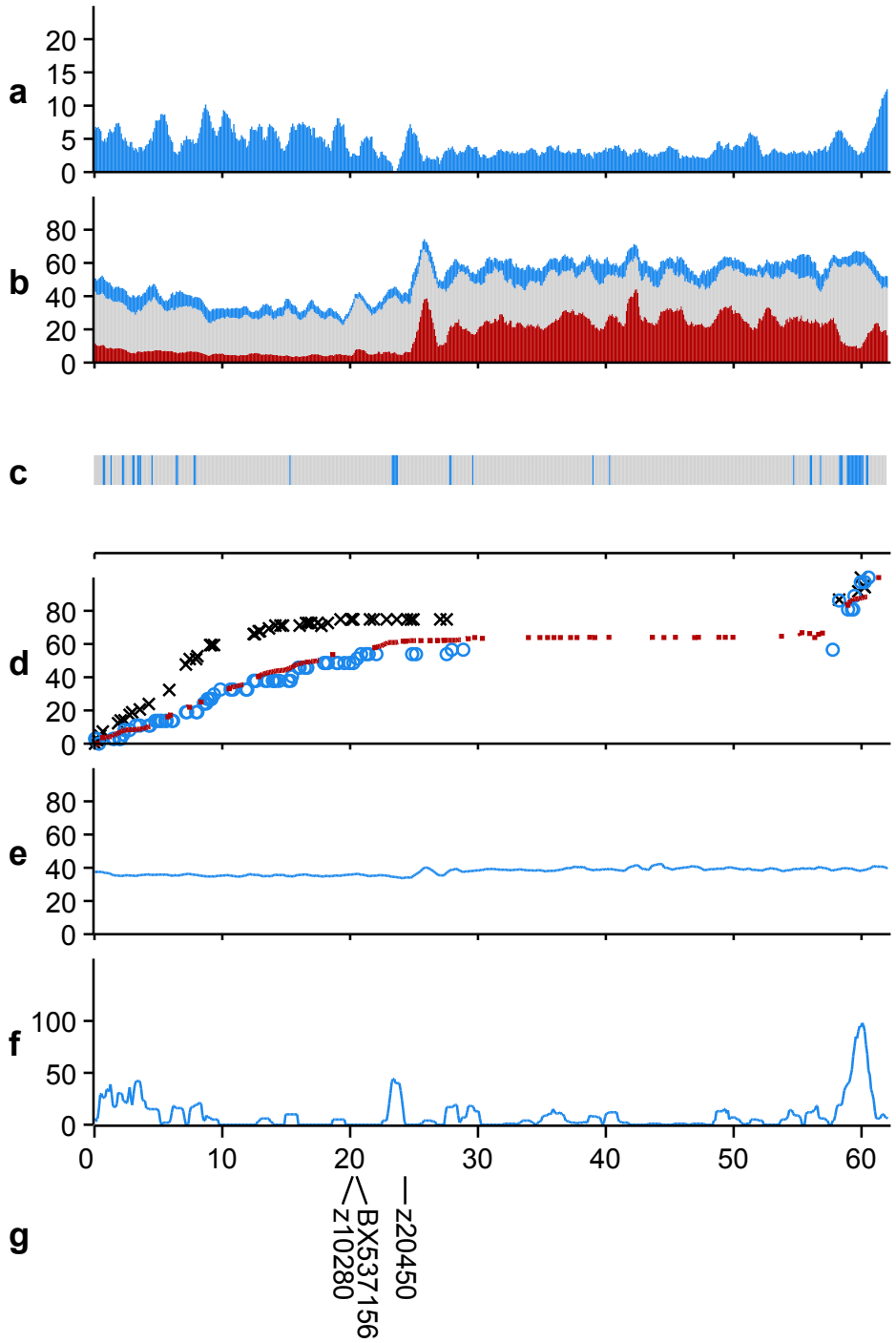
Supplementary Figure A1 | Landscape of Chromosome 1.
a, Percent coverage by exons of protein coding genes. **b**, Stacked repeat coverage, divided into Type I TEs (red), Type II TEs (grey) and other repeat types (blue) including dust, tandem and satellite repeats. **c**, Sequence composition (grey bars = clones, blue bars = whole genome shotgun contigs). **d**, Genetic marker placements (red = SATmap markers, blue = Heat Shock (HS) meiotic map markers, black = MGH meiotic map markers). Marker placements have been normalised to make the maps comparable. **e**, GC content. **f**, Gap density. **g**, Reported near centromeric markers (see Supplementary Information). The x-axis shows the chromosomal position in Mb. **a** and **b** were calculated as percentage coverage over 1Mb overlapping windows, with a 100kb shift between each window. **c**, **d**, **e** and **f** were calculated over 100 kb windows.



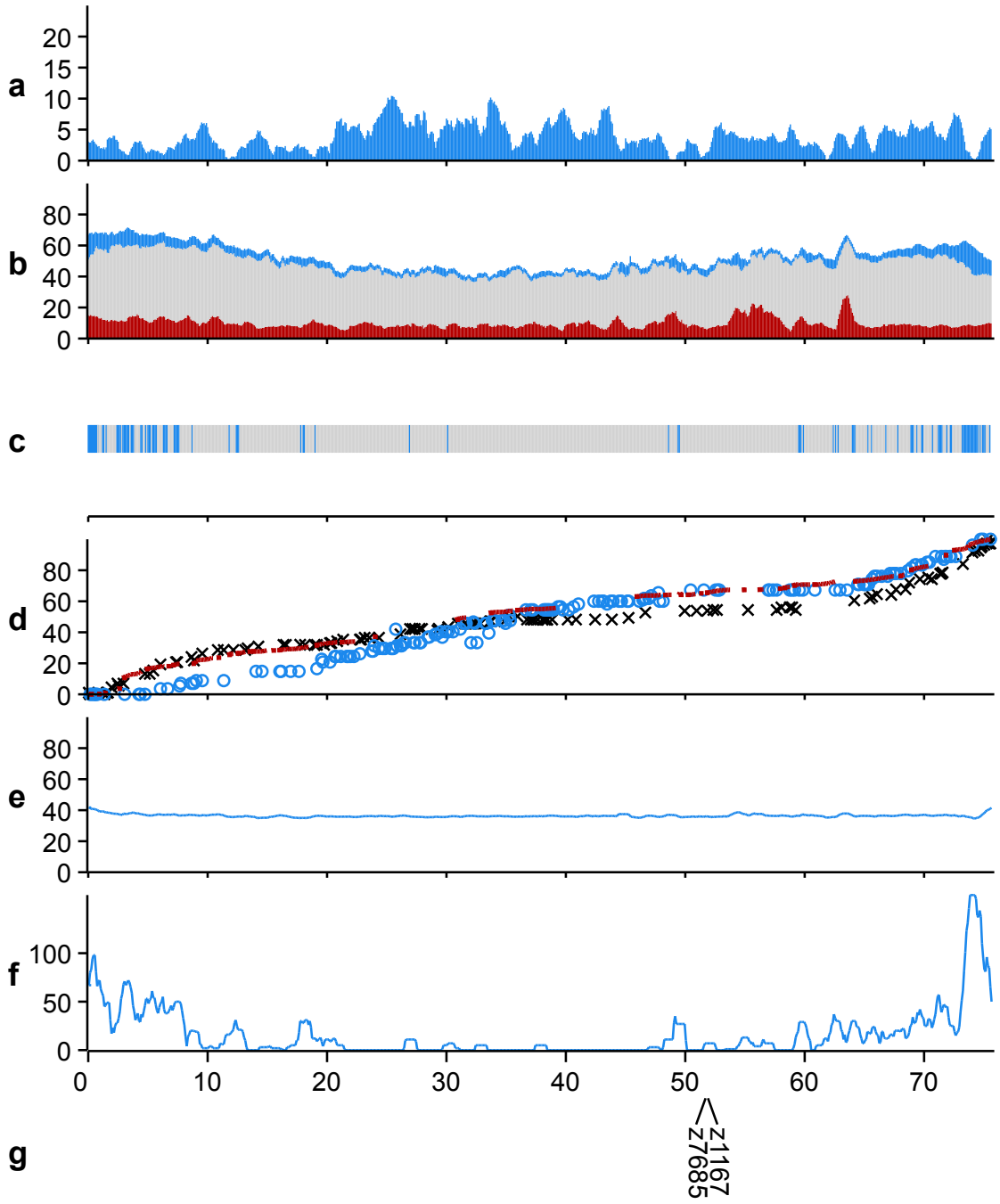
Supplementary Figure A2 | Landscape of Chromosome 2.
a, Percent coverage by exons of protein coding genes. **b**, Stacked repeat coverage, divided into Type I TEs (red), Type II TEs (grey) and other repeat types (blue) including dust, tandem and satellite repeats. **c**, Sequence composition (grey bars = clones, blue bars = whole genome shotgun contigs). **d**, Genetic marker placements (red = SATmap markers, blue = Heat Shock (HS) meiotic map markers, black = MGH meiotic map markers). Marker placements have been normalised to make the maps comparable. **e**, GC content. **f**, Gap density. **g**, Reported near centromeric markers (see Supplementary Information). The x-axis shows the chromosomal position in Mb. **a** and **b** were calculated as percentage coverage over 1Mb overlapping windows, with a 100kb shift between each window. **c**, **d**, **e** and **f** were calculated over 100 kb windows.



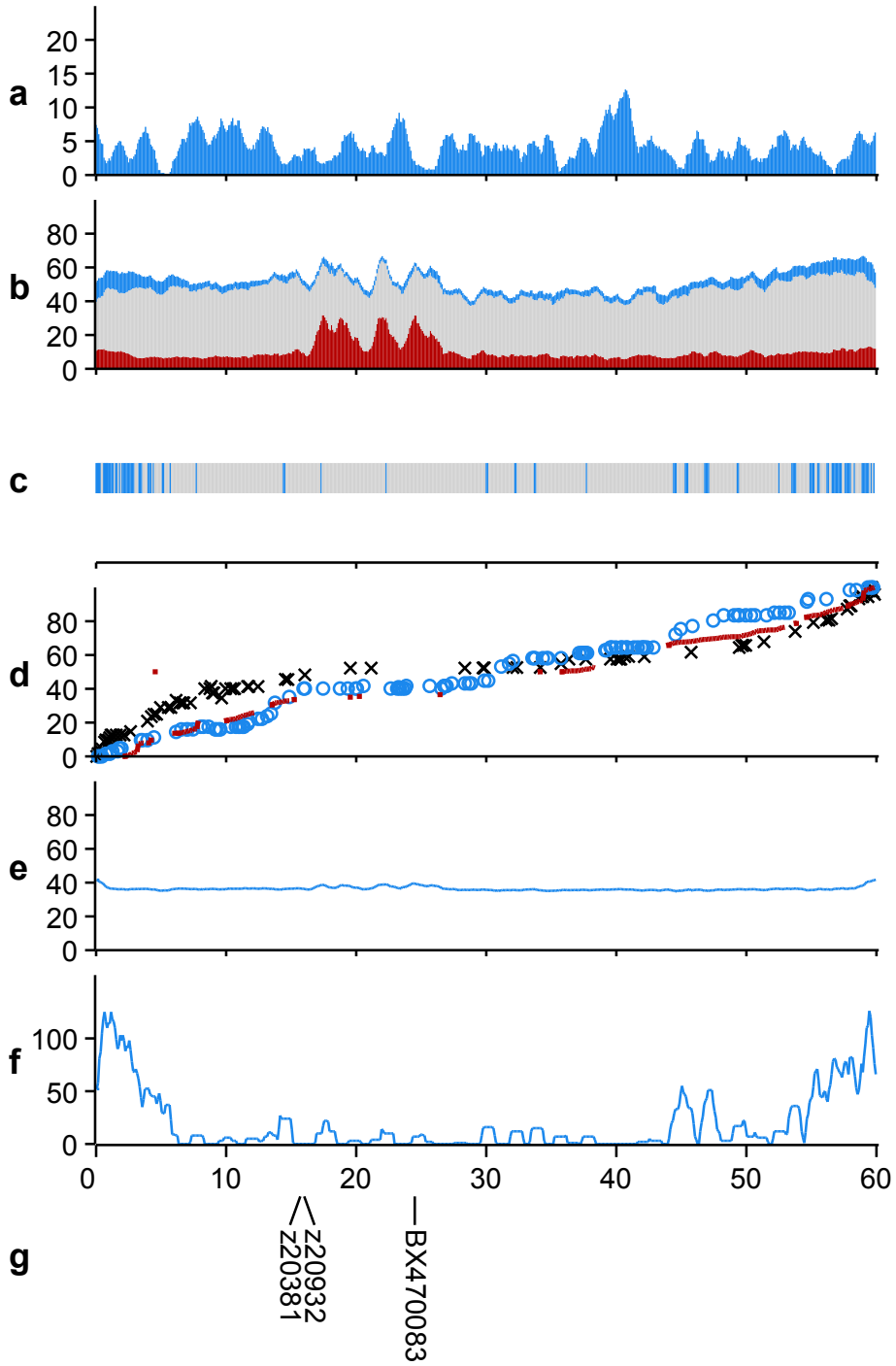
Supplementary Figure A3 | Landscape of Chromosome 3.
a, Percent coverage by exons of protein coding genes. **b**, Stacked repeat coverage, divided into Type I TEs (red), Type II TEs (grey) and other repeat types (blue) including dust, tandem and satellite repeats. **c**, Sequence composition (grey bars = clones, blue bars = whole genome shotgun contigs). **d**, Genetic marker placements (red = SATmap markers, blue = Heat Shock (HS) meiotic map markers, black = MGH meiotic map markers). Marker placements have been normalised to make the maps comparable. **e**, GC content. **f**, Gap density. **g**, Reported near centromeric markers (see Supplementary Information). The x-axis shows the chromosomal position in Mb. **a** and **b** were calculated as percentage coverage over 1Mb overlapping windows, with a 100kb shift between each window. **c**, **d**, **e** and **f** were calculated over 100 kb windows.



