

1 Structure and evolutionary history of a large family of NLR 2 proteins in the zebrafish

3
4
5 Kerstin Howe^{1*}, Philipp H. Schiffer^{2,3*}, Julia Zielinski², Thomas Wiehe², Gavin K. Laird¹,
6 John C. Marioni^{1,4}, Onuralp Soylemez^{5,6}, Fyodor Kondrashov^{5,6,7}, Maria Leptin^{2,3#}

7
8 ¹ Wellcome Trust Sanger Institute, Cambridge, United Kingdom

9 ² Institut für Genetik, Universität zu Köln, Köln, Deutschland

10 ³ The European Molecular Biology Laboratory, Heidelberg, Germany

11 ⁴ The European Molecular Biology Laboratory, The European Bioinformatics Institute
12 (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, United Kingdom

13 ⁵ Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG) 88 Dr.
14 Aiguader, 08003 Barcelona, Spain.

15 ⁶ Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain.

16 ⁷Institució Catalana de Recerca i Estudis Avançats (ICREA), 23 Pg. Lluís Companys, 08010
17 Barcelona, Spain.

18 *Contributed equally

19 **Keywords:** NACHT, B30.2, SPRY, Pysin, gene family expansion, gene conversion, innate
20 immune system, genome evolution

21 **Running title:** *Evolution and structure of the NLR-B30.2 family*

22
23 kerstin@sanger.ac.uk, ORCID: 0000-0003-2237-513X

24 philipp.schiffer@gmail.com, ORCID: 0000-0001-6776-0934

25 julia.zielinski@uk-koeln.de

26 twiehe@uni-koeln.de

27 gkl@sanger.ac.uk, ORCID: 0000-0003-2294-0135

28 marioni@ebi.ac.uk, ORCID: 0000-0001-9092-0852

29 onuralp@gmail.com, ORCID, 0000-0001-8308-6855

30 fyodor.kondrashov@crg.eu

31 mleptin@uni-koeln.de, ORCID: 0000-0001-7097-348X

32
33 #corresponding author:

34 Prof. Dr. Maria Leptin

35 EMBO Director

36 Director's Research Unit

37 EMBO/EMBL Heidelberg

38 Tel: +49 (0)6221 8891101

39 Meyerhofstraße 1

40 69117 Heidelberg, Germany

41
42 All online supplementary material can be found on Figshare under the accession number

43 <http://dx.doi.org/10.6084/m9.figshare.1473092>

44

45 **ABSTRACT**

46 Animals and plants have evolved a range of mechanisms for recognizing noxious substances
47 and organisms. A particular challenge, most successfully met by the adaptive immune system
48 in vertebrates, is the specific recognition of potential pathogens, which themselves evolve to
49 escape recognition. A variety of genomic and evolutionary mechanisms shape large families
50 of proteins dedicated to detecting pathogens and create the diversity of binding sites needed
51 for epitope recognition. One family involved in innate immunity are the NACHT-domain-and
52 Leucine-Rich-Repeat-containing (NLR) proteins. Mammals have a small number of NLR
53 proteins, which are involved in first-line immune defense and recognize several conserved
54 molecular patterns. However, there is no evidence that they cover a wider spectrum of
55 differential pathogenic epitopes. In other species, mostly those without adaptive immune
56 systems, NLRs have expanded into very large families. A family of nearly 400 NLR proteins
57 is encoded in the zebrafish genome. They are subdivided into four groups defined by their
58 NACHT and effector domains, with a characteristic overall structure that arose in fishes from
59 a fusion of the NLR domains with a domain used for immune recognition, the B30.2 domain.
60 The majority of the genes are located on one chromosome arm, interspersed with other large
61 multi-gene families, including a new family encoding proteins with multiple tandem arrays of
62 Zinc fingers. This chromosome arm may be a hot spot for evolutionary change in the
63 zebrafish genome. NLR genes not on this chromosome tend to be located near chromosomal
64 ends.

65 Extensive duplication, loss of genes and domains, exon shuffling and gene conversion acting
66 differentially on the NACHT and B30.2 domains have shaped the family. Its four groups,
67 which are conserved across the fishes, are homogenised within each group by gene
68 conversion, while the B30.2 domain is subject to gene conversion across the groups.
69 Evidence of positive selection on diversifying mutations in the B30.2 domain, probably
70 driven by pathogen interactions, indicates that this domain rather than the LRRs acts as a
71 recognition domain. The NLR-B30.2 proteins represent a new family with diversity in the
72 specific recognition module that is present in fishes in spite of the parallel existence of an
73 adaptive immune system.

74

75 INTRODUCTION

76 The need to adapt to new environments is a strong driving force for diversification during
77 evolution. In particular, pathogens, with their immense diversity and their ability to subvert
78 host defense mechanisms, force organisms to develop ways to recognize them and keep them
79 in check. The diversity and adaptability of pathogen recognition systems rely on a range of
80 genetic mechanisms, from somatic recombination, hypermutation and exon shuffling, to gene
81 conversion and gene duplication to generate the necessary spectrum of molecules.

82 The family of NACHT-domain (Koonin and Aravind 2000) and Leucine Rich Repeat
83 containing (NLR) proteins (reviewed in Proell et al., 2008, Ting et al., 2008) act as sensors
84 for sterile and pathogen-associated stress signals in all multicellular organisms. In
85 vertebrates, a set of seven conserved NLR proteins are shared across a wide range of species.
86 These are the sensor for apoptotic signals, APAF1, the transcriptional regulator CIITA, the
87 inflammasome and nodosome proteins NOD1, NOD2, NOD3/NlrC3, Nod9/NlrX1 and the as
88 yet functionally uncharacterized NachtP1 or NWD1 (Stein et al. 2007; Kufer and Sansonetti
89 2011). Other NLR proteins, which must have evolved independently of the conserved NLR
90 proteins, are shared by only a few species, or are unique to a species. Some non-vertebrates,
91 such as sea urchins or corals, have very large families of NLR-encoding genes (Bonardi et al.
92 2012), but an extreme example of species-specific expansion can be found in zebrafish (Stein
93 et al. 2007). Such species-specific gene family expansions suggest adaptive genome
94 evolution in response to specific environments, most probably different pathogens (Liu et al.
95 2013).

96 The zebrafish has become a widely used model system for the study of disease and
97 immunity (Rowe, Withey, and Neely 2014; