nature ecology & evolution

ARTICLES

https://doi.org/10.1038/s41559-020-1156-z

Deeply conserved synteny resolves early events in vertebrate evolution

Oleg Simakov[1,2,13] ✉, Ferdinand Marlétaz[1,11,13], Jia-Xing Yue[3,12], Brendan O'Connell[4], Jerry Jenkins[5], Alexander Brandt[6], Robert Calef[7], Che-Huang Tung[8], Tzu-Kai Huang[8], Jeremy Schmutz[5], Nori Satoh[9], Jr-Kai Yu[8], Nicholas H. Putnam[7], Richard E. Green[4] and Daniel S. Rokhsar[1,6,10] ✉

Although it is widely believed that early vertebrate evolution was shaped by ancient whole-genome duplications, the number, timing and mechanism of these events remain elusive. Here, we infer the history of vertebrates through genomic comparisons with a new chromosome-scale sequence of the invertebrate chordate amphioxus. We show how the karyotypes of amphioxus and diverse vertebrates are derived from 17 ancestral chordate linkage groups (and 19 ancestral bilaterian groups) by fusion, rearrangement and duplication. We resolve two distinct ancient duplications based on patterns of chromosomal conserved synteny. All extant vertebrates share the first duplication, which occurred in the mid/late Cambrian by autotetraploidization (that is, direct genome doubling). In contrast, the second duplication is found only in jawed vertebrates and occurred in the mid–late Ordovician by allotetraploidization (that is, genome duplication following interspecific hybridization) from two now-extinct progenitors. This complex genomic history parallels the diversification of vertebrate lineages in the fossil record.

In the 1970s, Ohno[1] proposed that vertebrates arose through a process involving one or more genome-wide duplications. This hypothesis received early support from the discovery of multiple vertebrate Hox clusters compared with one invertebrate cluster[2] and the finding of numerous vertebrate gene families with members distributed across multiple chromosomes[3,4]. Further evidence came from the discovery of paralogous (that is, duplicated) blocks of linked genes on multiple chromosomes within the human genome[5–8], culminating in the discovery of widespread quadruply conserved synteny of the human genome[9,10]. These studies support the so-called '2R' scenario of two rounds of whole-genome duplication during vertebrate evolution.

However, the number, timing and mechanism of these duplication events are still debated[3,10–14]. Alternatives to the 2R hypothesis include the recent proposal of a single whole-genome duplication with "additional large paralogy regions being the product of rare segmental duplications occurring both before and after", based on comparative analyses of the sea lamprey genome[13,15]. Others have suggested a series of large segmental duplications without any genome-wide events[16,17], although this is a minority view. Contributing to this uncertainty are discrepancies in the inferred chromosomal organization of the proto-vertebrate ancestor. By analysing gene linkages within and among selected bony vertebrate genomes (Euteleostomi), some authors have suggested the existence of 10–13 proto-vertebrate (that is, before any duplications) chromosomes[13,15,18–20], although other studies[10,14,21] have inferred 17 ancestral chromosomes.

Results and discussion

Amphioxus chromosomes reflect ancestral chordate linkages. As an invertebrate chordate whose lineage diverged before the emergence of vertebrates, amphioxus species have often served as a proxy for the ancestral proto-vertebrate condition[22], and provide a critical outgroup for analysing vertebrate-specific gene duplications[2–4,10] and the evolution of vertebrate gene regulation[23]. To robustly infer the proto-vertebrate karyotype and the genomic changes that accompanied the invertebrate-to-vertebrate transition, we produced a chromosome-scale genome assembly of amphioxus (the Florida lancelet *Branchiostoma floridae*). We combined existing shotgun data[10] with new in vitro[24] and in vivo[25] chromatin conformation capture sequences that enable megabase-scale scaffolds to be accurately linked together to reconstruct chromosomes[24,25] (Methods, Supplementary Notes 1 and 2 and Extended Data Fig. 1a). The resulting chromosome-scale assembly of *B. floridae* represents a substantial improvement over the original draft genome sequence, which achieved only megabase-scale scaffolds[10], and megabase-scale assemblies of other amphioxus species[23,26]. Our assembly assigns 94.5% of genes to the 19 *B. floridae* chromosomes BFL1–19. We validated the chromosome-scale accuracy of the new *B. floridae* assembly by generating a dense meiotic linkage map made from the F1 progeny of two wild parents (Supplementary Note 3 and Extended Data Fig. 1b)[10,22].

To examine the conservation of syntenic relationships, we constructed Oxford dot plots comparing the chromosomal positions of orthologous genes between genomes of amphioxus and multiple

[1]Molecular Genetics Unit, Okinawa Institute of Science and Technology Graduate University, Okinawa, Japan. [2]Department of Neuroscience and Developmental Biology, University of Vienna, Vienna, Austria. [3]Université Côte d'Azur, CNRS, INSERM, IRCAN, Nice, France. [4]Department of Biomolecular Engineering, University of California, Santa Cruz, Santa Cruz, CA, USA. [5]HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA. [6]Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA, USA. [7]Dovetail Genomics, Scotts Valley, CA, USA. [8]Institute of Cellular and Organismic Biology, Academia Sinica, Taipei, Taiwan. [9]Marine Genomics Unit, Okinawa Institute of Science and Technology Graduate University, Okinawa, Japan. [10]Chan Zuckerberg Biohub, San Francisco, CA, USA. [11]Present address: Centre for Life's Origins and Evolution, Department of Genetics, Evolution and Environment, University College London, London, UK. [12]Present address: State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University Cancer Center, Guangzhou, China. [13]These authors contributed equally: Oleg Simakov, Ferdinand Marlétaz. ✉e-mail: oleg.simakov@univie.ac.at; dsrokhsar@gmail.com

vertebrates (Fig. 1a and Extended Data Figs. 2 and 3) and invertebrates (Fig. 1b and Extended Data Fig. 4). These plots clearly display dense rectangular blocks of dots that represent units of deeply conserved synteny. Here, we use the original meaning of 'synteny'[27] to represent physical linkage without regard to gene order. The uniform distribution of orthologues within these blocks implies that while physical linkage is conserved, gene order within syntenic units has become largely scrambled since the amphioxus and other lineages diverged from each other more than half a billion years ago. This gene order scrambling is the result of accumulated inversions and other intra-chromosomal rearrangements over time, as observed between *Drosophila* species across increasing evolutionary distances[28,29].

Comparison of the chromosomes of amphioxus and the scallop *Patinopecten yessoensis* (chromosome code: PYE)[30]—the most complete chromosomal assembly of a marine invertebrate until amphioxus, albeit with only 80% of genes assigned to linkage groups—shows the remarkable stability of bilaterian chromosomes. Many amphioxus and scallop are in 1:1 correspondence, and others are related by the limited rearrangements described below (Fig. 1b). This observation directly confirms the deep conservation of synteny we previously hypothesized based on networks of conserved linkages observed among draft genomes of diverse invertebrates[21,30–33] (see Extended Data Fig. 4).

We identified 17 distinct patterns of conserved amphioxus–vertebrate synteny (Supplementary Note 4). Each pattern represents a group of genes whose linkage has been preserved since the divergence of the vertebrate and cephalochordate lineages, and is identified with an ancestral chordate linkage group (CLG). Each CLG is assigned a letter (A–Q, in decreasing order of number of genes) and, for ease of representation, colour, consistently used throughout. Vertical dashed lines show sharp boundaries between segments of different CLG ancestry in amphioxus; these are consistent in all comparisons among both vertebrates and invertebrates (Fig. 1 and Extended Data Figs. 2–4). The CLGs defined here by amphioxus–vertebrate comparisons also are found as intact units in the scallop (Fig. 1b and Supplementary Note 4), which implies that they represent even more ancient bilaterian or metazoan conserved syntenic units.

The amphioxus karyotype ($n = 19$) is derived from the 17 ancestral CLGs through a handful of large-scale rearrangements. Twelve amphioxus chromosomes (BFL1, 5–9, 11–15 and 17) are in 1:1 correspondence with CLGs and are therefore direct descendants of these ancestral units, albeit with extensive internal gene order rearrangement (Fig. 1). The remaining amphioxus chromosomes were formed through a small number of translocations between ancestral units. Three of the longer amphioxus chromosomes (BFL2, 3 and 4) are each derived from pairs of CLGs, with sharp boundaries between distinct patterns of conserved synteny across the chicken (chromosome code: GGA), gar (chromosome code: LOC), human (chromosome code: HSA), frog (chromosome code: XTR), sea lamprey and scallop (vertical dashed lines Fig. 1 and Extended Data Figs. 2 and 3). BFL2 exhibits an alternating block pattern of CLGJ

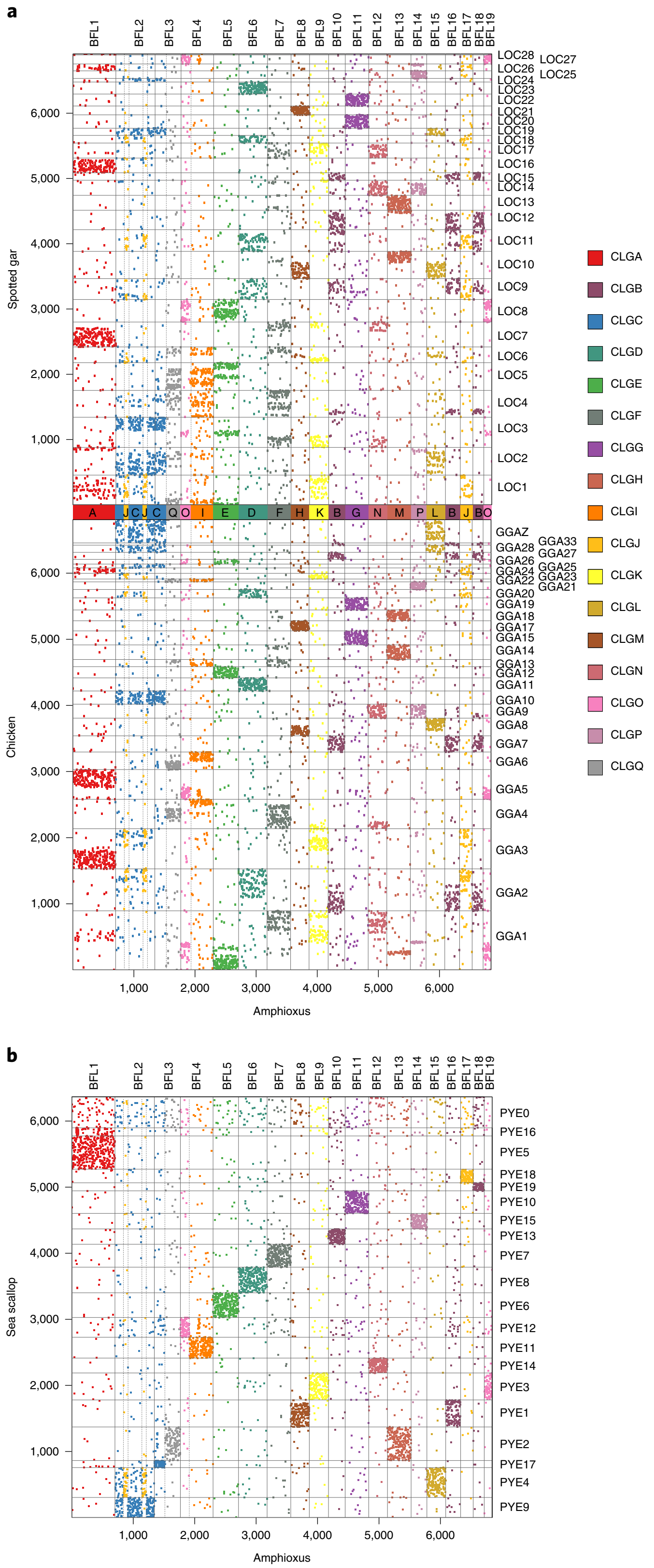


Fig. 1 | Conserved syntenies between amphioxus and various species. **a**, Oxford dot plot of orthologous genes between amphioxus and two representative bony vertebrates: spotted gar (*Lepisosteus oculatus*; top) and chicken (*Gallus gallus*; bottom). The axes show the index of 6,843 orthologous gene families anchored by mutual best hits from gar, chick, frog and human to amphioxus, with chromosome boundaries indicated. Dashed vertical lines show the location of synteny breakpoints for amphioxus that are consistent in comparisons with other vertebrate (Extended Data Figs. 2 and 3) and invertebrate genomes (see **b**; Extended Data Fig. 4). Genes are coloured according to this partitioning, defining 17 ancestral CLGs, with labels shown to the right. **b**, Mutual best-hit dot plot of amphioxus versus scallop, using the same colouring as in **a**. Syntenic discontinuities in amphioxus (indicated by the dashed lines) are consistent in the scallop. Note that CLGB (dark purple) is distributed across three pairs of homologous chromosomes, implying that this CLG existed as three distinct linkage groups in the scallop–amphioxus common ancestor.

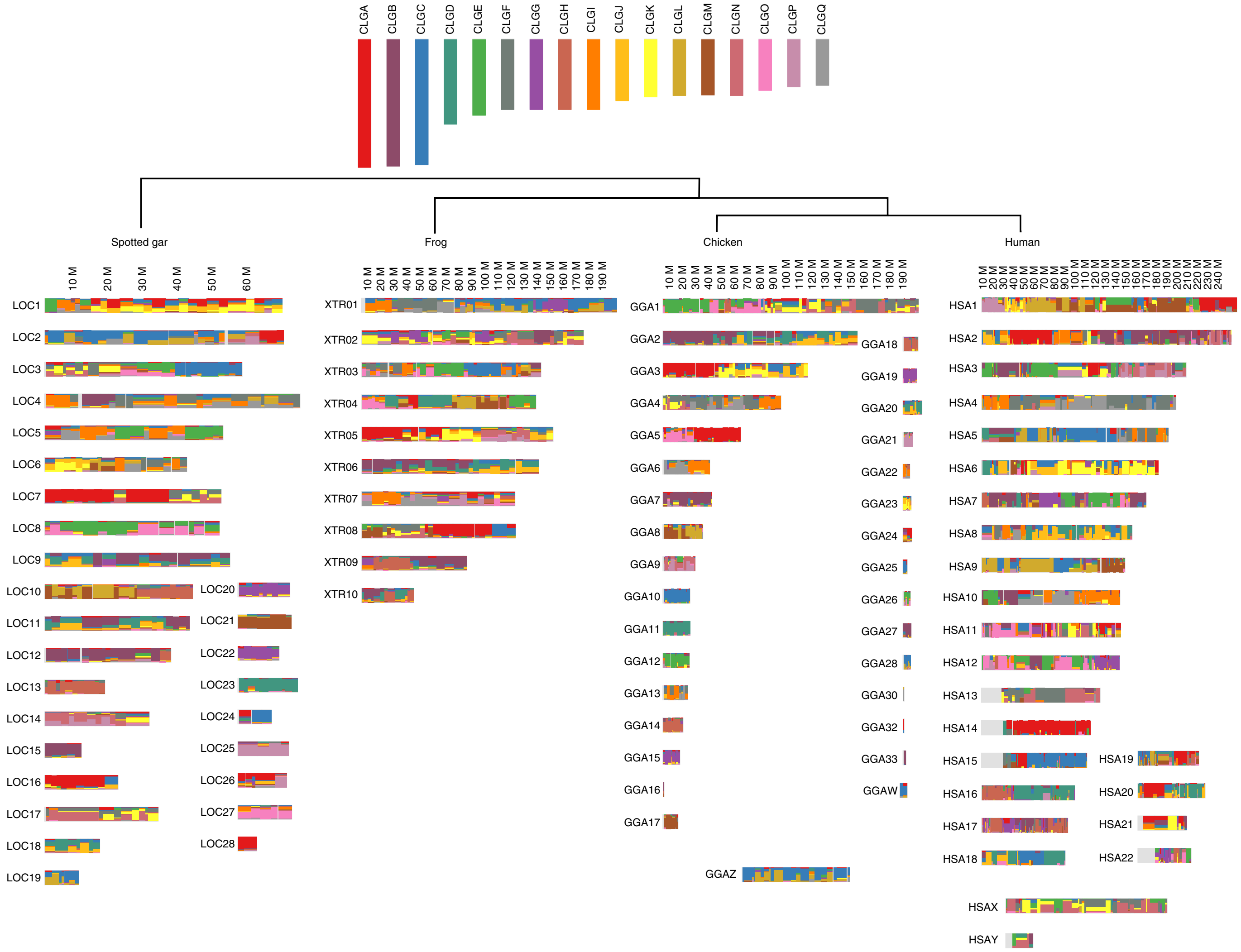


Fig. 2 | Contributions of the 17 ancestral CLGs to contemporary vertebrate genomes. The CLG ancestries of four jawed vertebrate genomes are shown by the local fraction of genes that are derived from each CLG, in windows of approximately 20 genes (see Methods). Note that, in contrast with Fig. 1, the chromosomal position is shown as physical coordinates (that is, base pairs), so area is not proportional to gene number. Colours are the same as in Fig. 1. The statistical significance of the associations between CLGs and vertebrate chromosomes is reported in Fig. 3.

and C ancestry that plausibly arose through a pair of overlapping inversions that occurred after a translocation involving a common ancestor with BFL17, which shares CLGJ ancestry with BFL2. Since the CLG boundaries remain sharp in amphioxus, these rearrangements must have occurred recently on the time scale of gene order scrambling. Alternately, sharp boundaries could represent current or historical centromeres, which interfere with mixing across arms.

The trio of amphioxus chromosomes BFL10, 16 and 18 together correspond to the ancestral CLGB. However, each of these chromosomes is associated with a different scallop chromosome (Fig. 1). It follows that these each represent a conserved ancestral bilaterian unit, implying at least 19 basic elements in the bilaterian ancestor. However, since orthologues of BFL10, 16 and 18 always occur mixed together in bony vertebrates and lamprey, we infer that these three elements fused before the origin of vertebrates, and for the purposes of our vertebrate-centric analysis treat them as a single CLG unit below. The stability of the amphioxus karyotype relative to these ancestral units is consistent with the megabase-scale conserved synteny (Supplementary Note 6) and minor karyotypic differences observed between *Branchiostoma* species[34]. The amphioxus genome shows no evidence of large-scale duplication.

Deep ancestry of vertebrate chromosomes. With the 17 ancestral chordate linkage units in hand, we can infer the sequence of genomic events that produced modern vertebrates. From Fig. 1, we can determine the distribution of CLG ancestry across the bony vertebrate genomes as shown in Fig. 2. Vertebrate chromosomes generally comprise one or more large blocks that are either: (1) descendants of single CLGs (dominated by a single colour; for example, the block of pure CLGA ancestry of the left arm of GGA3 and right arm of GGA5); or (2) mixtures of two or three CLGs formed by fusion and subsequent rearrangement (multiple overlapping or interleaved colours; for example, the CLGC and CLGL ancestry of GGAZ). Since chicken, spotted gar, and sea lamprey show fewer chromosomal rearrangements than human and frog, we focus on these genomes in the main text as more closely reflecting ancestral vertebrate genome organization.

We find that micro-chromosomes of the chicken, gar and lamprey (defined in ref. [35] as chromosomes shorter than 15 megabases) typically descend from single CLGs (Supplementary Note 6). Furthermore, micro-chromosomes of the chicken and spotted gar are often orthologous[34] (Fig. 3). For example, GGA28 and LOC19 both descend from CLGC and are orthologous (their 1:1 relationship is indicated by the double-headed arrow symbol in Fig. 3). These observations not only imply that such acrocentric micro-chromosomes were present in the common bony vertebrate ancestor[19,36], but also that they are relicts of even more ancient micro-chromosomes of the last common chordate ancestor, many of which are preserved in amphioxus. Remarkably, all CLGs appear at least once in unmixed form in the sea lamprey (see Extended Data Fig. 3), and nine out of 17 (CLGA, B, C, D, G, H, K, M and P) are also found in their ancestral unfused form in at least one bony vertebrate. This implies that at least one descendant of each original CLG has persisted (albeit with gene loss; see below) since the earliest periods of vertebrate evolution.

In contrast with micro-chromosomes, the longer metacentric macro-chromosomes of bony vertebrates are typically concatenations of segments with either distinct single CLG ancestry or blocks of mixed CLG ancestry (Fig. 2). Sharp boundaries between blocks of differing CLG ancestry represent either translocation boundaries or contemporary or ancient centromeres. The centromere scenario is consistent with the hypothesis that CLGs represent ancient chordate chromosome arms and implies that some metacentric vertebrate chromosomes arose through ancient Robertsonian fusions/translocations[37]. In this sense, the CLGs can be thought of as the vertebrate analogues of the Muller arms of *Drosophila*, which have maintained their integrity during fruit fly evolution despite considerable internal rearrangement[28,29].

Numerous instances of duplication, fusion and mixing can be seen in Figs. 1 and 2, and their pattern across genomes reveals the ancient dynamics of vertebrate chromosomes. Consider, for example, the distribution of genes with CLGE (green) and CLGO (pink) ancestry across bony vertebrate genomes. In Fig. 4a, chicken, gar, and frog chromosomes with E and O ancestry are arranged into five paralogous sets (Extended Data Fig. 5; see also Supplementary Note 6), revealing that these bony vertebrates generally have: (1) a pair of E-only segments (although one of the E-only segments is missing or dispersed and/or not detected in the chicken); (2) one O-only segment; and (3) two segments with mixed CLGE/CLGO ancestry. We interpret these blocks of mixed ancestry as arising from the past fusion of segments of E and O ancestry followed by a series of local rearrangements. The existence of pure E and O segments in the outgroups amphioxus and scallop indicates that pure CLG ancestry was the ancestral state and that these mixtures arose by fusion. This ancestral proto-vertebrate state is further corroborated by the absence of E–O fusions in the lamprey, which implies that the E–O fusion occurred on the bony vertebrate stem after divergence from the lamprey.

The cladogram on the left of Fig. 4a shows the most parsimonious derivation of these bony vertebrate segments from CLGE and CLGO chromosome ancestors. In particular, we note that it is more parsimonious for the two mixed-ancestry segments to arise by duplication after a common ancestral fusion/mixing than for two independent E–O fusion/mixing events to occur. Since all bony vertebrates possess the two fused segments, the duplication must have occurred before the tetrapod–gar divergence. Similarly, the co-existence of E-only, O-only and E–O fused segments implies duplication of E and O before the fusion/mixing event. The descendants of these early duplications are labelled 1 and 2, while the products of the second set of duplications are labelled α and β, in keeping with the notation detailed below. In this scenario, the β copy of O-1 has been lost (shown as a dashed pink rectangle).

Figure 3 extends this logic across vertebrate genomes to reveal a network of ancient duplications, fusions and mixing. Each cell in the table corresponds to the bony vertebrate descendent of an ancestral chordate unit, represented by a trio of orthologous chromosome segments of the chicken, spotted gar and frog (Methods). The cells are arranged according to CLG ancestry (rows), with paralogous copies in different columns. Solid lines enclose conserved linkages among CLGs (that is, juxtapositions or mixtures of CLG ancestry on orthologous vertebrate chromosomes) across bony vertebrates. In the vast majority of cases, segments on different chromosomes that descend from the same CLG arose by ancient duplication, as can be seen by the distribution of paralogous genes within jawed vertebrate genomes[9,10,19,20] (Extended Data Fig. 5). The alternative situation, in which CLG blocks are split across multiple chromosomes due to past translocations or fissions, is rarer but does occur[10,19], and can be inferred by parsimony when their orthologues are maintained as a single block in another jawed vertebrate genome, with amphioxus and scallop serving as outgroups. For example, the segments of LOC9 and 11 with CLGB, D and J ancestry are together orthologous to GGA2, and therefore probably arose by translocation after divergence of the gar and tetrapod lineages. Conversely, there are cases where two paralogous blocks with the same CLG ancestry are found on the same frog chromosome; consistent orthology with the gar and chicken allows these blocks to be identified and placed in different cells in Fig. 3.

Patterns of duplication and fusion in early vertebrate evolution. The hidden structure of vertebrate chromosomes revealed in Fig. 3 exhibits several remarkable patterns that: (1) imply two distinct tetraploidizations in the history of bony vertebrates; and (2) constrain the mechanisms and timing of those duplications. These observations lead us to propose a novel scenario for vertebrate palaeopolyploidy that is shown in Fig. 5. Since our inferences are derived from discrete patterns of conserved macro-synteny involving significant (see *P* values in Fig. 3 and Extended Data Figs. 6–8) conserved linkages of dozens to hundreds of genes, they are robust to phylogenetic artefacts of modelling sequence evolution.

First, the majority of CLGs (ten out of 17) are found in four descendent copies in bony vertebrate genomes (Fig. 3); the remainder are found in three copies. This pattern supports the 2R hypothesis if we allow for secondary chromosome loss via ancient aneuploidy of initially quadruply redundant copies. We note that gene loss, which can be extensive after genome duplication (see below), also reduces our power to detect statistically significant segments of conserved synteny, especially for the CLGs that contain fewer genes, so that some empty cells may be due to our inability to confidently detect them. Extensive gene loss after polyploidy also makes subsequent aneuploidy less disruptive and therefore more

Fig. 3 | Organization of bony vertebrate chromosomes after 2R. The majority of CLGs have four copies in bony vertebrates; the remainder have three. Organizing these copies by chromosome fusion (solid rectangles joining cells) and gene retention (numbers in cells) shows that chicken, spotted gar and frog chromosomes can be sorted into 'α–β' pairs that share the same patterns of CLG fusion, and these pairs themselves form '1–2' pairs. Bold dashed lines separating CLGA-2α and CLGB-1α from their fusions with other CLGs indicate either fusions in the α-lineage or fissions in β. Due to this ambiguity, the β pairings in these two rows are arbitrary. Similarly, the β copies for CLGG and CLGH are arbitrarily assigned to 2. In several cases (for example, CLGO) two distinct copies are found on the same chromosome of one species; these are indicated as a and b. Arrows imply that the entire source chromosome is orthologous to the target; double-headed arrows indicate reciprocal orthology; boxes indicate that segments of the chromosomes are orthologous; -- indicates undetected enrichment. The significance of associations between CLG and jawed vertebrate chromosomes was determined as described in Methods. Significance determined using 50-gene windows ($P < 0.01$) is indicated by an asterisk. Significance determined using 50-gene windows ($P < 0.05$), 100-gene windows ($P < 0.01$) and/or at the whole-chromosome level ($P < 0.05$) is determined by a plus sign. All *P* values were Bonferroni corrected.

plausible. Figure 3 is consistent with previous studies[9,10,14,19,20] that find approximately fourfold jawed vertebrate paralogy, as expected for a 2R scenario. Our findings are notably more extensive than the relatively limited jawed vertebrate paralogies recently reported in refs. [13,15], which led these authors to propose only a single whole-genome duplication during vertebrate evolution.

Chordate linkage group	1α	1β	2α	2β
CLGC	GGA10* ⇒ LOC3* XTR3* 0.301 : 0.329 0.337	GGA25* ⇔ LOC24* XTR8* 0.079 : 0.084 0.076	GGAZ* □ LOC2*/4* XTR1*b 0.344 : 0.328/0.073 0.354	GGA28* ⇔ LOC19* XTR1*a 0.142 : 0.144 0.183
CLGL	GGA8* ⇒ LOC10* XTR4* 0.431 : 0.419 0.425	GGA-- –LOC6* XTR3* -- : 0.170 0.150	GGAZ* □ LOC2*/4 XTR1*b 0.363 : 0.313/0.079 0.308	GGA28* ⇔ LOC19* XTR1*a 0.161 : 0.170 0.217
CLGM	GGA8* ⇒ LOC10* XTR4* 0.376 : 0.376 0.379	GGA-- –LOC6 XTR3 -- : 0.090 0.098	GGA17*⇔LOC21* XTR8* 0.506 : 0.489 0.409	--
CLGE	GGA12* ⇒ LOC5* XTR4* 0.370 : 0.367 0.313	GGA-- –LOC1* XTR8* -- : 0.156 0.128	GGA1* ⇐ LOC8* XTR3* 0.492 : 0.417 0.422	GGA26* ⇒ LOC3* XTR2+ 0.136 : 0.164 0.176
CLGO	GGA5* ⇒ LOC27* XTR4*/7* 0.450 : 0.384 0.300/(0.054)	--	GGA1* ⇐ LOC8* XTR3*/7* 0.463 : 0.452 0.305/(0.050)	GGA26+ ⇒ LOC3* XTR2+ 0.151 : 0.183 0.135
CLGI	GGA6* ⇒ LOC5* XTR7* 0.360 : 0.366 0.320	GGA22* ⇒ LOC1* XTR3b* 0.113 : 0.209 0.210	GGA4* □ LOC4*/2 XTR1* 0.260 : 0.283/0.089 0.320	GGA13* ⇒ LOC6* XTR3a* 0.148 : 0.142 0.190
CLGQ	GGA6* ⇒ LOC5* XTR7* 0.456 : 0.446 0.426	GGA22* ⇒ LOC1* XTR3b+ 0.078 : 0.169 0.159	GGA4* □ LOC4*/2 XTR1* 0.400 : 0.262/0.179 0.395	GGA13+ ⇒ LOC6+ XTR3a+ 0.156 : 0.190 0.132
CLGF	GGA1* □ LOC3*/17* XTR2* 0.401 : 0.179/0.154 0.328	GGA4b* □ LOC7* XTR8* 0.160 : 0.182 0.188	GGA4a* □ LOC4* XTR1* 0.394 : 0.373 0.377	GGA13* ⇒ LOC6* XTR3* 0.147 : 0.157 0.167
CLGK	GGA1* □ LOC3*/17* XTR2ac* 0.394 : 0.170/0.178 0.338	GGA4+ □ LOC7- XTR8* 0.137 : 0.117 0.135	See below	See below
CLGN	GGA1* □ LOC3*/17* XTR2* 0.494 : 0.142/0.287 0.451	GGA4* □ LOC7* XTR8* 0.247 : 0.232 0.242	GGA9* □ LOC14* XTR5* 0.357 : 0.366 0.373	--
CLGP	GGA21* ⇔ LOC25* XTR7* 0.437 : 0.444 0.451	GGA1* □ LOC26* XTR-- 0.141 : 0.094 --	GGA9* □ LOC14* XTR5* 0.315 : 0.323 0.369	
CLGA	GGA5*=LOC7*/2* XTR8* 0.400 : 0.369/0.117 0.435	GGA1 □ LOC3 XTR2 0.072 : 0.037 0.052	GGA3* □ LOC16*/1* XTR5* 0.446 : 0.276/0.161 0.388	GGA24* ⇒ LOC26* XTR7* 0.058 : 0.058 0.076
CLGK	See above	See above	GGA3* □ LOC1* XTR5* 0.446 : 0.372 0.406	GGA23* ⇒ LOC6* XTR2b* 0.161 : 0.215 0.178
CLGJ	GGA2* ⇐ LOC9*/11* XTR6* 0.370 : 0.177/0.177 0.338	GGA20* ⇔ LOC18* XTR10 0.174 : 0.184 0.104	GGA3* □ LOC1*/16 XTR5* 0.362 : 0.311/0.049 0.252	GGA23* ⇒ LOC6+ XTR2* 0.159 : 0.194 0.209
CLGD	GGA2* ⇐ LOC9*/11* XTR6* 0.389 : 0.273/0.173 0.372	GGA20* ⇔ LOC18* XTR10* 0.196 : 0.204 0.148	GGA11* ⇔ LOC23* XTR4* 0.437 : 0.421 0.427	--
CLGB	GGA2* ⇐ LOC9*/11* XTR6* 0.427 : 0.431* 0.353	GGA33* ⇒ LOC4* XTR2* 0.090 : 0.162 0.199	GGA7* □ LOC12* XTR9* 0.344 : 0.383 0.307	GGA27* ⇔ LOC15* XTR10* 0.196 : 0.216 0.189
CLGG	GGA15* ⇔ LOC20* XTR1* 0.441 : 0.443 0.480	--	GGA19* ⇔ LOC22* XTR2* 0.402 : 0.410 0.411	GGA5 □ LOC9 XTR9* 0.066 : 0.075 0.055
CLGH	GGA14* ⇔ LOC13* XTR9* 0.512 : 0.532 0.465	--	GGA18* ⇒ LOC10* XTR10* 0.386 : 0.413 0.298	GGA1* □ LOC12* XTR4* 0.181 : 0.096 0.167

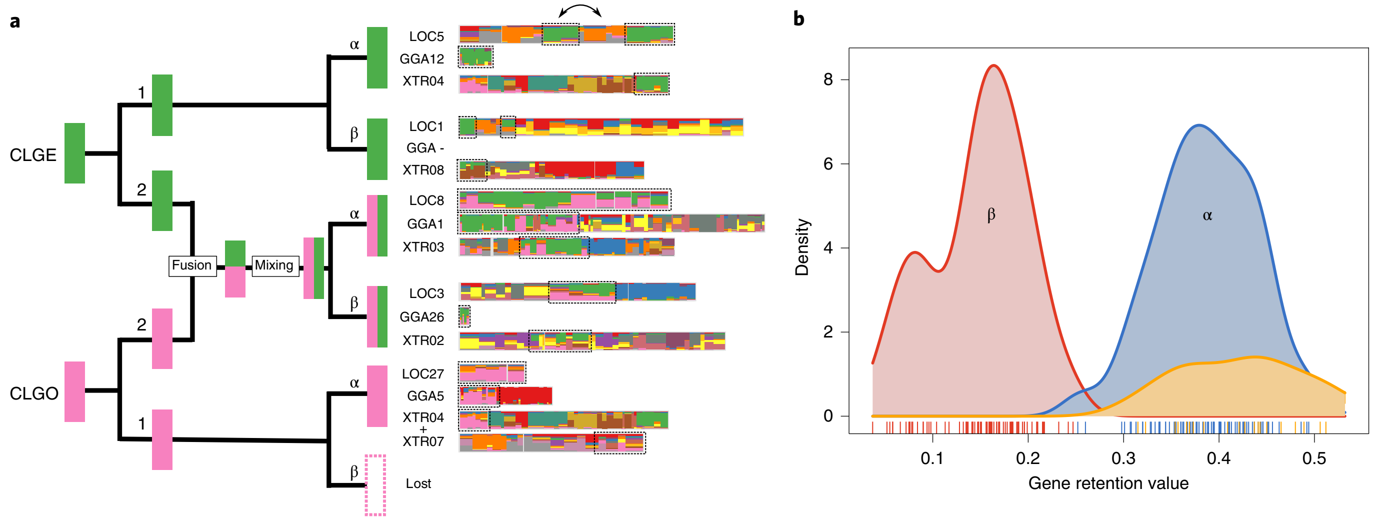


Fig. 4 | Duplications, fusions and mixing in bony vertebrates. a, Right: chromosomal descendants of CLGE (green) and CLGO (pink) are organized into five groups. Each chromosome is represented as in Fig. 2, with corresponding segments outlined by black dotted rectangles. The double-headed arrow indicates probable inversion that separated two CLG blocks. Within each group, segments with the CLGE and/or CLGO ancestry are orthologous among the chicken, gar and frog, and groups are paralogous to each other. Note that the frog chromosome XTR4 has distinct CLGE and CLGO segments with distinct ancestry (see Supplementary Note 6). Left: cladogram showing the most parsimonious evolutionary history leading to these vertebrate chromosomes, starting from CLGE and CLGO ancestors. This includes an early duplication (producing copies labelled 1 and 2), a fusion and subsequent mixing, and then a second duplication (producing copies labelled α and β). The CLGO-1β copy was not found, as indicated by a dashed pink rectangle. CLGE-1β was not found in chicken, as indicated by the dash. CLGO-1α was found split across XTR04 and XTR07, as indicated by the plus sign. **b**, Distribution of gene retention for the α and β segments listed in Fig. 3, with rug plot and kernel density estimator. The upper curves are for α–β pairs, whereas the orange curve is for α segments without β counterparts (presumed lost or possessing limited gene content and therefore undetected).

Second, Fig. 3 shows that bony vertebrate chromosomes with shared combinations of CLG ancestry generally appear in paralogous α–β pairs. Each α–β pair exhibits the same CLG linkages in the spotted gar, chick and frog, as shown by solid lines surrounding adjacent α and β cells in Fig. 3. (The two exceptions—A2/KJ2 and B1/DJ1—are shown with dashed lines in Fig. 3. These differences between α and β are plausibly accounted for by fusion in the α lineage after divergence from β.) These linked CLG combinations reflect ancient fusions or translocations[37], followed by more or less extensive mixing. By parsimony, we infer that the CLG fusions shared by paralogous α and β copies predate the genome-wide duplication that produced these α–β pairs, as shown in Fig. 4a. The α–β duplication also evidently preceded the divergence of bony vertebrate lineages, since α and β copies are found in both spotted gar and tetrapods. We reject the alternative scenario in which β copies independently fused in the same pattern as α copies ($P<10^{-10}$; Methods).

Third, each CLG generally participates in two α–β pairs of bony vertebrate segments (which we arbitrarily labelled 1 and 2 in Fig. 3). The genomic duplication that produced the 1–2 pairs therefore must have preceded the α–β duplication. We also infer from Fig. 3 that most CLG fusions in bony vertebrates occurred between the 1–2 and α–β duplications. CLGG and CLGH have evidently not fused; CLGI and CLGQ were either anciently joined before the 1–2 duplication, perhaps as arms of a single metacentric chromosome, or fused independently after it. This behaviour extends the scenario shown in Fig. 4a.

Finally, we can place the divergence of lamprey and bony vertebrate lineages relative to these duplications by noting that almost all of the CLG fusions observed in bony vertebrates are absent in the sea lamprey; indeed, most lamprey chromosomes are directly descended from single CLGs (Extended Data Figs. 3 and 8). This observation further corroborates the ancestral proto-vertebrate nature of these 17 syntenic units, and implies that most of the ancient bony vertebrate-specific fusions shown in Fig. 3 occurred after the divergence of the lamprey lineage. The exception is the C–L fusion found in a single lamprey chromosome (scaf_0006), which has a paralogous α–β pair in bony vertebrates. This observation suggests that either the C–L fusion occurred before the divergence of the lamprey and jawed vertebrate lineages, or convergent C–L fusions occurred in the two lineages. However, since unfused copies of CLGC and CLGL also exist in the lamprey, we infer that the 1–2 duplication occurred before the split between lamprey and jawed vertebrates. This duplication shared by lamprey and jawed vertebrates corresponds to the single event identified by Smith et al.[13]. The α–β duplication, however, occurred on the jawed vertebrate lineage after it split from lamprey. The fusions that intervene between the 1–2 and α–β events imply that the two duplications were temporally distinct (Fig. 5).

Asymmetrical paralogue retention and the mechanism of vertebrate genome duplication. Remarkably, we can infer the mechanisms of the temporally distinct 1–2 and α–β duplications by examining the chromosomal distribution of vertebrate orthologues relative to the unduplicated amphioxus genome. In the aftermath of a genome duplication, most duplicated genes are rapidly lost, so that paralogous segments retain a subset of the original gene complement[38–40]. However, the relative uniformity of gene loss depends on the underlying mechanism of genome duplication[41,42]. Following autotetraploidization (genome doubling within a species), we expect that gene losses should be evenly distributed across (initially homologous) duplicated chromosomes—a symmetry enforced by persistent tetrasomic inheritance immediately following autotetraploidy[43–45]. In contrast, allotetraploidization (genome doubling accompanying interspecific hybridization) is an inherently asymmetrical process that brings together the genomes

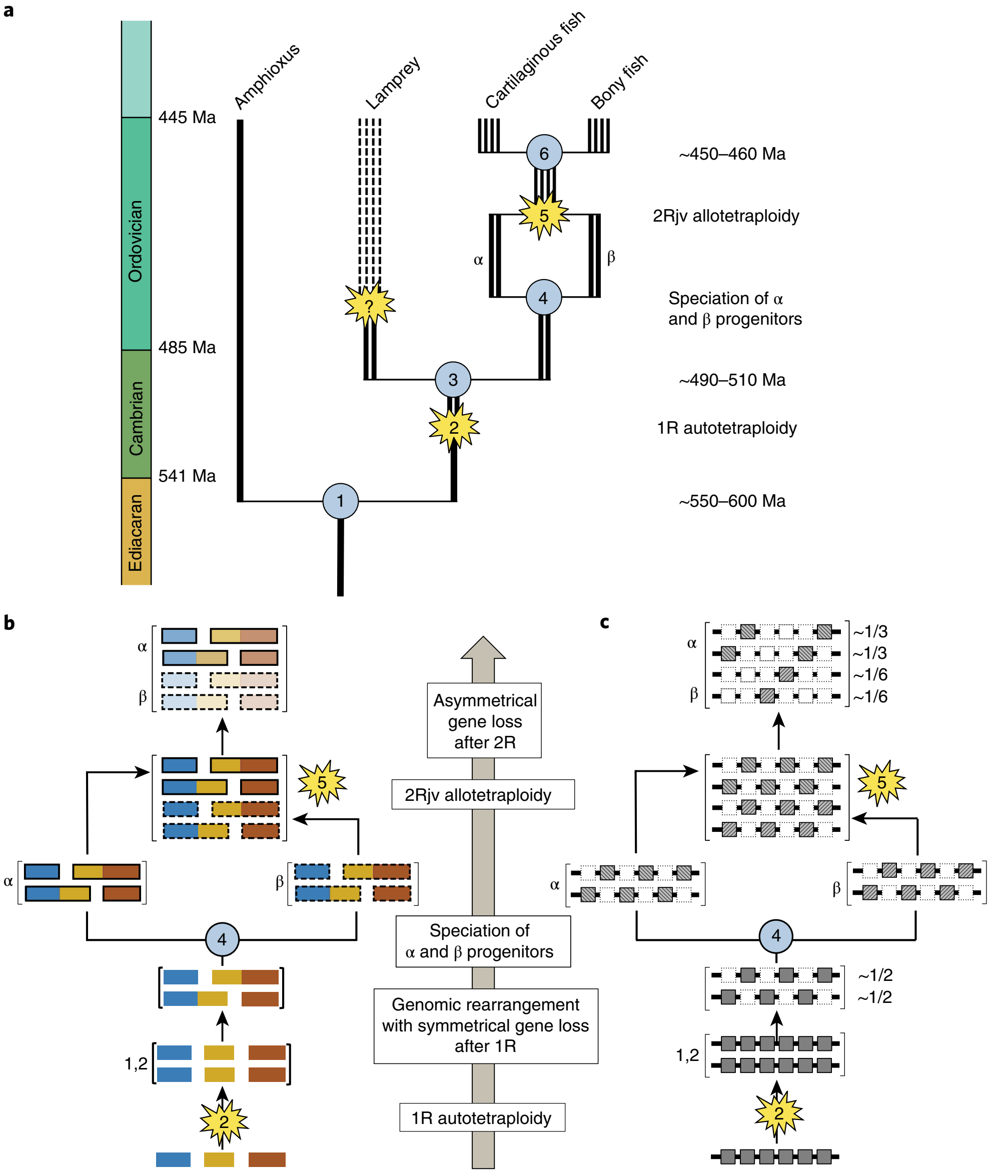


Fig. 5 | Auto- then allotetraploidy scenario for vertebrate evolution. Schematic of the auto- then allotetraploidy scenario described in the main text. **a**, Each line represents a chromosomal lineage. Single lines represent diploids, paired lines represent tetraploids, and so on, relative to the ancestral chordate chromosome complement. Dashed lines later in the lamprey lineage reflect one or more additional genomic duplications. Labelled nodes: (1) divergence of amphioxus and vertebrates (last common chordate ancestor); (2) 1R autotetraploidy, resulting in genome doubling; (3) divergence of (tetraploid) lamprey and gnathostome progenitor lineages; (4) speciation of palaeotetraploid gnathostome progenitors; (5) 2Rjv allotetraploidy, in which palaeotetraploid gnathostome progenitors hybridize to form the crown gnathostome lineage, which is quadrupled relative to the chordate ancestor; (6) divergence of extant jawed vertebrate lineages. The question mark indicates one or more additional duplication(s) that may have occurred in the lamprey lineage. **b**, Schematic showing the evolution of three ancestral CLGs. Relevant nodes are labelled as in **a**. Bold and dashed boundaries around chromosomes in the α and β lineages, respectively, represent divergences that accumulate in each lineage. Differential shading after node 5 indicates subsequent gene loss. **c**, Schematic of the evolutionary history of six linked chordate genes through vertebrate duplications. Gene loss is symmetrical after autotetraploidy (node 2) but asymmetrical after allotetraploidy (node 5). For simplicity in this diagram, gene order changes are not shown. Cross-hatching indicates independent differentiation in α and β lineages. Empty dashed boxes follow the fate of lost genes.

of two progenitors with distinct epigenetic landscapes, cytonuclear interactions and histories of transposable element activity. Allotetrapoids are therefore expected to show an asymmetrical distribution of gene losses, as observed in palaeo-allotetraploid plants[42] and frogs[41].

Each cell of Fig. 3 reports the retention fraction in the chicken, spotted gar and frog relative to the corresponding linkage group of the (unduplicated) chordate ancestor (using amphioxus segments to represent ancestral content; see Methods). Under a model in which (1) all chromosomal copies are equivalent (as expected, for example, for two successive autotetraploidies) and (2) redundancies are completely eliminated by gene loss, we would expect ~25% retention per segment (note that this simple picture neglects the relatively small fraction of genes retained in multiple copies[10], including well-known genes such as those in the Hox, Wnt and Fox families; see Methods). Instead, we observe a strikingly asymmetrical pattern in which genes are more than twice as likely to be retained on α segments than on paralogous β segments (Fig. 4b, Methods and

Supplementary Note 6). This asymmetry implies that the α–β duplication occurred through allotetraploidy. In contrast, no asymmetry is observed between 1–2 duplicates (combining the corresponding α–β descendants), which suggests that this earlier duplication occurred by autotetraploidy. Our full scenario is shown in Fig. 5.

The distribution of gene retention fractions across spotted gar, chicken and frog (Fig. 4b) is clearly bimodal (unimodality rejected by Hartigan's dip test; $D = 0.076$; $P = 2 \times 10^{-6}$). High-retention α copies retain >25% of ancestral CLG content relative to amphioxus (mean 38.9%; s.d. 4.8% across gar, chicken, and frog) while paralogous low-retention β copies retain <25% (mean 15.1%; s.d. 5.3%). The mean difference between the retention fractions of α–β pairs is significantly higher than a null model in which retention is uniformly random across pairs of chromosomes ($P = 1.4 \times 10^{-5}$; Supplementary Note 6). Notably, the assignment of vertebrate chromosome segments to their respective α and β columns in Fig. 3 based on high and low retention is consistent with the patterns of CLG fusion. Across orthologous segments, retention fractions are highly correlated between the spotted gar, chicken and frog (pairwise Pearson correlations: 94–96%), consistent with most gene losses occurring in the immediate aftermath of the α–β duplication before the divergence of bony vertebrates. In their analysis of vertebrate genome duplication, Smith et al.[15] often only detected the α signal (Supplementary Note 6), accounting for the predominance of two rather than four paralogous copies in their study.

The auto- then allotetraploidization model of vertebrate genome evolution. Our auto- then allotetraploidization model (Fig. 5) differs from previous proposals in both the timing and modes of genomic duplication during vertebrate evolution[3,11]. In our scenario, an initial 1R doubling occurred before the divergence of the lamprey and jawed vertebrate lineages via autotetraploidization, and was followed by a second duplication (here, called 2Rjv) in the jawed vertebrate lineage (that is, after the lamprey lineage had diverged) via allotetraploidization. Since Putnam et al.[10] and Venkatesh et al.[46,47] previously found extensive syntenic conservation between the genomes of the elephant shark and some bony vertebrates, we infer that the palaeo-allotetraploidy described here in bony vertebrates had already occurred before the last common gnathostome ancestor. A strong prediction of our model is thus that, when fully characterized, cartilaginous fish chromosomes will show the patterns of CLG fusion described in Fig. 3. Previous scenarios have suggested two rounds of allotetraploidization[3] or two rounds of autotetraploidization[11]. Several studies based on gene trees suggested that two duplications preceded the lamprey–jawed vertebrate split[12,14], although an earlier analysis could not resolve the position of the first duplication (1R) relative to this split[10].

The 1R event shared by all vertebrates is analogous to the autopolyploidizations described in salmonids, cyprinids (carps and their relatives) and sturgeons (reviewed in refs. [48,49]). One of the ensuing autotetraploid lineages gave rise to lampreys, which lack CLG fusions that are found across bony vertebrates (Extended Data Fig. 3). A second autotetraploid lineage, leading to the jawed vertebrates, experienced the series of chromosomal fusions described in Fig. 3. These rearrangements were probably associated with a period of genetic diploidization (that is, the transition from tetrasomic to disomic inheritance (the formation of consistent bivalents between specific homologous pairs))[38]. Two descendants of this second lineage later hybridized in an allotetraploidization event (2Rjv) to give the jawed vertebrate ancestor.

Although many vertebrate gene families do not follow a doubly bifurcating pattern, as expected in a simple 2R scenario[4,16], Furlong and Holland[11] have elegantly argued that this phenomenon could be explained by extensive homeologous recombination during two closely spaced autotetraploidies. This argument generalizes to our auto- then allotetraploidy model as long as the diploidization period following 1R extended into the α and β progenitors before 2Rjv. Such a long period of residual tetrasomy after 1R is plausible; in salmonid fish, polysomic inheritance (that is, ongoing homeologous recombination) has persisted for tens of millions of years[43,49]. Detailed analysis of Hox-bearing chromosomes by Lynch and Wagner[50] suggests that they experienced two chromosomal crossovers early in vertebrate evolution. Interestingly, the Hox gene clusters are found on CLGB, with HoxA and HoxD found on α segments and HoxB and C on β segments. However, unlike most other CLGs, we cannot uniquely pair specific α and β copies based on chromosomal linkage in the absence of a pattern of paired fusion involving CLGB. This suggests that the ancestral Hox-bearing chromosomes (as well as CLGA, G and H) could have experiences a prolonged period of homeologous interaction and exchange, consistent with Lynch and Wagner's finding.

Although here we have described 17 CLGs by comparing the chromosomes of amphioxus and vertebrates, confirming ref. [10], several previous studies based only on comparisons among vertebrate genomes have variously inferred the existence of 10–13 ancestral vertebrate chromosomes[13,15,18–20]. Comparisons among vertebrate genomes without reference to an outgroup such as amphioxus are likely to miss the fusions that we have documented here, since anciently fused regions appear as single conserved syntenic blocks in comparisons among jawed vertebrates. Neglect of the complex history of fusions and mixing after duplication, and the reduced power to detect significant conserved synteny among paralogous regions with extensive gene loss (that is, the β segments of Fig. 3), appear to account for the discrepancies between previously published characterizations of ancestral vertebrate proto-chromosomes based on comparisons among or within vertebrates and our analyses (see also Supplementary Note 6 for several specific case studies). It is also important to consider that different portions of a vertebrate chromosome can have different ancestry. Specifically when assessed in 50-gene windows rather than at the whole-chromosome scale, we find statistically significant CLG–chicken and CLG–lamprey associations that were not found in earlier whole-chromosome comparisons of the chicken and sea lamprey (Extended Data Figs. 6 and 7 and Supplementary Note 6). Remarkably, in contrast with refs. [13,15], we find that lamprey chromosomes can be grouped into 17 clusters that correspond to our 17 CLGs (Extended Data Fig. 8; compare with Fig. 5b of ref. [13]). Finally, the original proposal[10] of 17 CLGs has found renewed support from analysis of the reconstructed ancestral amniote karyotype[14]. However, we note that this study used the non-chromosomal draft assembly of the amphioxus genome[10] for its outgroup, and so made the same assumptions as ref. [10].

Conclusion

Our analyses imply a novel scenario for vertebrate evolution and the events that shaped vertebrate genomes (Fig. 5). First, we show that 17 ancient CLGs[10] are stable chromosomal units and that relicts of these units are readily detectable as either intact micro-chromosomes or large chromosomal segments in vertebrates, amphioxus, and even molluscs. Second, the jawed vertebrate lineage experienced two temporally and mechanistically distinct genome-wide duplications. The first—an autotetraploidization (1R)—preceded the divergence of lamprey and jawed vertebrate lineages[51] ~490 million years ago (Ma). On the jawed vertebrate stem lineage, 1R was followed by a series of chromosomal fusions that preceded a second genome-wide duplication. On the lamprey lineage, 1R was followed by fewer and largely distinct fusions. Other studies[52,53] have suggested that additional large-scale duplication events occurred in the lamprey lineage (as indicated by the node with a question mark in Fig. 5a), consistent with the one-to-many conserved synteny that we observe between amphioxus and lamprey (Extended Data Fig. 3), and the many-to-many relationship found between lamprey and chicken chromosomes[13,15]. Ongoing diploidization in both lineages after 1R, including gene loss[38] and homeologous recombination[11], is consistent with the complex orthology relationships observed between lamprey and gnathostome genes (for example, refs. [52,54]). Disentangling these additional duplications specific to either the lamprey or cyclostome lineages will require further study. Third, the observation of asymmetrical gene retention implies that the second whole-genome duplication in the jawed vertebrate lineage (2Rjv) was an allotetraploidization (whole-genome duplication after interspecific hybridization). This second duplication preceded[10] the divergence of bony and cartilaginous fish (~438–465 Ma)[55,56]. Our bounds on the timing of 1R and 2Rjv suggest that the extensive Ordovician diversification[51] of early jawless and armoured fish (443–485 Ma) probably occurred during the period of diploidization after 1R, which was marked by chromosomal fusions and rearrangements that allowed new regulatory linkages to be explored. Hybridization of two related 1R descendants accompanied by genome duplication then established the lineage that subsequently gave rise to all living jawed vertebrates.

Methods

Chromosome-scale genome assembly and annotation. To produce a chromosome-scale assembly of amphioxus, we (1) reassembled existing shotgun data and then (2) ordered and oriented the resulting assembly with in vitro and in vivo chromatin conformation capture data.

We used ARACHNE[57] to assemble the approximately tenfold redundant Sanger whole-genome shotgun sequence that was previously generated[10] from a single diploid Florida lancelet (*B. floridae*). As in ref. [10], the two highly divergent haplotypes were assembled apart. For the present assembly, we resolved these diploid redundancies with HaploMerger2 (ref. [58]) to produce a single reference haplotype for further analysis (Supplementary Note 1). In the present assembly, the total assembled contig length improved from 480–489 megabases and the L50 contig length doubled from 25.6–52 kb.

To achieve a chromosome-scale assembly of amphioxus, we obtained long-range linkages using in vitro (Chicago library[24,25]) and in vivo (Hi-C) chromatin conformation capture libraries, as described in ref. [25] and Supplementary Note 2. We used the HiRise pipeline (Dovetail Genomics) to scaffold the haplomerged amphioxus assembly with both types of chromatin conformation capture data (Supplementary Note 3). The resulting assembly of 19 chromosomal scaffolds (accounting for 94.5% of the assembled sequence) was validated by constructing a genetic map from light shotgun sequencing of 96 progeny from an F1 cross (Supplementary Note 4).

The protein-coding genes of amphioxus were annotated using the EVM pipeline[59], incorporating recent transcriptome data[60] for *B. floridae* (Supplementary Note 5). The process yielded 28,192 protein-coding loci containing 5,108 distinct Pfam domains—an increase of 6.5% relative to the 4,797 Pfam domains in the annotation of the earlier sub-chromosomal assembly[10].

Mutual best-hit orthology. To develop sets of high-confidence orthologous genes between amphioxus and selected vertebrate genomes, we performed mutual (that is, reciprocal) best BLASTp searches (Supplementary Note 6). Mutual best hits provides a high-confidence set of orthologues with minimal additional bioinformatics processing, and was used to define CLGs and identify syntenically orthologous units between amphioxus and vertebrate genomes.

Definition of 17 CLGs. We partitioned the amphioxus chromosomes into 17 CLGs using the data from Fig. 1 and Extended Data Fig. 2, as follows. First, we identified boundaries along the amphioxus chromosomes between blocks of distinct vertebrate conserved synteny. Consider only genes i in amphioxus with mutual best hits in the comparator species, and define $\mathbf{x}_a(i)$ as the synteny indicator vector of gene i that takes the value of 1 if the gene has its orthologue in chromosome a of the comparator, and 0 otherwise. To identify boundaries at which the conserved synteny changes, we computed the left and right windowed averages of the synteny indicator vector:

$$\mathbf{X}_a^{\mathrm{L}}(i) = \frac{1}{W}\sum\nolimits_{j=i-W+1,i} \mathbf{x}_a(j) \text{ and } \mathbf{X}_a^{\mathrm{R}}(i) = \frac{1}{W}\sum\nolimits_{j=i,i+W-1} \mathbf{x}_a(j)$$

The squared Euclidean norm of the difference $D(i, i+1) = \mathbf{X}_a^{\mathrm{R}}(i+1) + \mathbf{X}_a^{\mathrm{L}}(i)$ then measures the discontinuity of conserved synteny between genes i and $i+1$. Using as comparators the vertebrates chicken, gar or frog, or the invertebrate scallop, and the window size $W=25$, D shows spikes at discontinuous boundaries.

We identified local peaks in D as intra-chromosomal boundaries in amphioxus between distinct patterns of conserved synteny. Consistent with the patterns seen in Fig. 1 and Extended Data Figs. 2 and 3, no boundaries were detected for most amphioxus chromosomes, but we identified four such synteny breakpoints in BFL2 and one boundary each in BFL3 and BFL4. These boundaries are consistent among vertebrates and scallops and consensus positions are indicated by vertical dashed lines in Fig. 1 and Extended Data Figs. 2–4. In several cases, the synteny indicator vectors of the chromosomal segments defined by the above procedure were closely aligned with each other. In these cases, the amphioxus segments were combined into a single syntenic unit (see further discussion in Supplementary Note 6). The resulting 17 CLGs (defined by the amphioxus genes contained in the corresponding segments) agree with the 17 putative ancestral linkage groups defined by clustering megabase-scale scaffolds (Supplementary Table 7) based on statistically significant patterns of conserved synteny with humans[10].

Gene families. To allow for analysis of (unlinked) gene duplication in vertebrates relative to the chordate ancestor, we also constructed families of orthologous chordate genes through sequence-based clustering analyses, as described in ref. [10]. For the purposes of assessing the retention of gene duplicates after whole-genome duplication (see below), we counted only one gene family member per chromosome. With this counting, linked (for example, tandem) gene duplications produced by local processes do not lead to increased retention values, but unlinked duplications plausibly created through chromosome or genome-scale events are counted.

Distribution of CLG ancestry across vertebrate chromosomes. Visualization of the local CLG ancestry across vertebrate chromosomes (Fig. 2) was based on orthologous chordate gene families. The height of each coloured bar in Fig. 2 was determined by the fraction of genes with the corresponding CLG ancestry in a window of at least 20 genes. Since gene density varies across vertebrate chromosomes, windows can have different physical sizes on Fig. 2, which shows base-pair position along chromosomes. To subdivide chromosomes according to their CLG ancestry, we searched through each chromosome for the largest peak in D, as defined above. This search was iterated with the condition that additional breakpoints were at least 20 genes away from a previously determined breakpoint. This process produces a partitioning of each chromosome into windows of at least 20 genes with relatively homogeneous CLG ancestry. Note that multiple CLGs can contribute to the same vertebrate chromosomal region, since blocks of CLG ancestry can overlap due to fusion and subsequent mixing, as described in the main text.

Significance testing of syntenic associations between amphioxus and vertebrate genomes. We tested for significant associations between amphioxus and vertebrate genomes using a variation of the method described in ref. [10] and later applied by Smith and Keinath[13] in their comparison of lamprey linkage groups with jawed vertebrate chromosomes. The null hypothesis was that orthologous genes are randomly distributed across the two genomes, with a Bonferroni correction for the total number of pairwise tests. While Smith and Keinath[13] used gene families defined by collecting high-scoring hits among lamprey and selected vertebrates, we used the 6,843 mutual best hits between amphioxus and each of the four bony vertebrates: spotted gar, chicken, frog and human. Mutual best hits provide a conservative set of orthologues. For mutual best-hit orthologues, the equivalent null distribution for the distribution of orthologues shared between chromosomes (or between windows within chromosomes) is hypergeometric (Supplementary Note 6). The significance of associations between CLGs and vertebrate chromosomes, including a multiple test correction for the number of associations tested, is shown in Extended Data Fig. 6.

Based on Figs. 1 and 2 and Extended Data Fig. 3, we noted that, especially for vertebrate macro-chromosomes, orthologues from a given amphioxus chromosome or CLG are not uniformly distributed along the sequence, but rather appear to be concentrated in sub-chromosomal windows. To test whether sub-chromosomal windows show significant enrichments relative to the null model, we applied the same hypergeometric test to sliding windows of 50 genes (Supplementary Note 6) and accounted for the increased multiple testing with a Bonferroni correction based on the number of window tests. The resulting P values are shown in Extended Data Fig. 7. We note that this is a more sensitive test than the chromosome-scale test of Smith and Keinath, since associations that are not significant at the chromosome scale can be significant when tested with 50-gene windows. In Fig. 3, asterisks represent significant associations ($P<0.01$ after Bonferroni multiple test correction) between CLGs and 50-gene windows; whereas plus signs indicate significance using 50-gene windows ($P<0.05$), 100-gene windows ($P < 0.01$) and/or at the whole chromosome level ($P<0.05$). Rows of Fig. 3 are also strongly supported by paralogy within vertebrate genomes.

Definition of retention fraction. After genome duplications, unlinked duplicates may be lost. The retention fraction for a CLG on a vertebrate chromosome or chromosome segment is defined as the ratio of the number of CLG orthologues (that is, gene family members) found on that chromosome to the number of CLG-defining genes in amphioxus. Since a gene can be retained in multiple unlinked copies, the total retention of a CLG across all chromosomes of a vertebrate species can exceed one. As noted above, gene families are defined such that linked duplicates (arising by tandem duplication) are not counted towards the retention fraction; only unlinked duplicates are relevant for analysis of whole-genome duplications except in special circumstances of translocations that combine paralogous chromosomes from the same CLG on the same vertebrate chromosome. These translocations are identified by parsimony and comparison among vertebrate genomes, as described in the main text.

Significance testing of asymmetry between high and low retention. To assess the significance of the difference between paired high- and low-retention classes, we compared the high and low means across the spotted gar and chicken against a null model in which retention rates in cells of Fig. 3 were chosen at random (that is, with only one retention class). To capture the broad distribution of values, we used a uniform distribution ranging from 0–0.532 (twice the overall retention mean). We chose pairs from this distribution and assigned the larger value as α and the smaller value as β. The difference between group means was normally distributed (confirmed by one million bootstrap simulations), and the observed group mean difference (high − low) was 7.16 standard deviations from the null model mean (Supplementary Note 7). To provide a conservative estimate of significance, we excluded pairs for which no β segment was found in Fig. 3. These pairs were excluded since coding these missing β segments as having a retention rate of zero only increased the high–low group difference. Similarly, using a normal distribution of retention rates (with mean and variance computed from observations) produced a null distribution of high–low means that was even farther from observed differences, so our use of the uniform distribution was conservative.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The raw data and genome assembly are available from the National Center for Biotechnology Information BioProject under the accession code PRJNA412957. Processed data are available from https://bitbucket.org/viemet/public.

Code availability

The custom code used in this study is available from https://bitbucket.org/viemet/public.

Received: 20 December 2017; Accepted: 19 February 2020;
Published online: 20 April 2020

References

1. Ohno, S. *Evolution by Gene Duplication* (Springer, 1970).
2. Garcia-Fernández, J. & Holland, P. W. Archetypal organization of the amphioxus Hox gene cluster. *Nature* **370**, 563–566 (1994).
3. Spring, J. Vertebrate evolution by interspecific hybridisation—are we polyploid? *FEBS Lett.* **400**, 2–8 (1997).
4. Escriva, H., Holland, N. D., Gronemeyer, H., Laudet, V. & Holland, L. Z. The retinoic acid signaling pathway regulates anterior/posterior patterning in the nerve cord and pharynx of amphioxus, a chordate lacking neural crest. *Development* **129**, 2905–2916 (2002).
5. Pebusque, M.-J., Coulier, F., Birnbaum, D. & Pontarotti, P. Ancient large-scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution. *Mol. Biol. Evol.* **15**, 1145–1159 (1998).
6. Abi-Rached, L., Gilles, A., Shiina, T., Pontarotti, P. & Inoko, H. Evidence of en bloc duplication in vertebrate genomes. *Nat. Genet.* **31**, 100–105 (2002).
7. Lundin, L.-G., Larhammar, D. & Hallböök, F. Numerous groups of chromosomal regional paralogies strongly indicate two genome doublings at the root of the vertebrates. *J. Struct. Funct. Genomics* **3**, 53–63 (2003).
8. Hokamp, K., McLysaght, A. & Wolfe, K. H. The 2R hypothesis and the human genome sequence. *J. Struct. Funct. Genomics* **3**, 95–110 (2003).
9. Dehal, P. & Boore, J. L. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.* **3**, e314 (2005).
10. Putnam, N. H. et al. The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**, 1064–1071 (2008).
11. Furlong, R. F. & Holland, P. W. H. Were vertebrates octoploid? *Phil. Trans. R. Soc. Lond. B Biol. Sci.* **357**, 531–544 (2002).
12. Kuraku, S. & Meyer, A. The evolution and maintenance of Hox gene clusters in vertebrates and the teleost-specific genome duplication. *Int. J. Dev. Biol.* **53**, 765–773 (2009).
13. Smith, J. J. & Keinath, M. C. The sea lamprey meiotic map improves resolution of ancient vertebrate genome duplications. *Genome Res.* **25**, 1081–1090 (2015).
14. Sacerdot, C., Louis, A., Bon, C., Berthelot, C. & Roest Crollius, H. Chromosome evolution at the origin of the ancestral vertebrate genome. *Genome Biol.* **19**, 166 (2018).
15. Smith, J. J. et al. The sea lamprey germline genome provides insights into programmed genome rearrangement and vertebrate evolution. *Nat. Genet.* **50**, 270–277 (2018).
16. Friedman, R. & Hughes, A. L. Pattern and timing of gene duplication in animal genomes. *Genome Res.* **11**, 1842–1847 (2001).
17. Naz, R., Tahir, S. & Abbasi, A. A. An insight into the evolutionary history of human MHC paralogon. *Mol. Phylogenet. Evol.* **110**, 1–6 (2017).
18. Kohn, M. et al. Reconstruction of a 450-My-old ancestral vertebrate protokaryotype. *Trends Genet.* **22**, 203–210 (2006).
19. 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 (2007).
20. Murat, F., Van de Peer, Y. & Salse, J. Decoding plant and animal genome plasticity from differential paleo-evolutionary patterns and processes. *Genome Biol. Evol.* **4**, 917–928 (2012).
21. Simakov, O. et al. Hemichordate genomes and deuterostome origins. *Nature* **527**, 459–465 (2015).
22. Holland, L. Z., Laudet, V. & Schubert, M. The chordate amphioxus: an emerging model organism for developmental biology. *Cell. Mol. Life Sci.* **61**, 2290–2308 (2004).
23. Marlétaz, F. et al. Amphioxus functional genomics and the origins of vertebrate gene regulation. *Nature* **564**, 64–70 (2018).
24. Putnam, N. H. et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **26**, 342–350 (2016).
25. Burton, J. N. et al. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).
26. Huang, S. et al. Decelerated genome evolution in modern vertebrates revealed by analysis of multiple lancelet genomes. *Nat. Commun.* **5**, 5896 (2014).
27. Renwick, J. H. The mapping of human chromosomes. *Annu. Rev. Genet.* **5**, 81–120 (1971).
28. Sturtevant, A. H. & Novitski, E. The homologies of the chromosome elements in the genus *Drosophila*. *Genetics* **26**, 517–541 (1941).
29. Ranz, J. M. et al. Principles of genome evolution in the *Drosophila melanogaster* species group. *PLoS Biol.* **5**, e152 (2007).
30. Wang, S. et al. Scallop genome provides insights into evolution of bilaterian karyotype and development. *Nat. Ecol. Evol.* **1**, 120 (2017).
31. Putnam, N. H. et al. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* **317**, 86–94 (2007).
32. Simakov, O. et al. Insights into bilaterian evolution from three spiralian genomes. *Nature* **493**, 526–531 (2013).
33. Hall, M. R. et al. The crown-of-thorns starfish genome as a guide for biocontrol of this coral reef pest. *Nature* **544**, 231–234 (2017).
34. Wang, C., Zhang, S. & Chu, J. G-banding patterns of the chromosomes of amphioxus *Branchiostoma belcheri tsingtauense*. *Hereditas* **141**, 2–7 (2004).
35. Burt, D. W. Origin and evolution of avian microchromosomes. *Cytogenet. Genome Res.* **96**, 97–112 (2002).
36. Braasch, I. et al. The spotted gar genome illuminates vertebrate evolution and facilitates human–teleost comparisons. *Nat. Genet.* **48**, 427–437 (2016).
37. Schubert, I. & Lysak, M. A. Interpretation of karyotype evolution should consider chromosome structural constraints. *Trends Genet.* **27**, 207–216 (2011).
38. Wolfe, K. H. Yesterday's polyploids and the mystery of diploidization. *Nat. Rev. Genet.* **2**, 333–341 (2001).
39. 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 (2004).
40. Wolfe, K. H. & Shields, D. C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**, 708–713 (1997).
41. Session, A. M. et al. Genome evolution in the allotetraploid frog *Xenopus laevis*. *Nature* **538**, 336–343 (2016).
42. Garsmeur, O. et al. Two evolutionarily distinct classes of paleopolyploidy. *Mol. Biol. Evol.* **31**, 448–454 (2014).
43. Lien, S. et al. The Atlantic salmon genome provides insights into rediploidization. *Nature* **533**, 200–205 (2016).
44. Martin, K. J. & Holland, P. W. H. Enigmatic orthology relationships between Hox clusters of the African butterfly fish and other teleosts following ancient whole-genome duplication. *Mol. Biol. Evol.* **31**, 2592–2611 (2014).
45. Robertson, F. M. et al. Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification. *Genome Biol.* **18**, 111 (2017).
46. Venkatesh, B. et al. Elephant shark genome provides unique insights into gnathostome evolution. *Nature* **505**, 174–179 (2014).
47. Venkatesh, B. et al. Survey sequencing and comparative analysis of the elephant shark (*Callorhinchus milii*) genome. *PLoS Biol.* **5**, e101 (2007).
48. Braasch, I. & Postlethwait, J. H. in *Polyploidy and Genome Evolution* (eds Soltis, P. S. & Soltis, D. E.) 341–383 (Springer, 2012).
49. Allendorf, F. W. & Thorgaard, G. H. in *Evolutionary Genetics of Fishes* (ed. Turner, B.) 1–53 (Springer, 1984).
50. Lynch, V. J. & Wagner, G. P. Multiple chromosomal rearrangements structured the ancestral vertebrate Hox-bearing protochromosomes. *PLoS Genet.* **5**, e1000349 (2009).
51. Janvier, P. Facts and fancies about early fossil chordates and vertebrates. *Nature* **520**, 483–489 (2015).