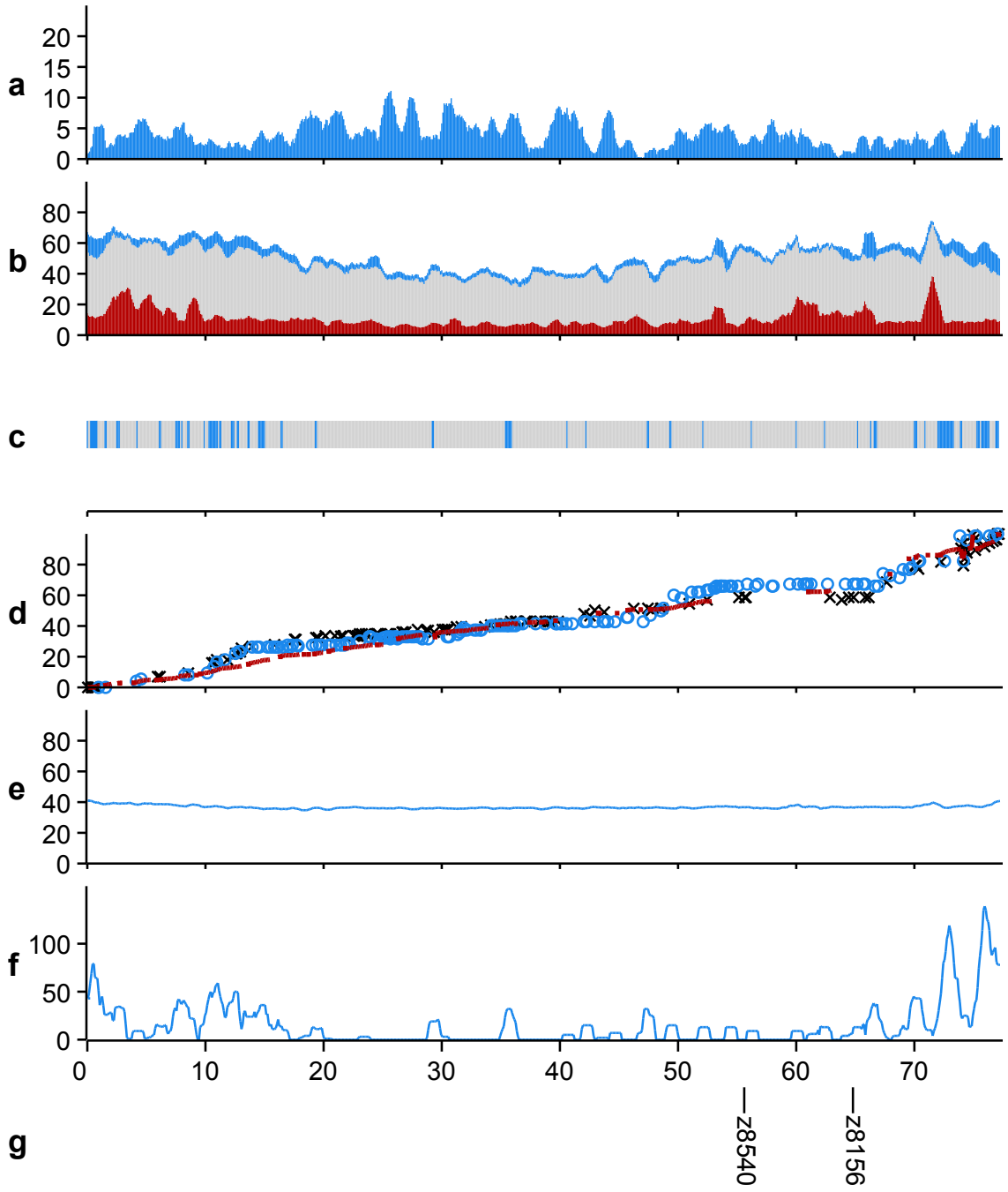
Supplementary Figure A4 | Landscape of Chromosome 4.
a, Percent coverage by exons of protein coding genes. **b**, Stacked repeat coverage, divided into Type I TEs (red), Type II TEs (grey) and other repeat types (blue) including dust, tandem and satellite repeats. **c**, Sequence composition (grey bars = clones, blue bars = whole genome shotgun contigs). **d**, Genetic marker placements (red = SATmap markers, blue = Heat Shock (HS) meiotic map markers, black = MGH meiotic map markers). Marker placements have been normalised to make the maps comparable. **e**, GC content. **f**, Gap density. **g**, Reported near centromeric markers (see Supplementary Information). The x-axis shows the chromosomal position in Mb. **a** and **b** were calculated as percentage coverage over 1Mb overlapping windows, with a 100kb shift between each window. **c**, **d**, **e** and **f** were calculated over 100 kb windows.



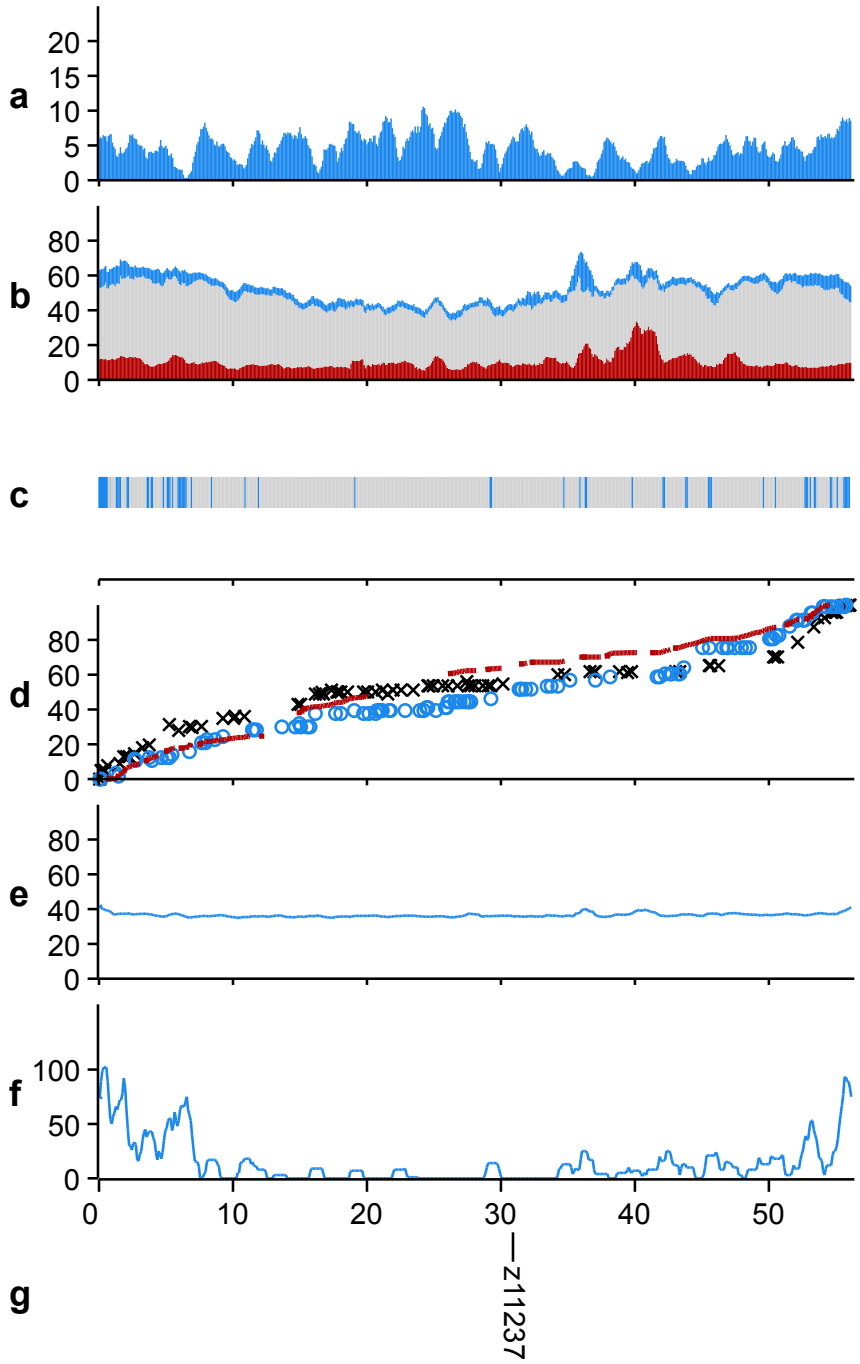
Supplementary Figure A5 | Landscape of Chromosome 5.
a, Percent coverage by exons of protein coding genes. **b**, Stacked repeat coverage, divided into Type I TEs (red), Type II TEs (grey) and other repeat types (blue) including dust, tandem and satellite repeats. **c**, Sequence composition (grey bars = clones, blue bars = whole genome shotgun contigs). **d**, Genetic marker placements (red = SATmap markers, blue = Heat Shock (HS) meiotic map markers, black = MGH meiotic map markers). Marker placements have been normalised to make the maps comparable. **e**, GC content. **f**, Gap density. **g**, Reported near centromeric markers (see Supplementary Information). The x-axis shows the chromosomal position in Mb. **a** and **b** were calculated as percentage coverage over 1Mb overlapping windows, with a 100kb shift between each window. **c**, **d**, **e** and **f** were calculated over 100 kb windows.



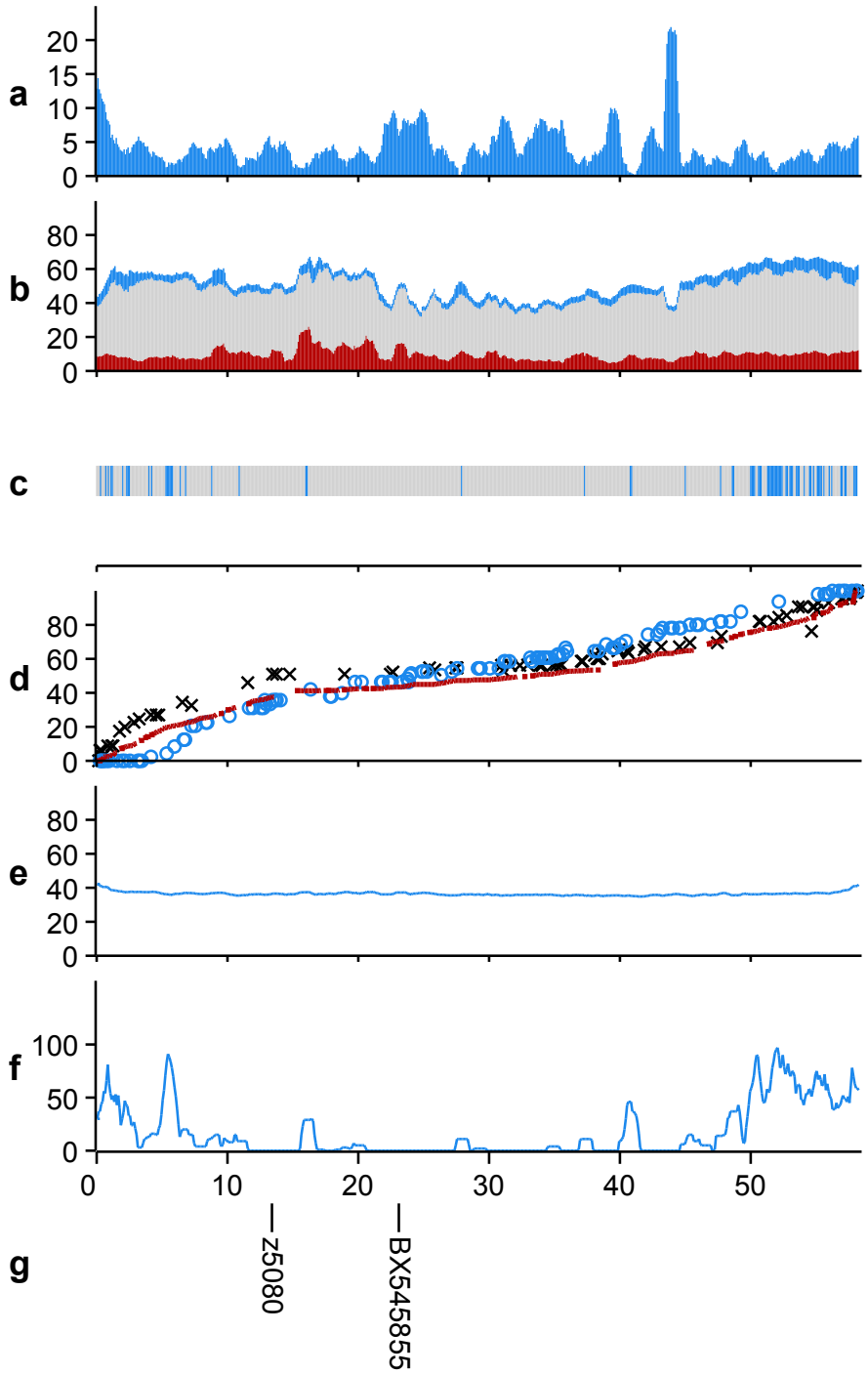
Supplementary Figure A6 | Landscape of Chromosome 6.
a, Percent coverage by exons of protein coding genes. **b**, Stacked repeat coverage, divided into Type I TEs (red), Type II TEs (grey) and other repeat types (blue) including dust, tandem and satellite repeats. **c**, Sequence composition (grey bars = clones, blue bars = whole genome shotgun contigs). **d**, Genetic marker placements (red = SATmap markers, blue = Heat Shock (HS) meiotic map markers, black = MGH meiotic map markers). Marker placements have been normalised to make the maps comparable. **e**, GC content. **f**, Gap density. **g**, Reported near centromeric markers (see Supplementary Information). The x-axis shows the chromosomal position in Mb. **a** and **b** were calculated as percentage coverage over 1Mb overlapping windows, with a 100kb shift between each window. **c**, **d**, **e** and **f** were calculated over 100 kb windows.



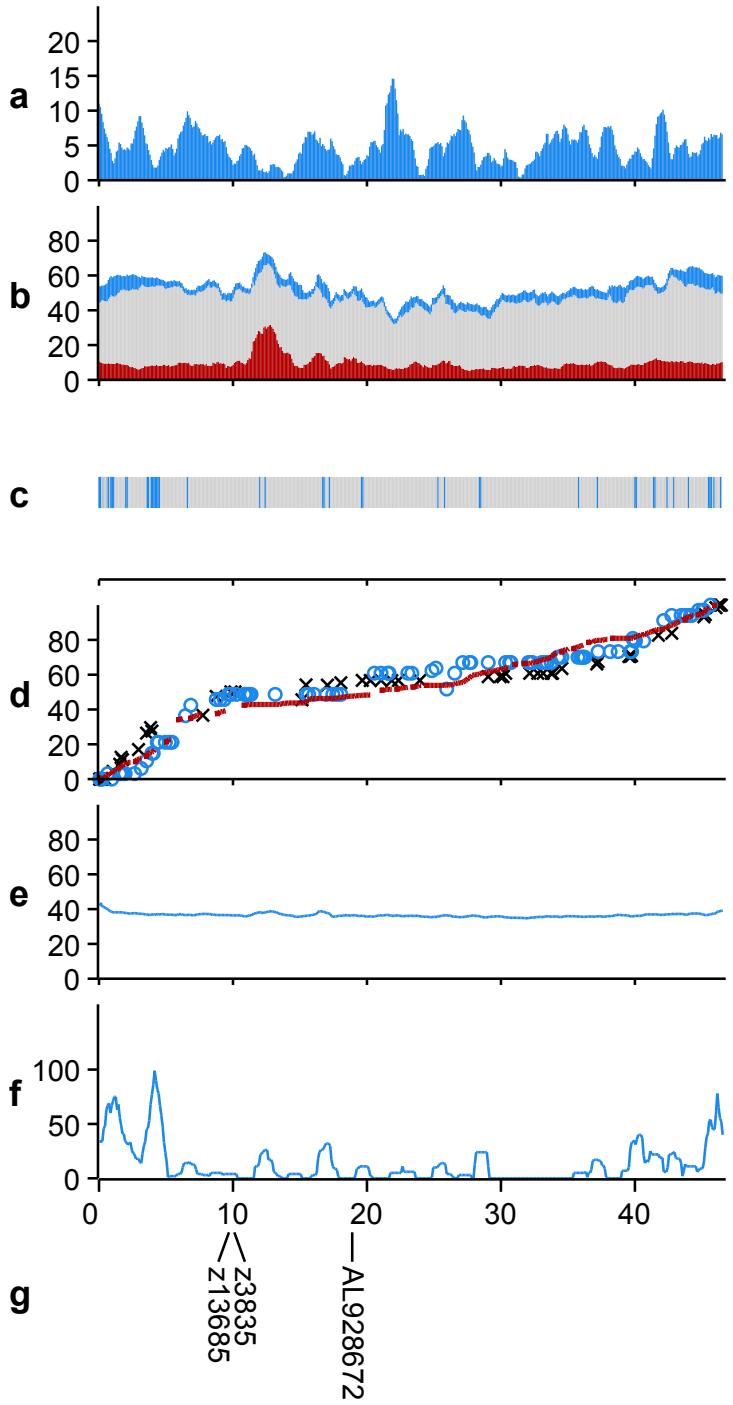
Supplementary Figure A7 | Landscape of Chromosome 7.
a, Percent coverage by exons of protein coding genes. **b**, Stacked repeat coverage, divided into Type I TEs (red), Type II TEs (grey) and other repeat types (blue) including dust, tandem and satellite repeats. **c**, Sequence composition (grey bars = clones, blue bars = whole genome shotgun contigs). **d**, Genetic marker placements (red = SATmap markers, blue = Heat Shock (HS) meiotic map markers, black = MGH meiotic map markers). Marker placements have been normalised to make the maps comparable. **e**, GC content. **f**, Gap density. **g**, Reported near centromeric markers (see Supplementary Information). The x-axis shows the chromosomal position in Mb. **a** and **b** were calculated as percentage coverage over 1Mb overlapping windows, with a 100kb shift between each window. **c**, **d**, **e** and **f** were calculated over 100 kb windows.



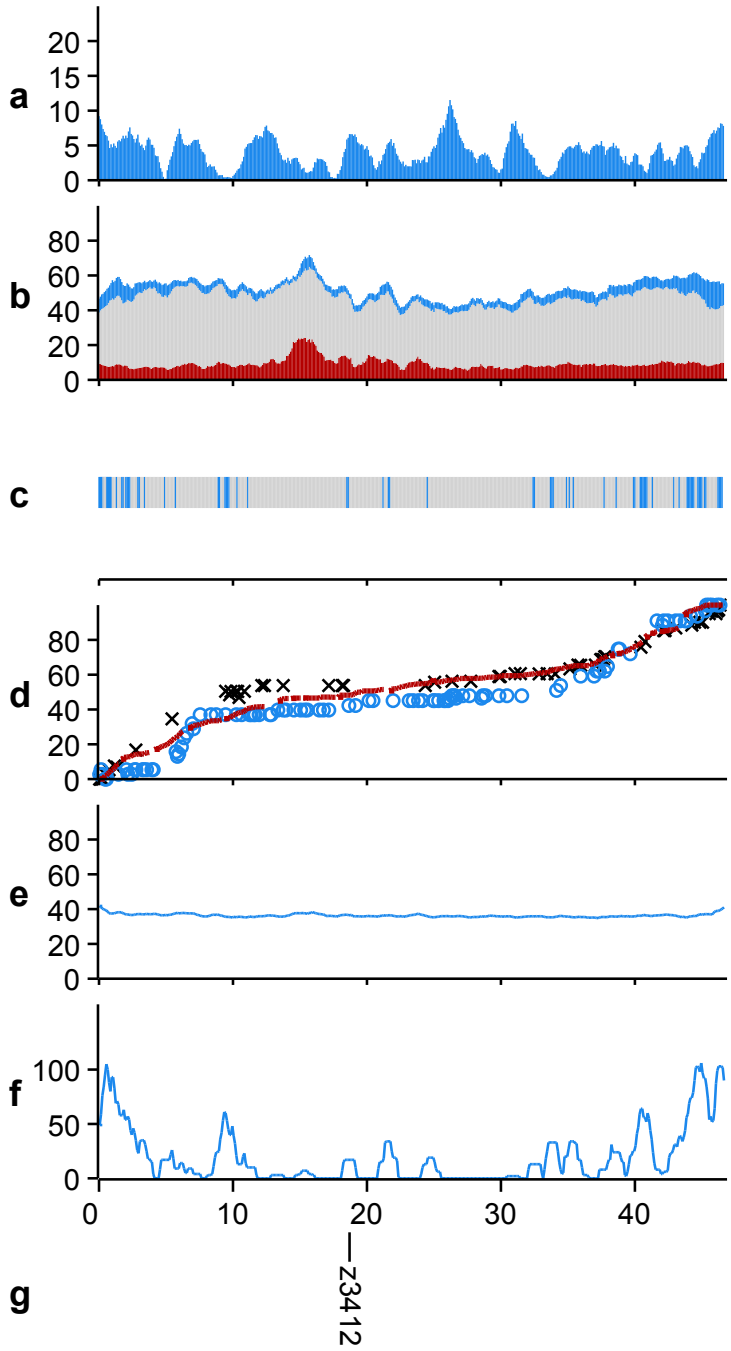
Supplementary Figure A8 | Landscape of Chromosome 8.
a, Percent coverage by exons of protein coding genes. **b**, Stacked repeat coverage, divided into Type I TEs (red), Type II TEs (grey) and other repeat types (blue) including dust, tandem and satellite repeats. **c**, Sequence composition (grey bars = clones, blue bars = whole genome shotgun contigs). **d**, Genetic marker placements (red = SATmap markers, blue = Heat Shock (HS) meiotic map markers, black = MGH meiotic map markers). Marker placements have been normalised to make the maps comparable. **e**, GC content. **f**, Gap density. **g**, Reported near centromeric markers (see Supplementary Information). The x-axis shows the chromosomal position in Mb. **a** and **b** were calculated as percentage coverage over 1Mb overlapping windows, with a 100kb shift between each window. **c**, **d**, **e** and **f** were calculated over 100 kb windows.