52. Mehta, T. K. et al. Evidence for at least six Hox clusters in the Japanese lamprey (*Lethenteron japonicum*). *Proc. Natl Acad. Sci. USA* **110**, 16044–16049 (2013).
53. Smith, J. J. et al. Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution. *Nat. Genet.* **45**, 415–421 (2013).
54. Kuraku, S. Palaeophylogenomics of the vertebrate ancestor—impact of hidden paralogy on hagfish and lamprey gene phylogeny. *Integr. Comp. Biol.* **50**, 124–129 (2010).
55. Irisarri, I. et al. Phylotranscriptomic consolidation of the jawed vertebrate timetree. *Nat. Ecol. Evol.* **1**, 1370–1378 (2017).
56. Brazeau, M. D. & Friedman, M. The origin and early phylogenetic history of jawed vertebrates. *Nature* **520**, 490–497 (2015).
57. Jaffe, D. B. et al. Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13**, 91–96 (2003).
58. Huang, S. et al. HaploMerger: reconstructing allelic relationships for polymorphic diploid genome assemblies. *Genome Res.* **22**, 1581–1588 (2012).
59. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
60. Hu, H. et al. Constrained vertebrate evolution by pleiotropic genes. *Nat. Ecol. Evol.* **1**, 1722–1730 (2017).

Acknowledgements

We thank J. Graves, J. Gerhart and S. Brenner for useful discussions. Work in Okinawa was supported by internal funds of OIST for the Molecular Genetics Unit (D.S.R., O.S. and F.M.) and the Marine Genomics Unit (N.S.). O.S. was supported by the Austrian Science Fund (grants P30686-B29 and P32190). J.-X.Y. was supported by a postdoctoral fellowship from Fondation ARC pour la Recherche sur le Cancer (PDF20150602803). J.-K.Y., C.-H.T. and T.-K.H. were supported by grants from Academia Sinica (AS-98-CDA-L06) and the Ministry of Science and Technology, Taiwan (104-2923-B-001-002-MY3 and 105-2628-B-001-003-MY3). R.E.G. and B.O. were supported by a grant from the Gordon and Betty Moore Foundation (GBMF3804). A.B. was supported by a training grant from the US National Human Genome Research Institute (T32HG000047). D.S.R. is grateful for support from the Marthella Foskett Brown Chair in Biological Sciences. The work conducted by the US Department of Energy Joint Genome Institute (J.S. and J.J.) was supported by the Office of Science of the US Department of Energy under contract number DE-AC02-05CH11231. Computational analyses were conducted on the OIST High Performance Computing Cluster and the Life Science Compute Cluster at the University of Vienna.

Author contributions

O.S., F.M. and D.S.R. conceived of and led the study. J.-X.Y., J.J. and J.S. assembled the shotgun sequence and separated the haplotypes. B.O. and R.E.G. carried out the chromatin conformation capture. R.C. and N.H.P. analysed the chromatin conformation data and produced the chromosome-scale assembly. C.-H.T., T.-K.H. and J.-K.Y. performed and raised the cross. N.S., F.M. and A.B. generated and analysed F1 the genetic map. F.M. annotated the genome. A.B. performed the statistical analyses. O.S. and D.S.R. analysed the synteny.

Competing interests

D.S.R. is a member of the Scientific Advisory Board of Dovetail Genomics. R.E.G. is the founder of Dovetail Genomics. N.H.P. is an employee of Dovetail Genomics. D.S.R., R.E.G. and N.H.P. are all shareholders in Dovetail Genomics. The other authors declare no competing interests.

Additional information

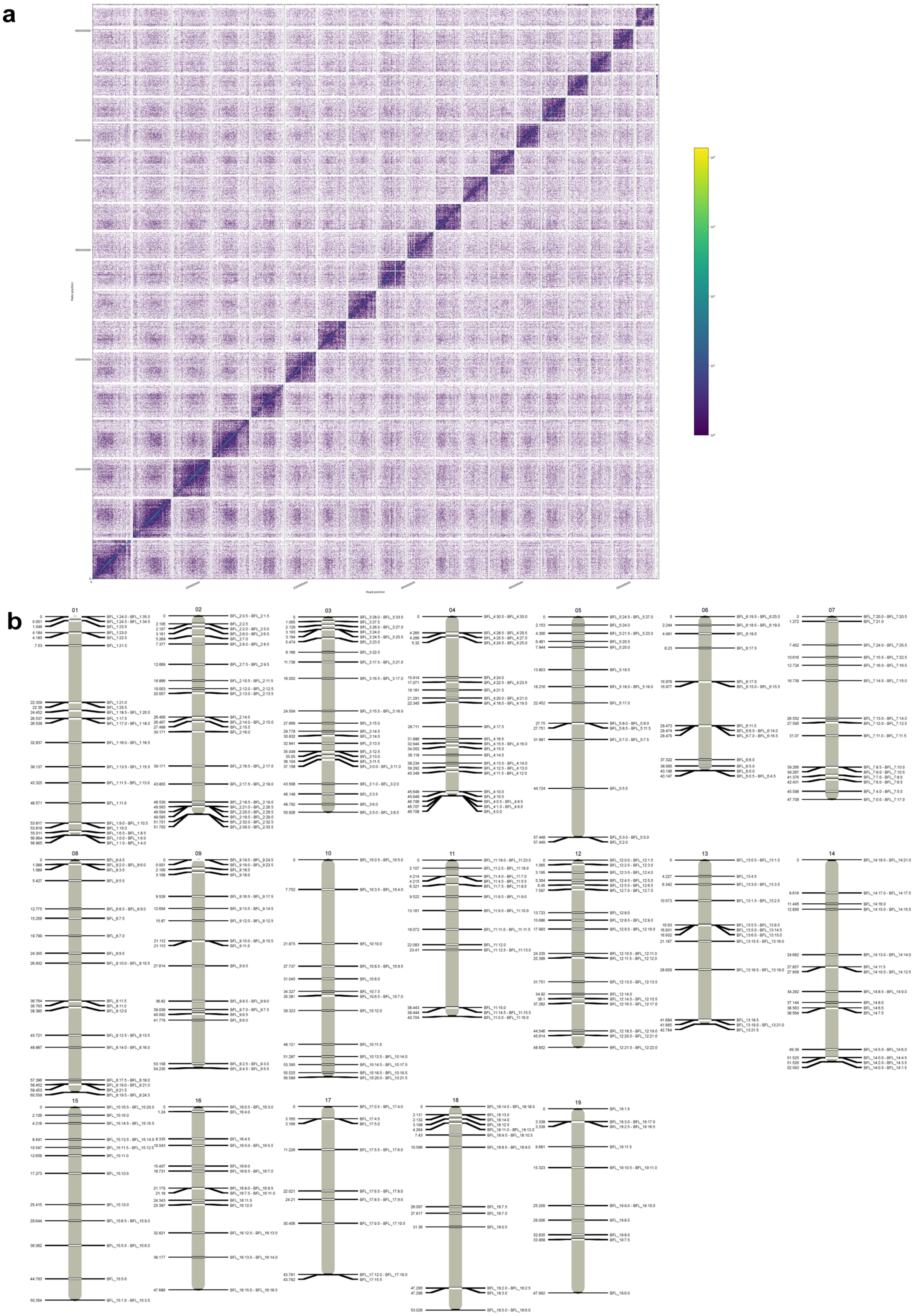
Extended data is available for this paper at https://doi.org/10.1038/s41559-020-1156-z.

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-020-1156-z.

Correspondence and requests for materials should be addressed to O.S. or D.S.R.

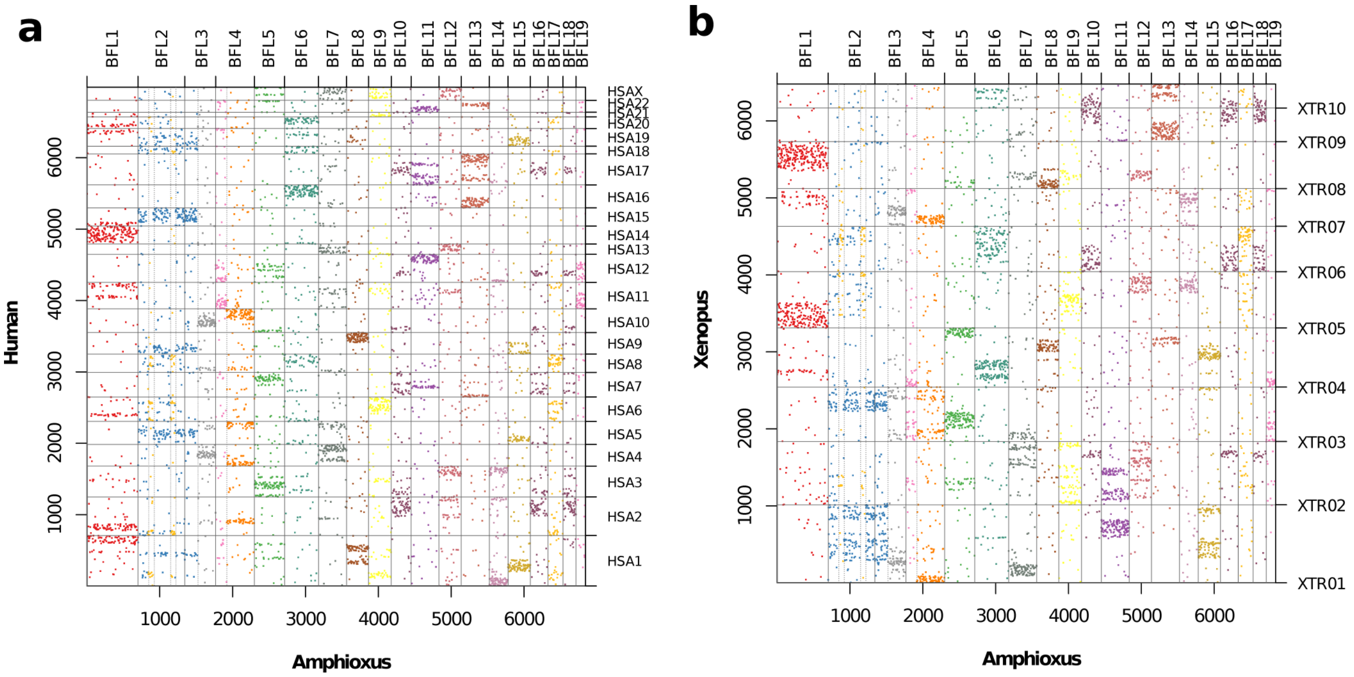
Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