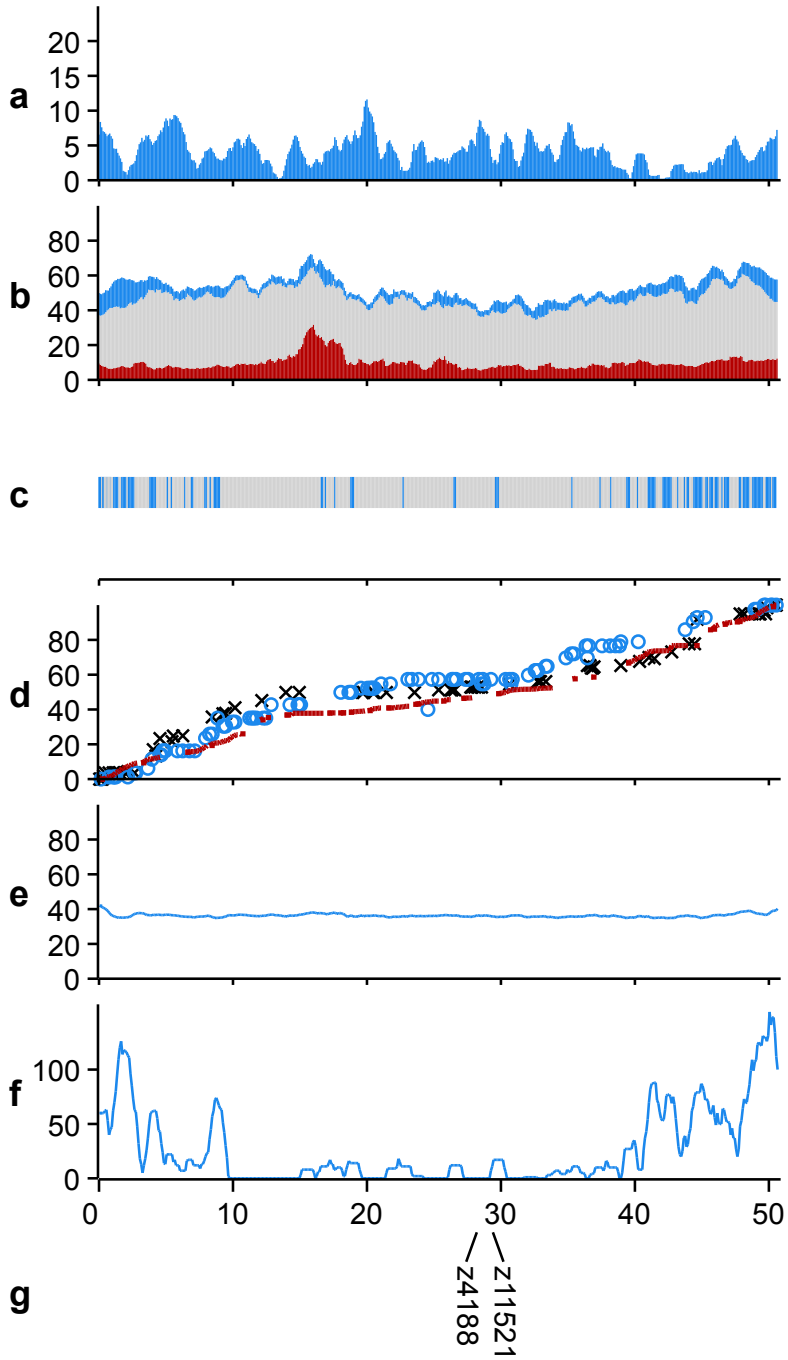
Supplementary Figure A9 | Landscape of Chromosome 9.
a, Percent coverage by exons of protein coding genes. **b**, Stacked repeat coverage, divided into Type I TEs (red), Type II TEs (grey) and other repeat types (blue) including dust, tandem and satellite repeats. **c**, Sequence composition (grey bars = clones, blue bars = whole genome shotgun contigs). **d**, Genetic marker placements (red = SATmap markers, blue = Heat Shock (HS) meiotic map markers, black = MGH meiotic map markers). Marker placements have been normalised to make the maps comparable. **e**, GC content. **f**, Gap density. **g**, Reported near centromeric markers (see Supplementary Information). The x-axis shows the chromosomal position in Mb. **a** and **b** were calculated as percentage coverage over 1Mb overlapping windows, with a 100kb shift between each window. **c**, **d**, **e** and **f** were calculated over 100 kb windows.



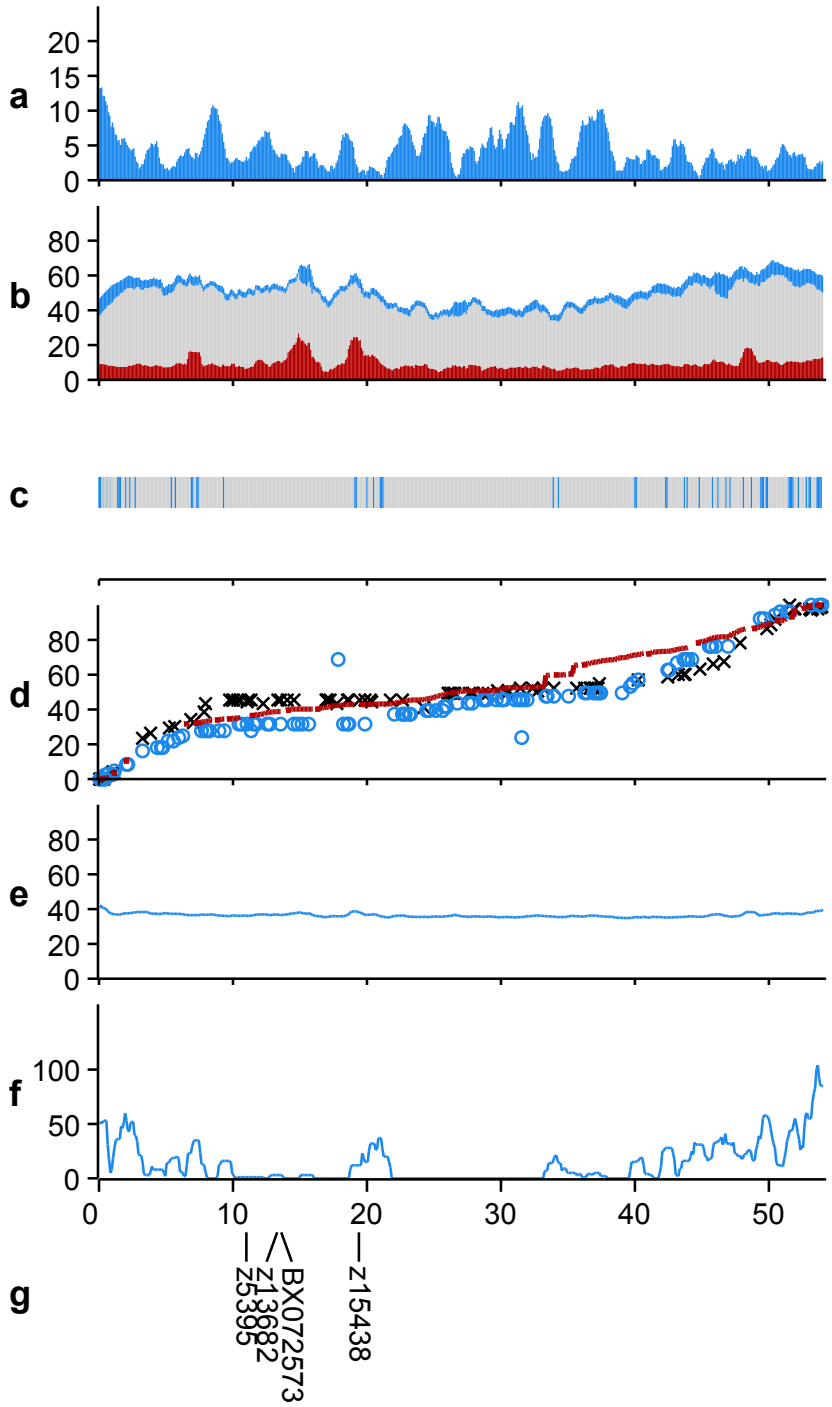
Supplementary Figure A10 | Landscape of Chromosome 10.
a, Percent coverage by exons of protein coding genes. **b**, Stacked repeat coverage, divided into Type I TEs (red), Type II TEs (grey) and other repeat types (blue) including dust, tandem and satellite repeats. **c**, Sequence composition (grey bars = clones, blue bars = whole genome shotgun contigs). **d**, Genetic marker placements (red = SATmap markers, blue = Heat Shock (HS) meiotic map markers, black = MGH meiotic map markers). Marker placements have been normalised to make the maps comparable. **e**, GC content. **f**, Gap density. **g**, Reported near centromeric markers (see Supplementary Information). The x-axis shows the chromosomal position in Mb. **a** and **b** were calculated as percentage coverage over 1Mb overlapping windows, with a 100kb shift between each window. **c**, **d**, **e** and **f** were calculated over 100 kb windows.



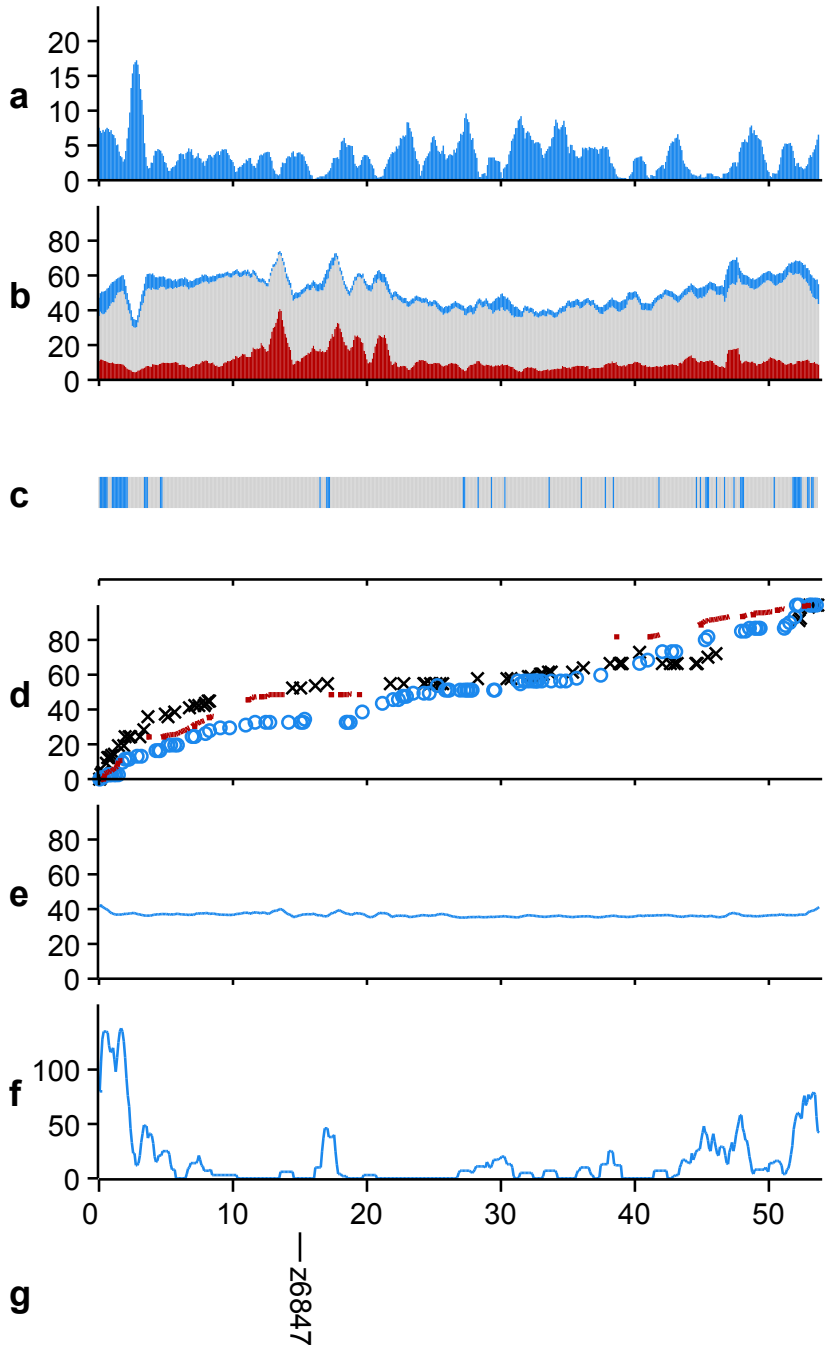
Supplementary Figure A11 | Landscape of Chromosome 11.
a, Percent coverage by exons of protein coding genes. **b**, Stacked repeat coverage, divided into Type I TEs (red), Type II TEs (grey) and other repeat types (blue) including dust, tandem and satellite repeats. **c**, Sequence composition (grey bars = clones, blue bars = whole genome shotgun contigs). **d**, Genetic marker placements (red = SATmap markers, blue = Heat Shock (HS) meiotic map markers, black = MGH meiotic map markers). Marker placements have been normalised to make the maps comparable. **e**, GC content. **f**, Gap density. **g**, Reported near centromeric markers (see Supplementary Information). The x-axis shows the chromosomal position in Mb. **a** and **b** were calculated as percentage coverage over 1Mb overlapping windows, with a 100kb shift between each window. **c**, **d**, **e** and **f** were calculated over 100 kb windows.



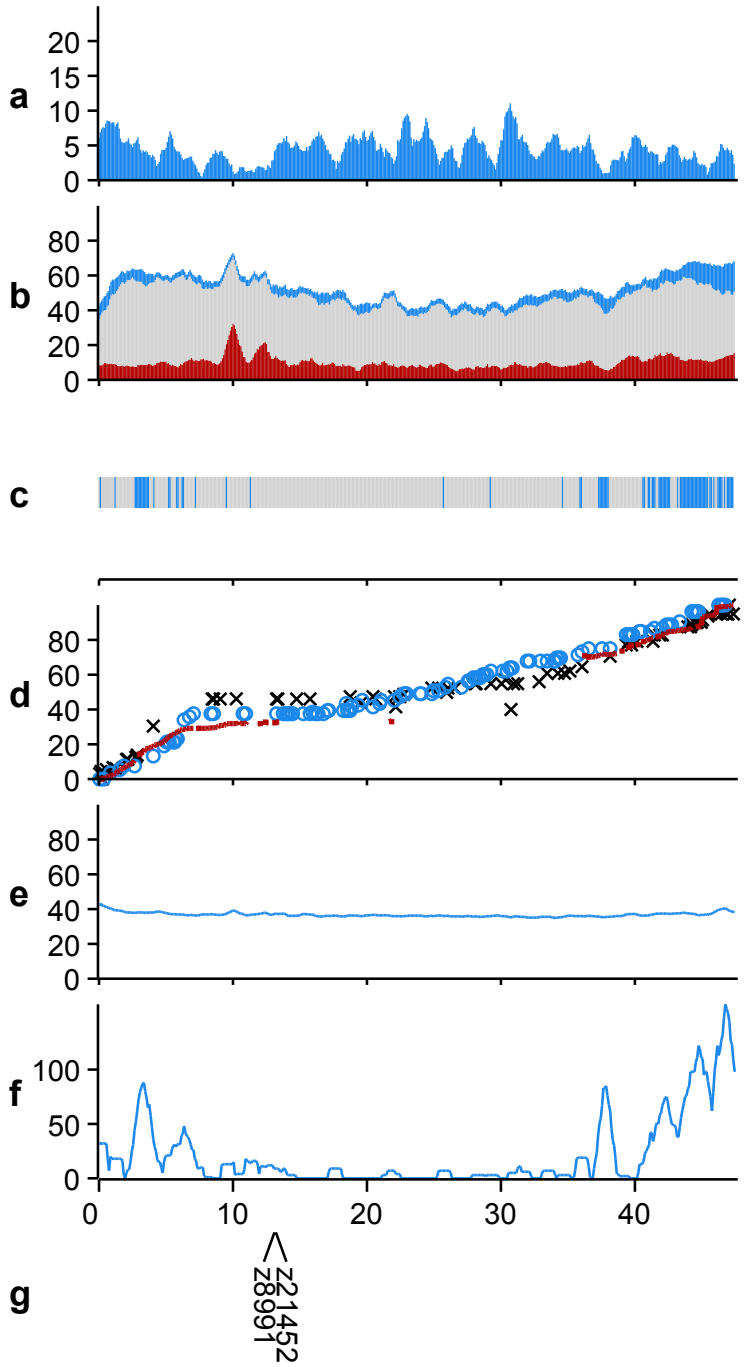
Supplementary Figure A12 | Landscape of Chromosome 12.
a, Percent coverage by exons of protein coding genes. **b**, Stacked repeat coverage, divided into Type I TEs (red), Type II TEs (grey) and other repeat types (blue) including dust, tandem and satellite repeats. **c**, Sequence composition (grey bars = clones, blue bars = whole genome shotgun contigs). **d**, Genetic marker placements (red = SATmap markers, blue = Heat Shock (HS) meiotic map markers, black = MGH meiotic map markers). Marker placements have been normalised to make the maps comparable. **e**, GC content. **f**, Gap density. **g**, Reported near centromeric markers (see Supplementary Information). The x-axis shows the chromosomal position in Mb. **a** and **b** were calculated as percentage coverage over 1Mb overlapping windows, with a 100kb shift between each window. **c**, **d**, **e** and **f** were calculated over 100 kb windows.



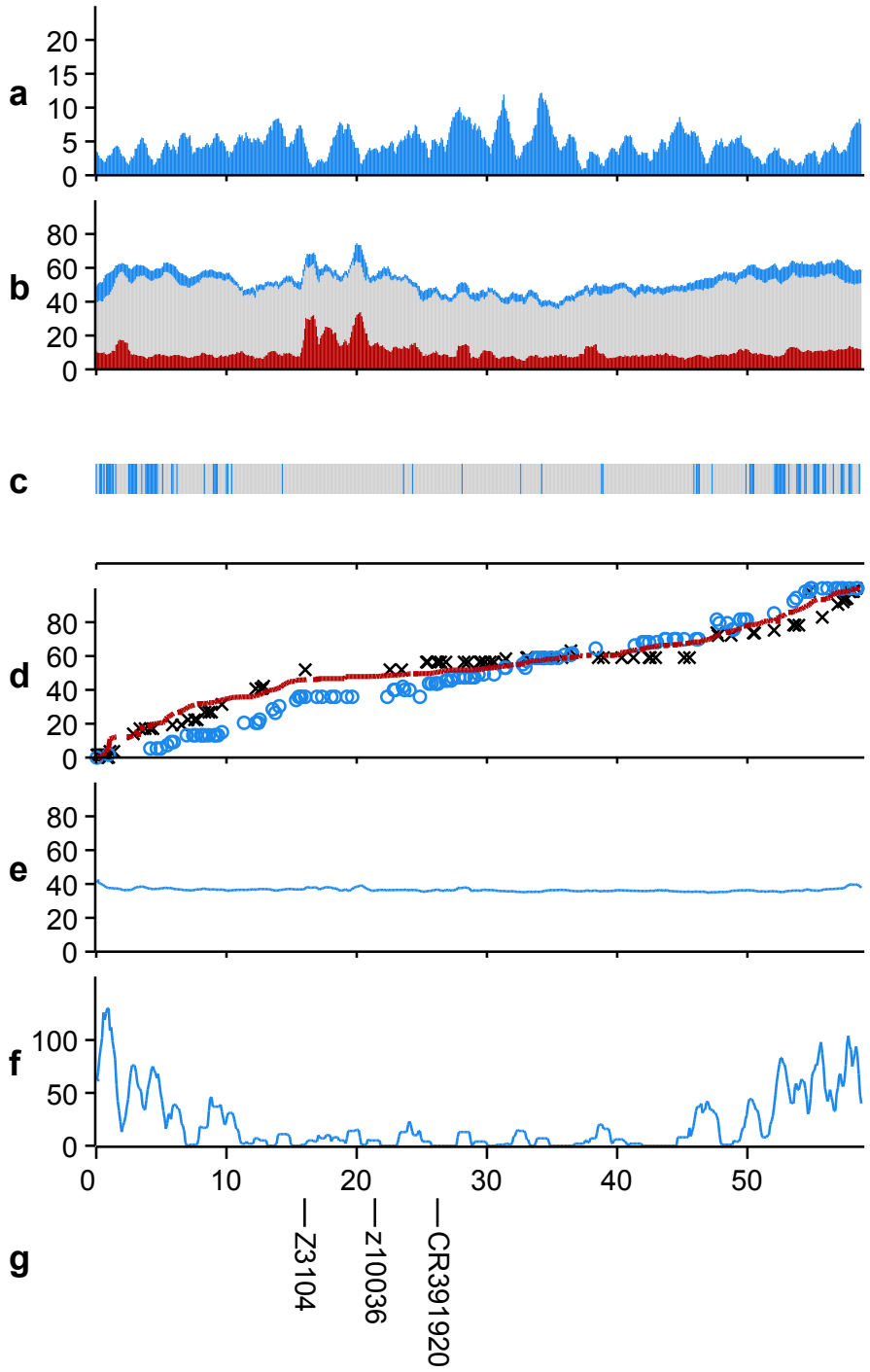
Supplementary Figure A13 | Landscape of Chromosome 13.
a, Percent coverage by exons of protein coding genes. **b**, Stacked repeat coverage, divided into Type I TEs (red), Type II TEs (grey) and other repeat types (blue) including dust, tandem and satellite repeats. **c**, Sequence composition (grey bars = clones, blue bars = whole genome shotgun contigs). **d**, Genetic marker placements (red = SATmap markers, blue = Heat Shock (HS) meiotic map markers, black = MGH meiotic map markers). Marker placements have been normalised to make the maps comparable. **e**, GC content. **f**, Gap density. **g**, Reported near centromeric markers (see Supplementary Information). The x-axis shows the chromosomal position in Mb. **a** and **b** were calculated as percentage coverage over 1Mb overlapping windows, with a 100kb shift between each window. **c**, **d**, **e** and **f** were calculated over 100 kb windows.



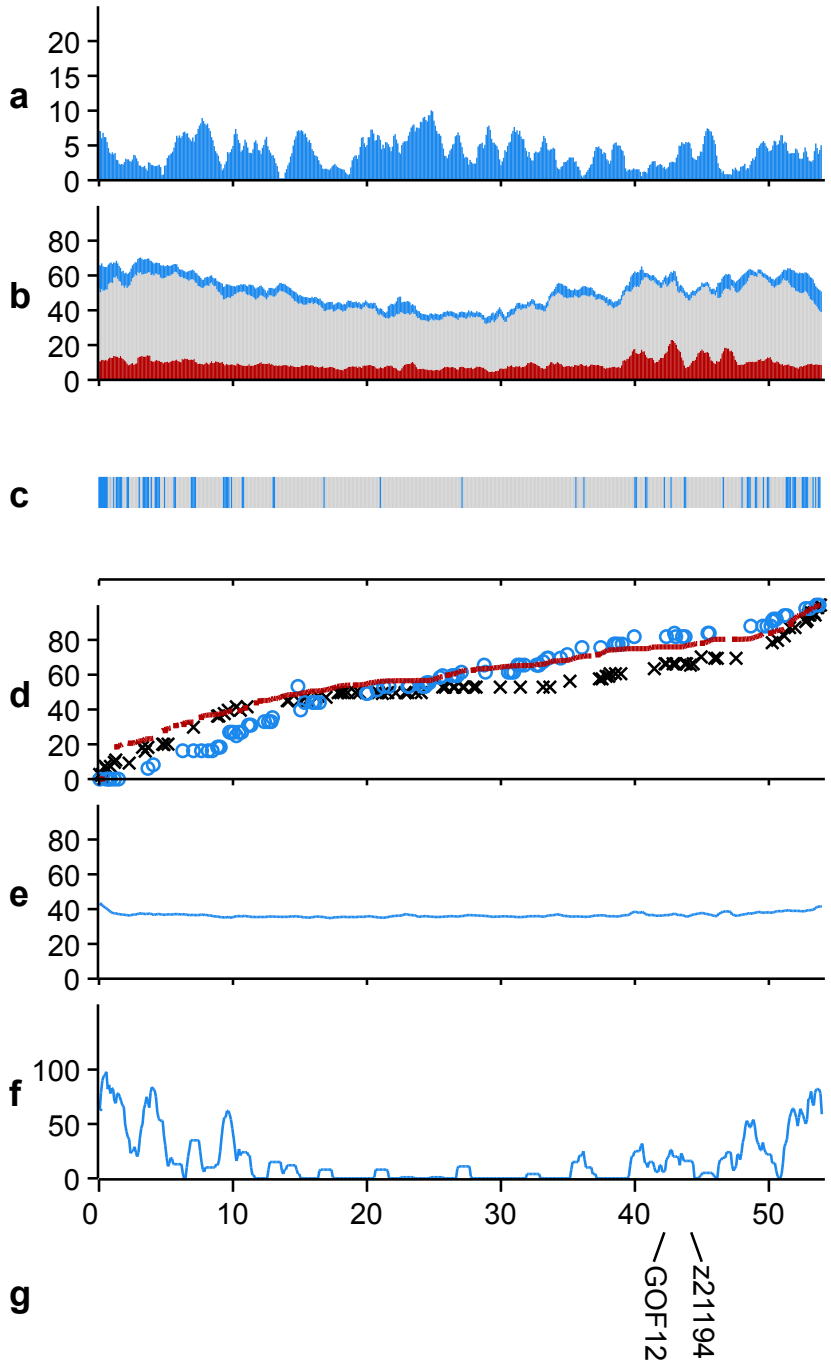
Supplementary Figure A14 | Landscape of Chromosome 14.
a, Percent coverage by exons of protein coding genes. **b**, Stacked repeat coverage, divided into Type I TEs (red), Type II TEs (grey) and other repeat types (blue) including dust, tandem and satellite repeats. **c**, Sequence composition (grey bars = clones, blue bars = whole genome shotgun contigs). **d**, Genetic marker placements (red = SATmap markers, blue = Heat Shock (HS) meiotic map markers, black = MGH meiotic map markers). Marker placements have been normalised to make the maps comparable. **e**, GC content. **f**, Gap density. **g**, Reported near centromeric markers (see Supplementary Information). The x-axis shows the chromosomal position in Mb. **a** and **b** were calculated as percentage coverage over 1Mb overlapping windows, with a 100kb shift between each window. **c**, **d**, **e** and **f** were calculated over 100 kb windows.