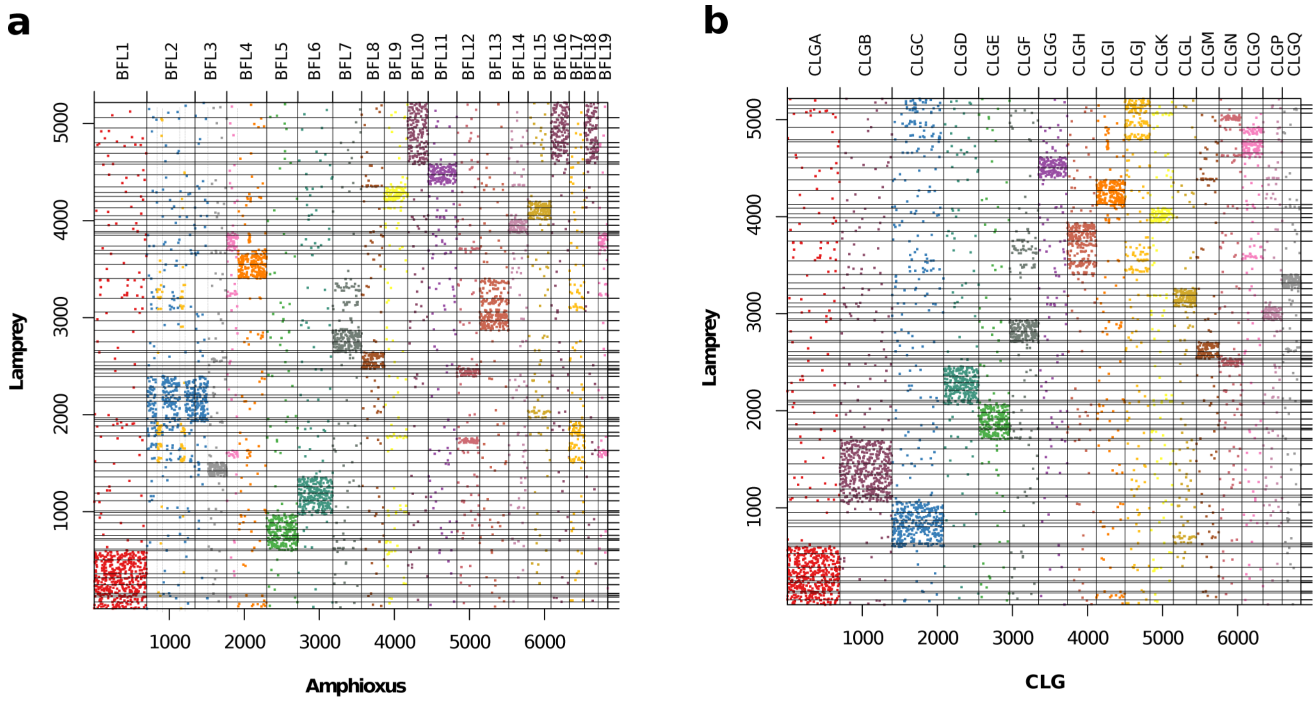


Extended Data Fig. 1 | See next page for caption.

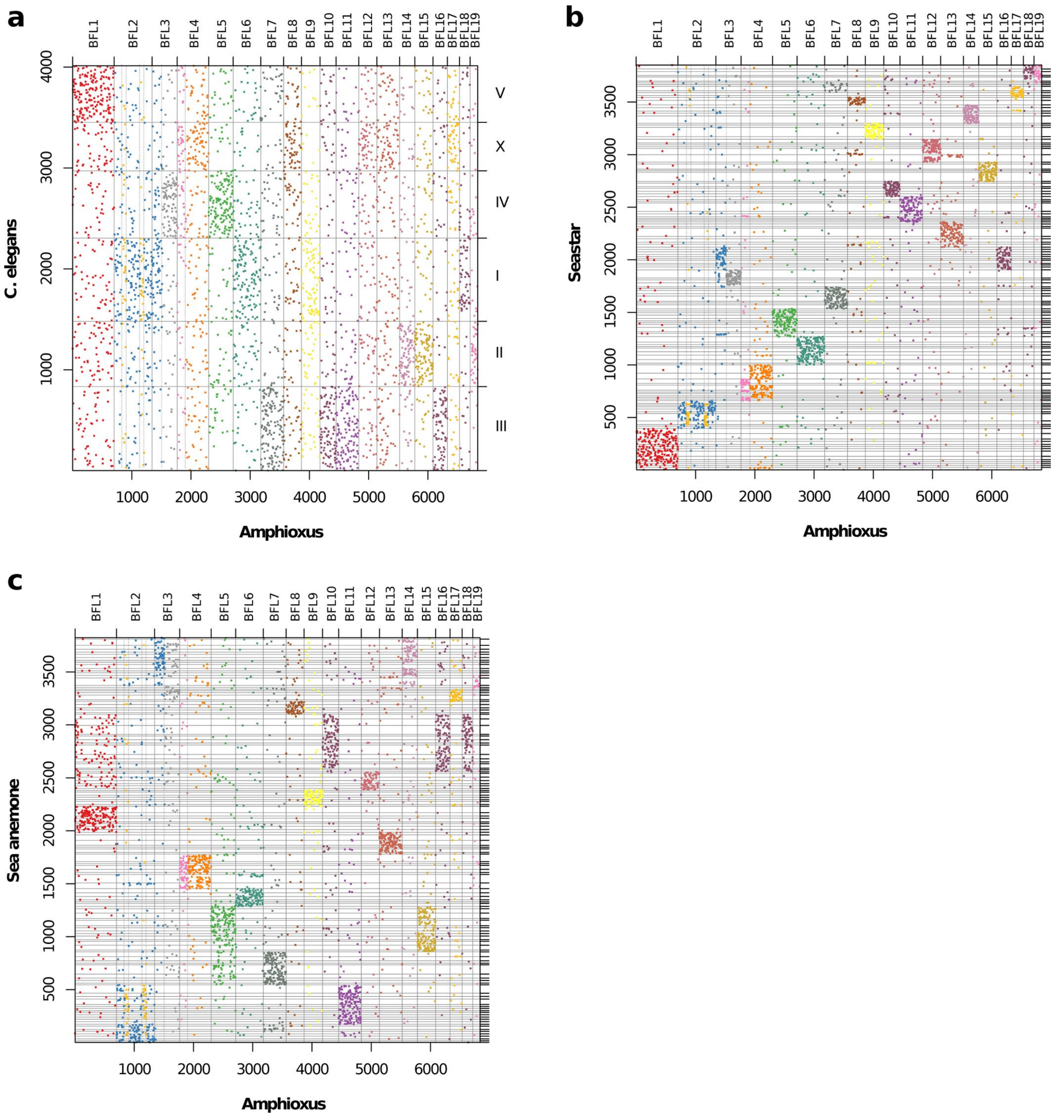
Extended Data Fig. 1 | Chromatin and genetic maps of amphioxus genome. (**a**): Chromatin conformation capture contact map for amphioxus genome assembly. Density of read-pairs representing three-dimensional chromatin contacts are shown as a heat map. (**b**): Maternal meiotic linkage map of amphioxus from a 96 progeny F1 cross. Markers represent phased 500 kb windows of the chromosomal assembly; consecutive windows are combined when there is no evidence for recombination in the genotyped progeny. Amphioxus linkage groups and the 19 longest assembled scaffolds are in 1:1 correspondence, confirming the Hi-C-based chromosome-scale assembly. (See Supplementary Note 4.).



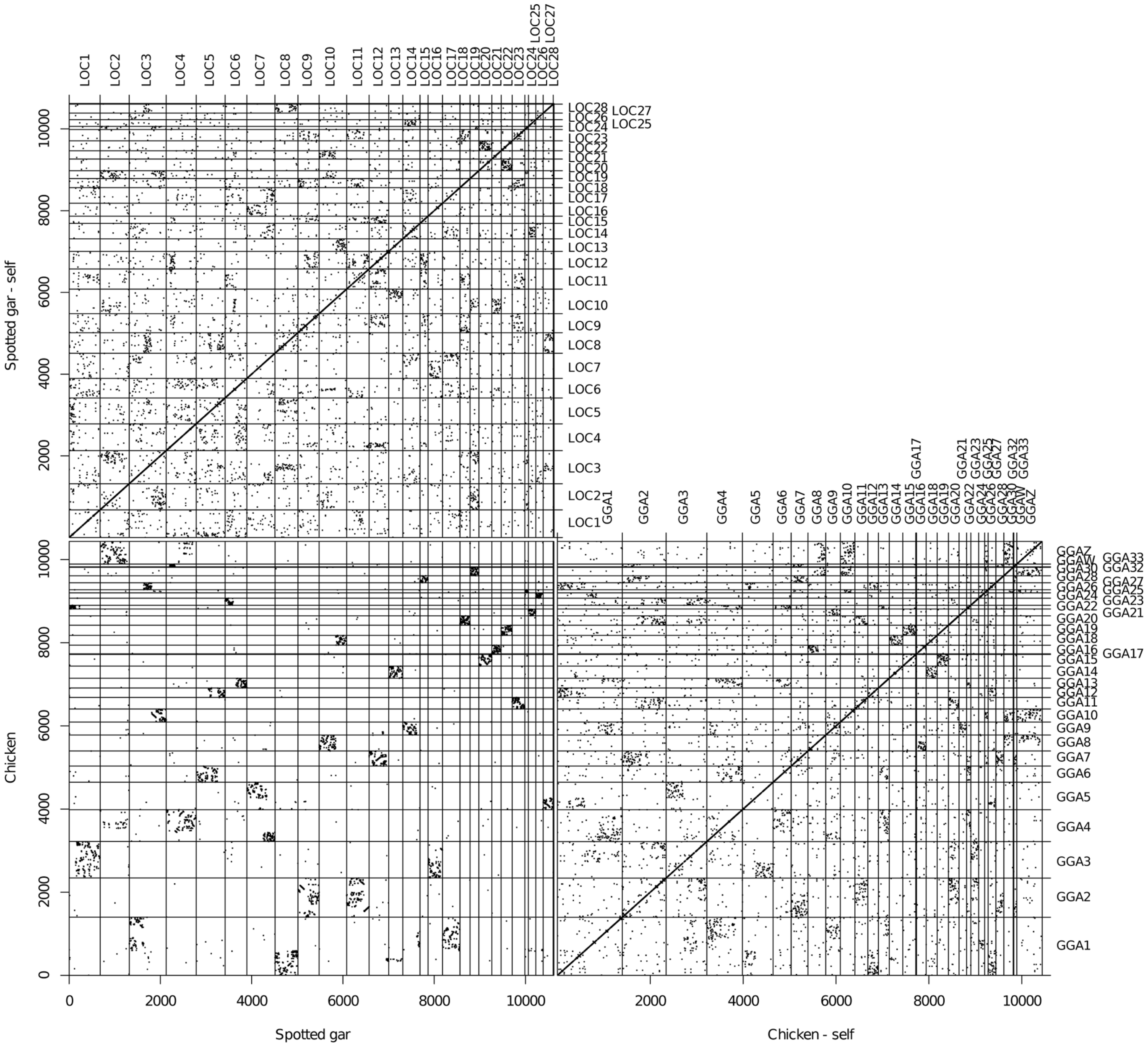
Extended Data Fig. 2 | Dot-plots showing conserved syntenies between amphioxus and human and frog. Dots represent mutual best hits between amphioxus and frog (*Xenopus tropicalis*, XTR) and human (*Homo sapiens*, HSA). Only mutual-best-hits involving the 6,843 genes of Fig. 1 are considered. (These gene families are anchored by mutual best hits between the four jawed vertebrate representatives and amphioxus.) Genes are colored based on their CLG membership as in main Fig. 1. Horizontal and vertical solid lines represent chromosome boundaries; vertical dashed lines represent inferred synteny breakpoints in amphioxus as in Fig. 1 (Methods).