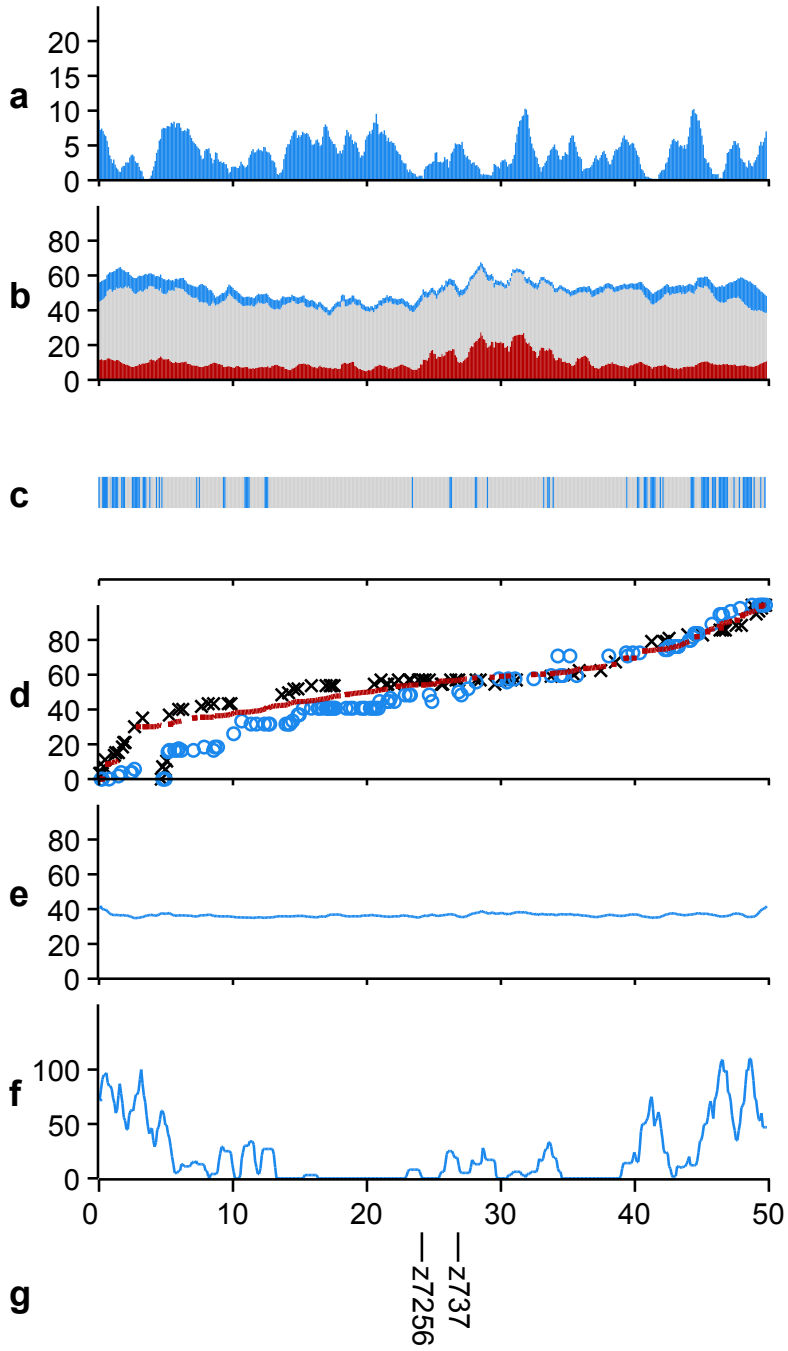
Supplementary Figure A15 | Landscape of Chromosome 15.
a, Percent coverage by exons of protein coding genes. **b**, Stacked repeat coverage, divided into Type I TEs (red), Type II TEs (grey) and other repeat types (blue) including dust, tandem and satellite repeats. **c**, Sequence composition (grey bars = clones, blue bars = whole genome shotgun contigs). **d**, Genetic marker placements (red = SATmap markers, blue = Heat Shock (HS) meiotic map markers, black = MGH meiotic map markers). Marker placements have been normalised to make the maps comparable. **e**, GC content. **f**, Gap density. **g**, Reported near centromeric markers (see Supplementary Information). The x-axis shows the chromosomal position in Mb. **a** and **b** were calculated as percentage coverage over 1Mb overlapping windows, with a 100kb shift between each window. **c**, **d**, **e** and **f** were calculated over 100 kb windows.



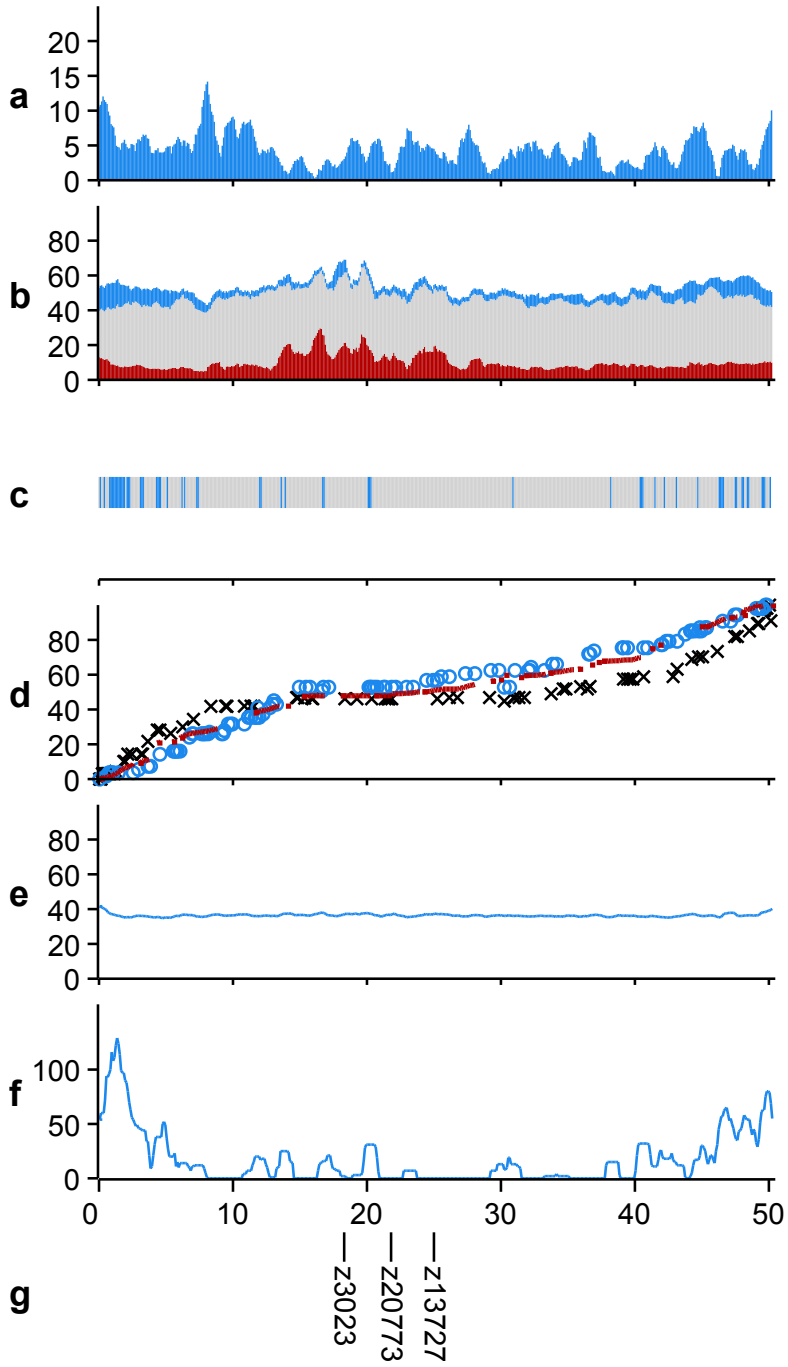
Supplementary Figure A16 | Landscape of Chromosome 16.
a, Percent coverage by exons of protein coding genes. **b**, Stacked repeat coverage, divided into Type I TEs (red), Type II TEs (grey) and other repeat types (blue) including dust, tandem and satellite repeats. **c**, Sequence composition (grey bars = clones, blue bars = whole genome shotgun contigs). **d**, Genetic marker placements (red = SATmap markers, blue = Heat Shock (HS) meiotic map markers, black = MGH meiotic map markers). Marker placements have been normalised to make the maps comparable. **e**, GC content. **f**, Gap density. **g**, Reported near centromeric markers (see Supplementary Information). The x-axis shows the chromosomal position in Mb. **a** and **b** were calculated as percentage coverage over 1Mb overlapping windows, with a 100kb shift between each window. **c**, **d**, **e** and **f** were calculated over 100 kb windows.



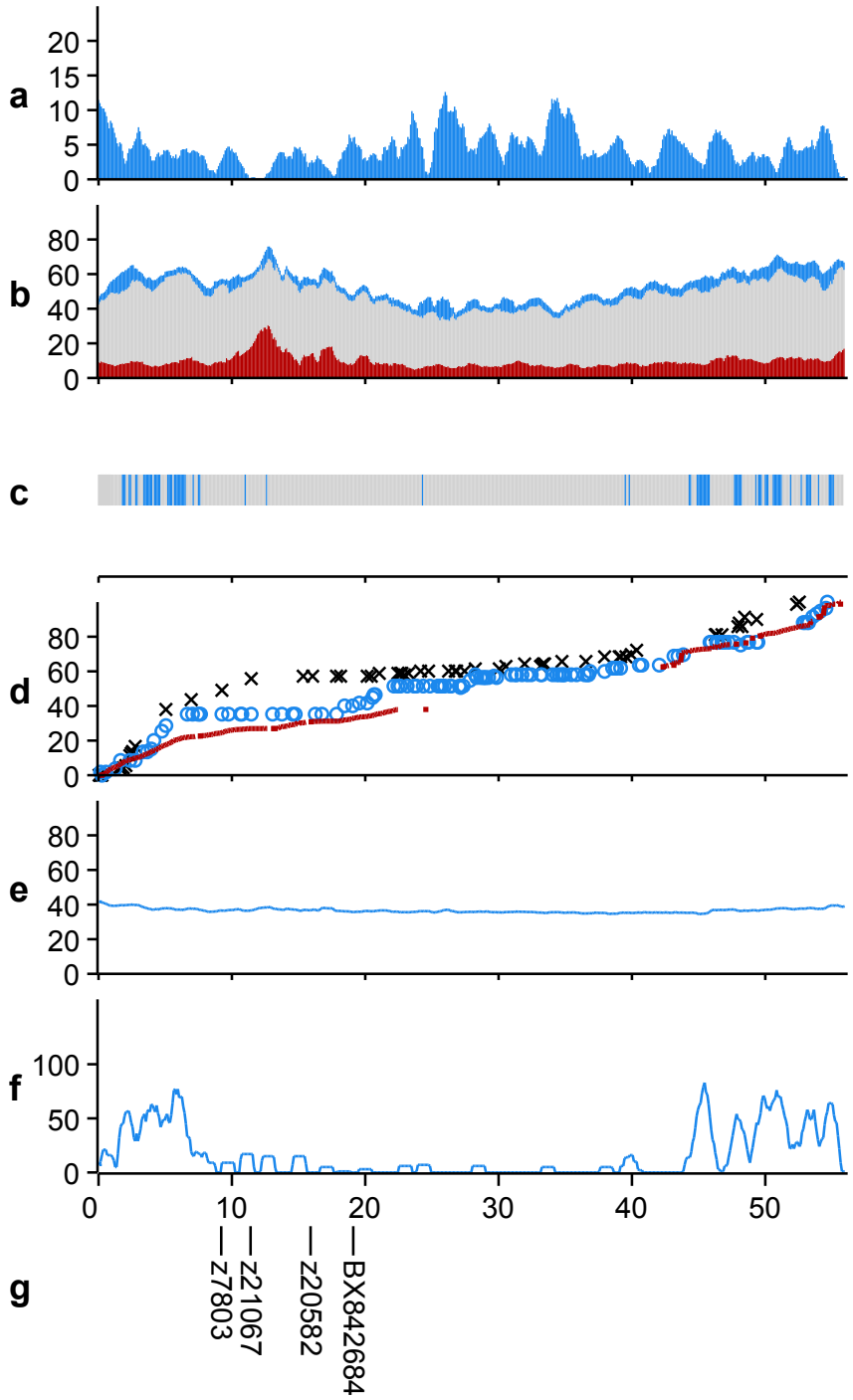
Supplementary Figure A17 | Landscape of Chromosome 17.
a, Percent coverage by exons of protein coding genes. **b**, Stacked repeat coverage, divided into Type I TEs (red), Type II TEs (grey) and other repeat types (blue) including dust, tandem and satellite repeats. **c**, Sequence composition (grey bars = clones, blue bars = whole genome shotgun contigs). **d**, Genetic marker placements (red = SATmap markers, blue = Heat Shock (HS) meiotic map markers, black = MGH meiotic map markers). Marker placements have been normalised to make the maps comparable. **e**, GC content. **f**, Gap density. **g**, Reported near centromeric markers (see Supplementary Information). The x-axis shows the chromosomal position in Mb. **a** and **b** were calculated as percentage coverage over 1Mb overlapping windows, with a 100kb shift between each window. **c**, **d**, **e** and **f** were calculated over 100 kb windows.



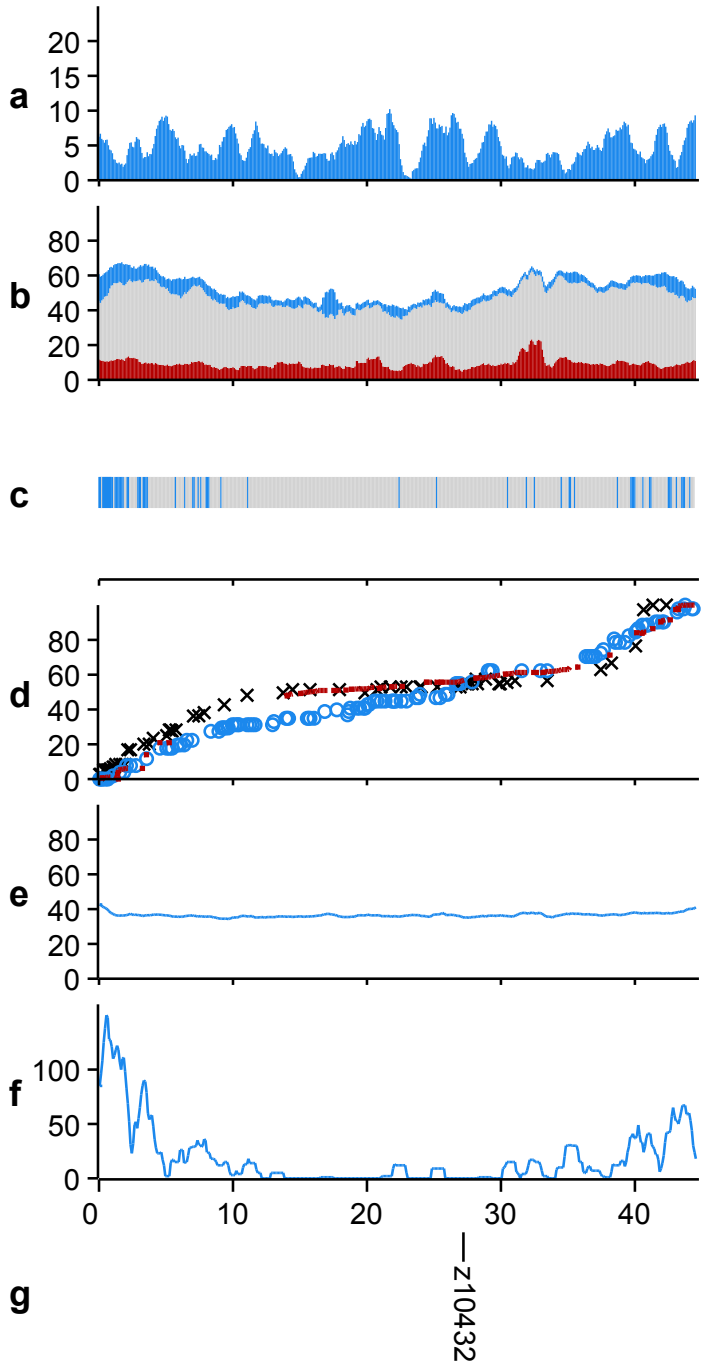
Supplementary Figure A18 | Landscape of Chromosome 18.
a, Percent coverage by exons of protein coding genes. **b**, Stacked repeat coverage, divided into Type I TEs (red), Type II TEs (grey) and other repeat types (blue) including dust, tandem and satellite repeats. **c**, Sequence composition (grey bars = clones, blue bars = whole genome shotgun contigs). **d**, Genetic marker placements (red = SATmap markers, blue = Heat Shock (HS) meiotic map markers, black = MGH meiotic map markers). Marker placements have been normalised to make the maps comparable. **e**, GC content. **f**, Gap density. **g**, Reported near centromeric markers (see Supplementary Information). The x-axis shows the chromosomal position in Mb. **a** and **b** were calculated as percentage coverage over 1Mb overlapping windows, with a 100kb shift between each window. **c**, **d**, **e** and **f** were calculated over 100 kb windows.



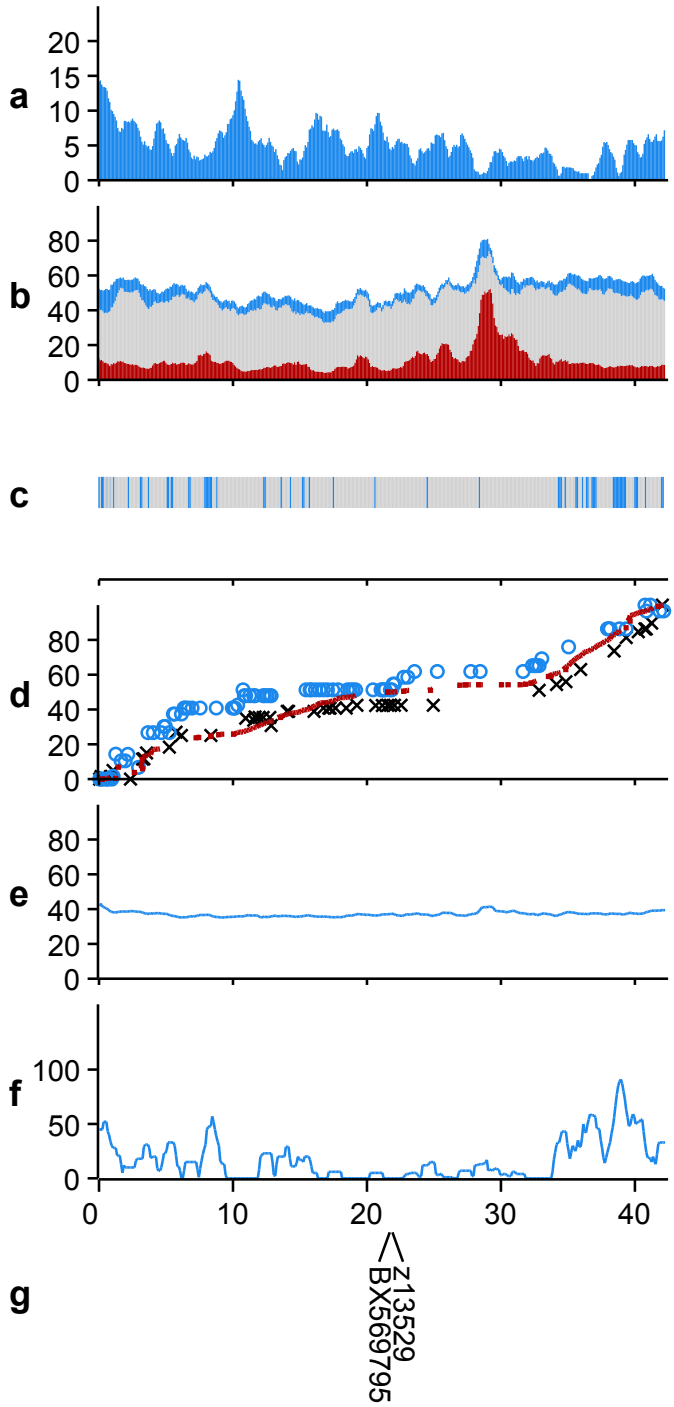
Supplementary Figure A19 | Landscape of Chromosome 19.
a, Percent coverage by exons of protein coding genes. **b**, Stacked repeat coverage, divided into Type I TEs (red), Type II TEs (grey) and other repeat types (blue) including dust, tandem and satellite repeats. **c**, Sequence composition (grey bars = clones, blue bars = whole genome shotgun contigs). **d**, Genetic marker placements (red = SATmap markers, blue = Heat Shock (HS) meiotic map markers, black = MGH meiotic map markers). Marker placements have been normalised to make the maps comparable. **e**, GC content. **f**, Gap density. **g**, Reported near centromeric markers (see Supplementary Information). The x-axis shows the chromosomal position in Mb. **a** and **b** were calculated as percentage coverage over 1Mb overlapping windows, with a 100kb shift between each window. **c**, **d**, **e** and **f** were calculated over 100 kb windows.