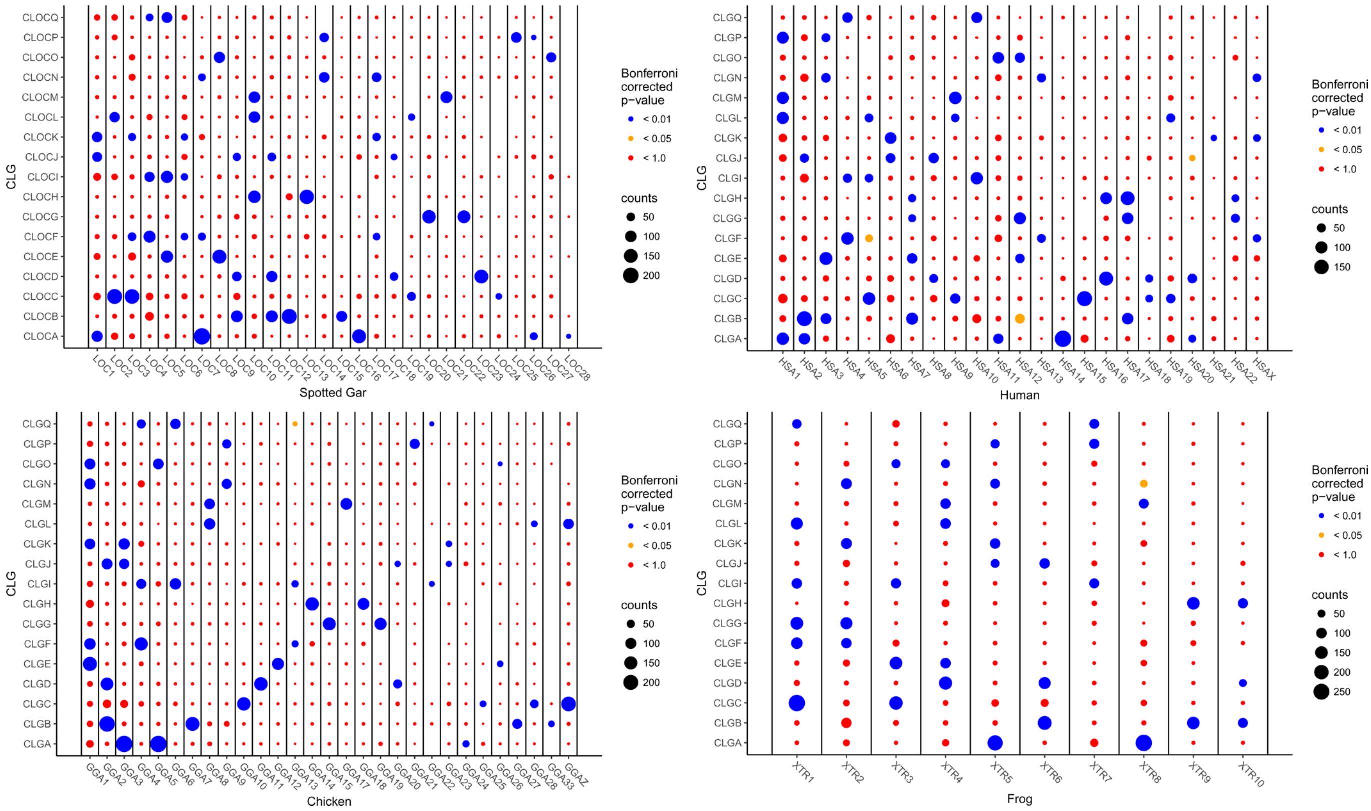
Extended Data Fig. 3 | Dot-plots showing conserved syntenies between lamprey and amphioxus. Dots represent mutual best hits between the (germline) genome of the sea lamprey (*Petromyzon marinus*) and amphioxus. Genes are colored based on their CLG assignment, and lamprey chromosomes are sorted according to their CLG content. Panel a shows these distribution of orthologous genes vs. amphioxus chromosomes, revealing the same discontinuities (vertical dashed lines) in amphioxus-lamprey synteny as found for amphioxus-bony vertebrate comparisons shown in Fig. 1 and Extended Data Fig. 2. Panel b shows these same orthologous gene pairs versus CLGs.



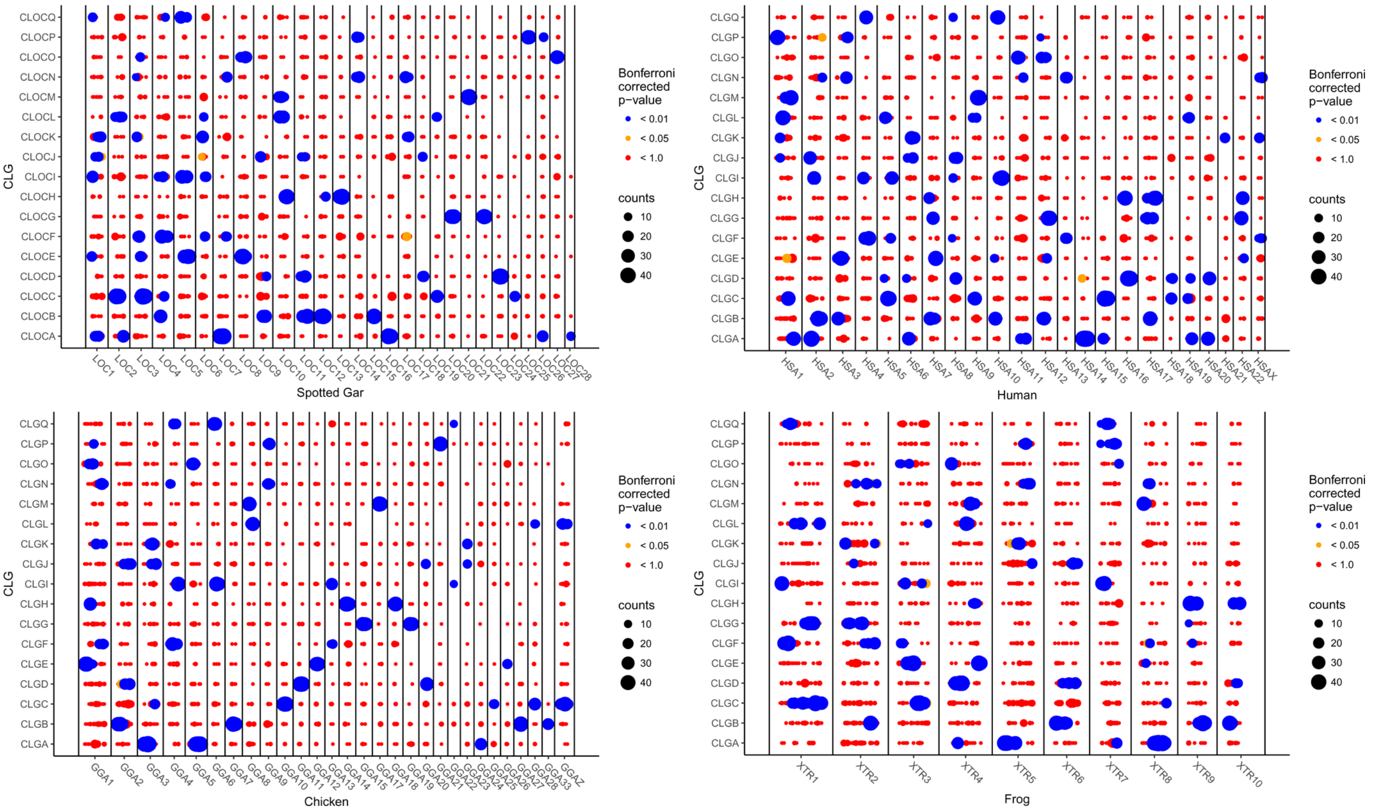
Extended Data Fig. 4 | Dot-plots showing conserved syntenies between amphioxus and selected invertebrates. Dots represent mutual best hits between amphioxus and the genomes of the Crown Of Thorns sea star Acanthaster planci, the soil nematode Caenorhabditis elegans, and the starlet sea anemone Nematostella vectensis. The N. vectensis and A. planci genomes are not yet assembled into chromosomes, and only scaffolds containing 20 or more genes (counting only mutual-best-hit vs. amphioxus) are shown. Scaffolds are sorted based on clustering using similarity of their CLG content. Vertical dashed lines are as shown in and Fig. 1, with the same CLG-based coloring, showing that the partitioning of amphioxus found using jawed vertebrates is also consistent with diverse invertebrates, and that sea star and sea anemone scaffolds can be grouped according to conserved synteny with amphioxus. C. elegans chromosomes arose by fusion, translocation, and mixing of the ancestral bilaterian units that are still retained in amphioxus.



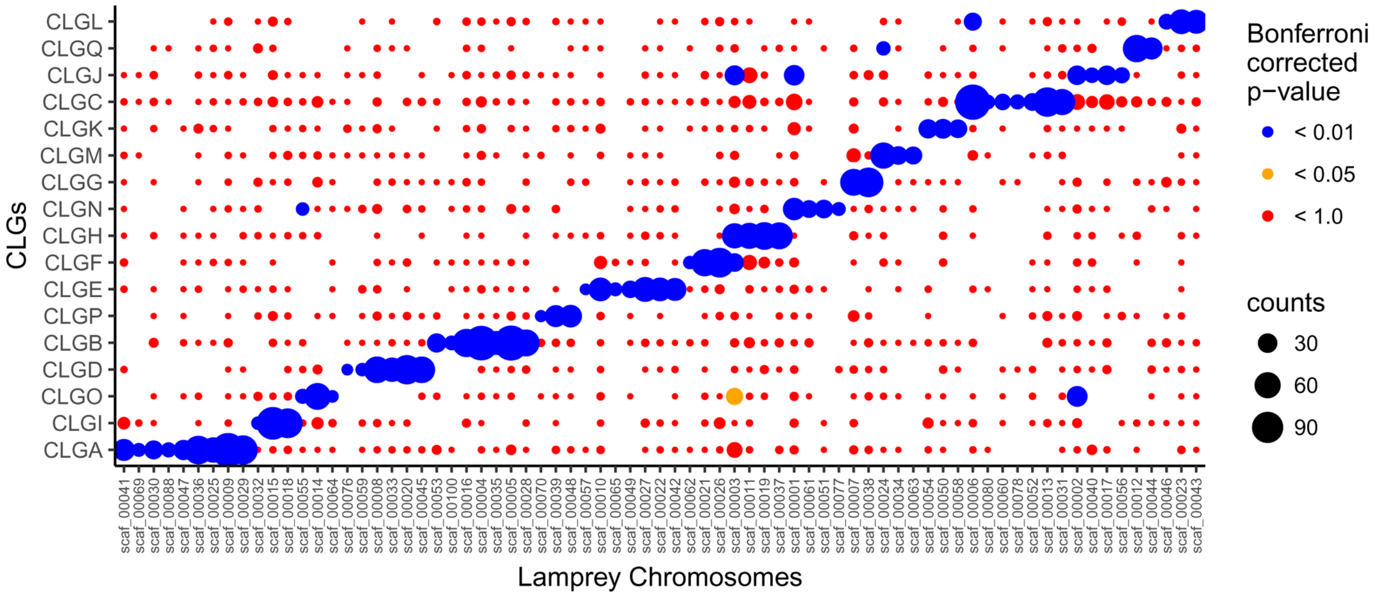
Extended Data Fig. 5 | Chicken-spotted gar orthologs and paralogs. "Oxford' dotpot between chicken (Gallus gallus, GGA) and spotted gar (Lepisosteus oculatus, LOC). Dots in the lower left corner represent mutual best hits between chicken and spotted gar, showing the clear orthologous blocks conserved synteny that allows chicken and spotted gar chromosome segments to be placed in correspondence with each other. Upper left and lower right show intra-genomic non-self best hits that identify paralogous regions within the chicken and spotted gar genome, respectively. Paralogous chromosomal regions share the same chordate linkage group ancestry, but arose through duplication.



Extended Data Fig. 6 | Oxford grid between bony vertebrate chromosomes and chordate linkage groups (CLGs). Circles represent the number of orthologous genes between human (Homo sapiens, HSA), chicken (Gallus gallus, GGA), frog (Xenopus tropicalis, XTR), and spotted gar (Lepisosteus oculatus, LOC) and the seventeen chordate linkage groups (CLGs) described in the text. Orthology is operationally defined by mutual best hits, restricted to 6,843 gene families anchored by mutual best hits of the four jawed vertebrates to amphioxus. The area of each circle is proportional to the number of orthologous genes for each chromosome-CLG pair, and the color indicates the significance of the association relative to a null model in which the position of the orthologous genes are randomly shuffled (Methods).



Extended Data Fig. 7 | Oxford grid showing associations between 50 gene segments of bony vertebrate chromosomes and chordate linkage groups (CLGs). Given the evident localization of orthologs along bony vertebrate chromosomes shown in the 'Oxford' dotplots of Fig. 1 and Extended Data Figs. 2 and 3, we assessed the significance of associations between sub-chromosomal regions and the chordate linkage groups. Each vertebrate chromosome was divided into overlapping 50 gene windows (offset by 25 genes). Only 6,843 genes with amphioxus-bony vertebrate mutual best hits are used. Circle areas are proportional to the number of orthologous genes for each chromosome-CLG pair, and the color indicates the significance of the association relative to a null model in which the position of the orthologous genes are randomly shuffled. Comparing Extended Data Figs. 6 and 7 shows additional significant associations that are missed based on whole chromosome analyses.



Extended Data Fig. 8 | Oxford grid between sea lamprey germline chromosomes and chordate linkage groups (CLGs). Circles show the number of orthologous genes between germline chromosomes of sea lamprey (Petromyzon marinus, PMA, denoted scaff_XXXXX following Smith et al.) and the seventeen CLGs described in the main text. As in Extended Data Fig. 6 the area of each circle is proportional to the number of orthologous genes for each chromosome-CLG pair, and the color indicates the significance of the association relative to a null model in which the position of the orthologous genes are randomly shuffled. Sea lamprey chromosomes and CLGs are both sorted to exhibit the striking correspondence between them. Each of the 17 CLGs is represented by at least one lamprey chromosome, and typically 6-8 lamprey chromosomes are associated with teach CLG. Compare Fig. 4b of Smith and Keinath 2015, which compares lamprey chromosomes to 'putative ancestral linkage groups' derived by Putnam 2008 through clustering of amphioxus scaffolds These putative ancestral linkage groups are in 1:1 correspondence with the CLGs shown here to be represented as large chromosomal segments of amphioxus.

natureresearch

Corresponding author(s): Simakov, Rokhsar

Last updated by author(s): Feb 14, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The raw data and assembly are available under the NCBI Project PRJNA412957
Data analysis	Data was analysed using software described in the Methods section. Custom code and processed data is available under https://bitbucket.com/viemet/public

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

PRJNA412957

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	N/A
Data exclusions	no exclusion
Replication	See Methods and Supplementary Information for exact procedure
Randomization	See Methods and Supplementary Information for exact procedure
Blinding	N/A

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	□	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology
□	☒	Animals and other organisms
☒	□	Human research participants
☒	□	Clinical data

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Branchiostoma floridae
Wild animals	N/A
Field-collected samples	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.