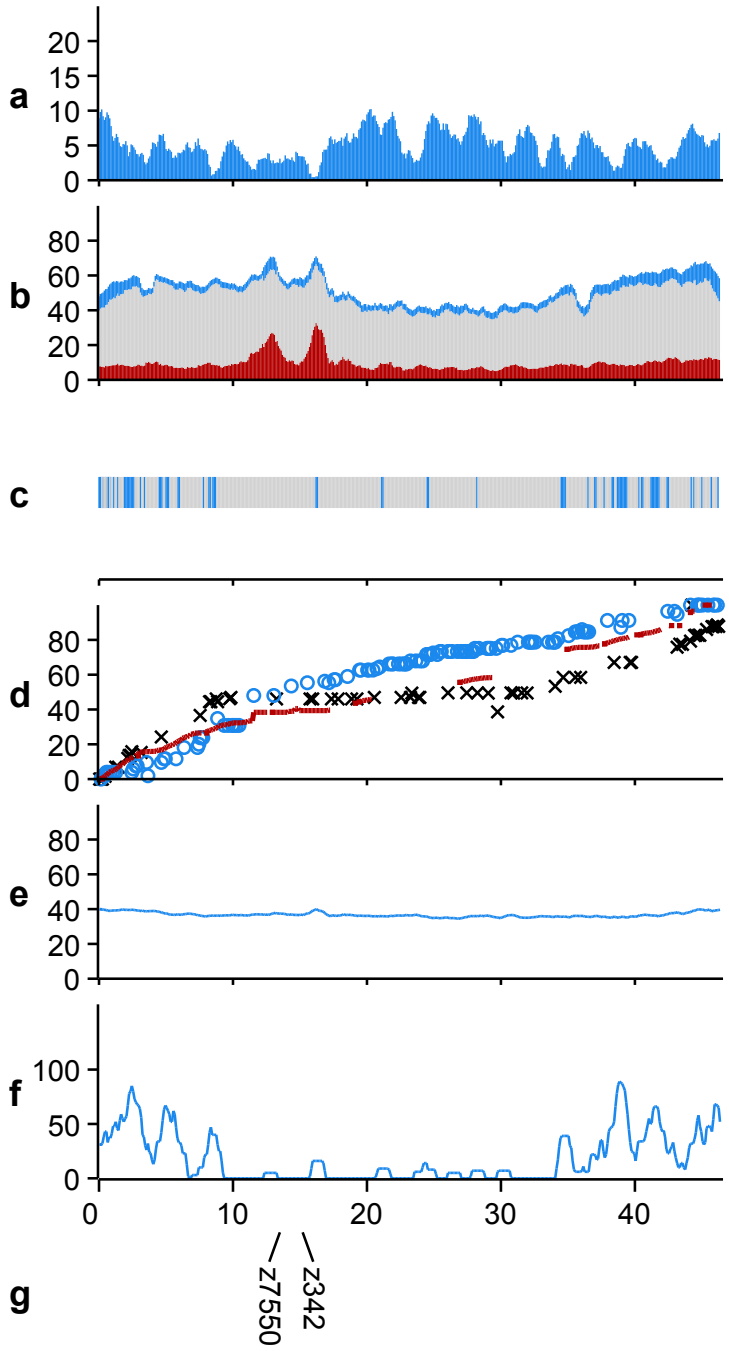
Supplementary Figure A20 | Landscape of Chromosome 20.
a, Percent coverage by exons of protein coding genes. **b**, Stacked repeat coverage, divided into Type I TEs (red), Type II TEs (grey) and other repeat types (blue) including dust, tandem and satellite repeats. **c**, Sequence composition (grey bars = clones, blue bars = whole genome shotgun contigs). **d**, Genetic marker placements (red = SATmap markers, blue = Heat Shock (HS) meiotic map markers, black = MGH meiotic map markers). Marker placements have been normalised to make the maps comparable. **e**, GC content. **f**, Gap density. **g**, Reported near centromeric markers (see Supplementary Information). The x-axis shows the chromosomal position in Mb. **a** and **b** were calculated as percentage coverage over 1Mb overlapping windows, with a 100kb shift between each window. **c**, **d**, **e** and **f** were calculated over 100 kb windows.



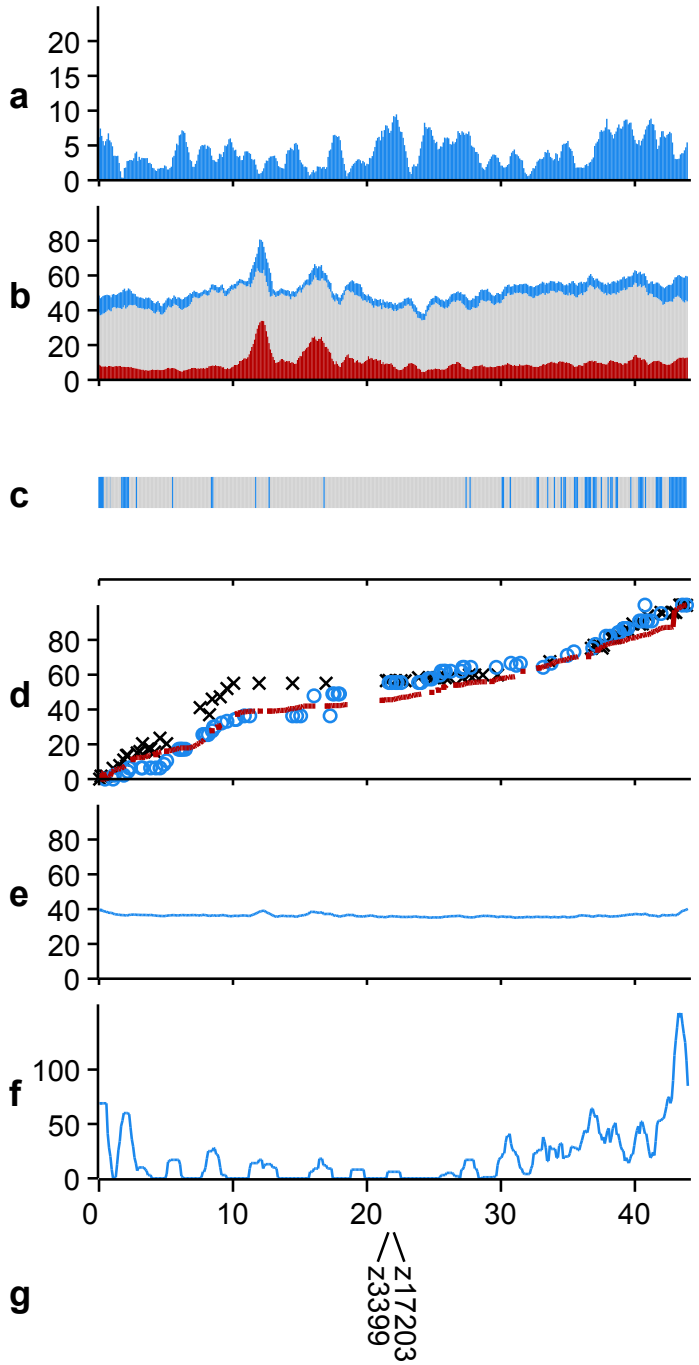
Supplementary Figure A21 | Landscape of Chromosome 21.
a, Percent coverage by exons of protein coding genes. **b**, Stacked repeat coverage, divided into Type I TEs (red), Type II TEs (grey) and other repeat types (blue) including dust, tandem and satellite repeats. **c**, Sequence composition (grey bars = clones, blue bars = whole genome shotgun contigs). **d**, Genetic marker placements (red = SATmap markers, blue = Heat Shock (HS) meiotic map markers, black = MGH meiotic map markers). Marker placements have been normalised to make the maps comparable. **e**, GC content. **f**, Gap density. **g**, Reported near centromeric markers (see Supplementary Information). The x-axis shows the chromosomal position in Mb. **a** and **b** were calculated as percentage coverage over 1Mb overlapping windows, with a 100kb shift between each window. **c**, **d**, **e** and **f** were calculated over 100 kb windows.



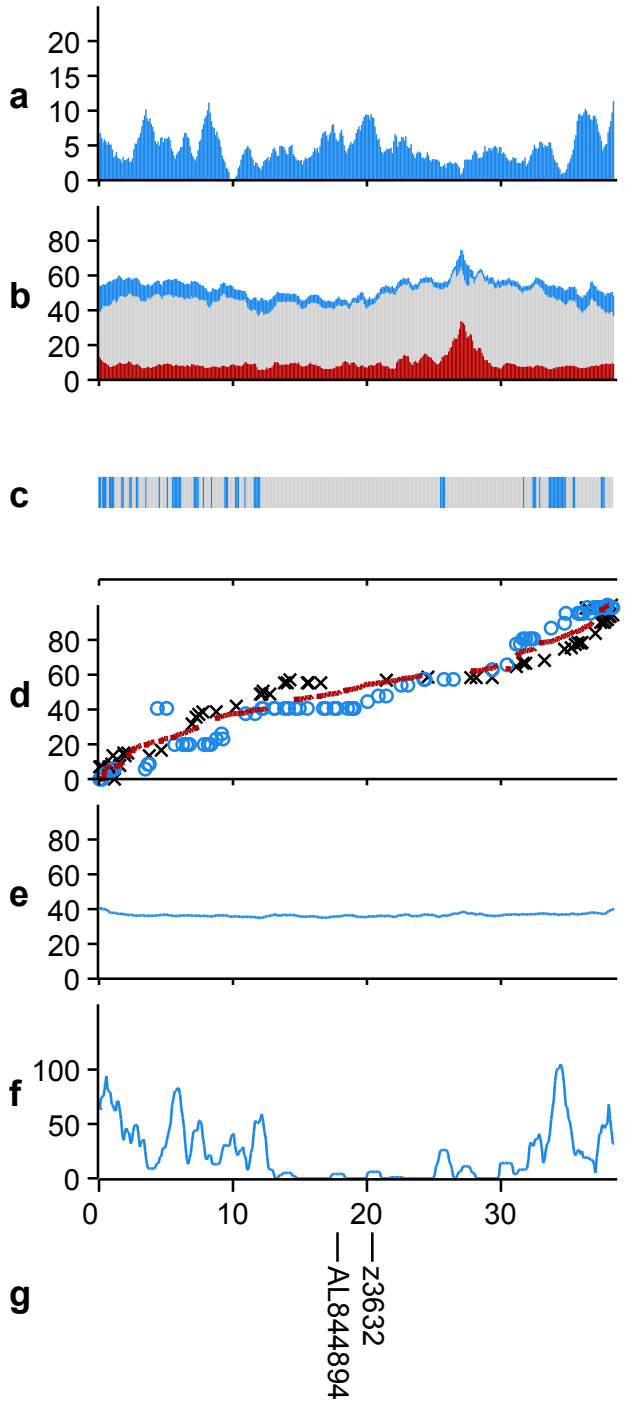
Supplementary Figure A22 | Landscape of Chromosome 22.
a, Percent coverage by exons of protein coding genes. **b**, Stacked repeat coverage, divided into Type I TEs (red), Type II TEs (grey) and other repeat types (blue) including dust, tandem and satellite repeats. **c**, Sequence composition (grey bars = clones, blue bars = whole genome shotgun contigs). **d**, Genetic marker placements (red = SATmap markers, blue = Heat Shock (HS) meiotic map markers, black = MGH meiotic map markers). Marker placements have been normalised to make the maps comparable. **e**, GC content. **f**, Gap density. **g**, Reported near centromeric markers (see Supplementary Information). The x-axis shows the chromosomal position in Mb. **a** and **b** were calculated as percentage coverage over 1Mb overlapping windows, with a 100kb shift between each window. **c**, **d**, **e** and **f** were calculated over 100 kb windows.



Supplementary Figure A23 | Landscape of Chromosome 23.
a, Percent coverage by exons of protein coding genes. **b**, Stacked repeat coverage, divided into Type I TEs (red), Type II TEs (grey) and other repeat types (blue) including dust, tandem and satellite repeats. **c**, Sequence composition (grey bars = clones, blue bars = whole genome shotgun contigs). **d**, Genetic marker placements (red = SATmap markers, blue = Heat Shock (HS) meiotic map markers, black = MGH meiotic map markers). Marker placements have been normalised to make the maps comparable. **e**, GC content. **f**, Gap density. **g**, Reported near centromeric markers (see Supplementary Information). The x-axis shows the chromosomal position in Mb. **a** and **b** were calculated as percentage coverage over 1Mb overlapping windows, with a 100kb shift between each window. **c**, **d**, **e** and **f** were calculated over 100 kb windows.



Supplementary Figure A24 | Landscape of Chromosome 24.
a, Percent coverage by exons of protein coding genes. **b**, Stacked repeat coverage, divided into Type I TEs (red), Type II TEs (grey) and other repeat types (blue) including dust, tandem and satellite repeats. **c**, Sequence composition (grey bars = clones, blue bars = whole genome shotgun contigs). **d**, Genetic marker placements (red = SATmap markers, blue = Heat Shock (HS) meiotic map markers, black = MGH meiotic map markers). Marker placements have been normalised to make the maps comparable. **e**, GC content. **f**, Gap density. **g**, Reported near centromeric markers (see Supplementary Information). The x-axis shows the chromosomal position in Mb. **a** and **b** were calculated as percentage coverage over 1Mb overlapping windows, with a 100kb shift between each window. **c**, **d**, **e** and **f** were calculated over 100 kb windows.



Supplementary Figure A25 | Landscape of Chromosome 25.
a, Percent coverage by exons of protein coding genes. **b**, Stacked repeat coverage, divided into Type I TEs (red), Type II TEs (grey) and other repeat types (blue) including dust, tandem and satellite repeats. **c**, Sequence composition (grey bars = clones, blue bars = whole genome shotgun contigs). **d**, Genetic marker placements (red = SATmap markers, blue = Heat Shock (HS) meiotic map markers, black = MGH meiotic map markers). Marker placements have been normalised to make the maps comparable. **e**, GC content. **f**, Gap density. **g**, Reported near centromeric markers (see Supplementary Information). The x-axis shows the chromosomal position in Mb. **a** and **b** were calculated as percentage coverage over 1Mb overlapping windows, with a 100kb shift between each window. **c**, **d**, **e** and **f** were calculated over 100 kb windows.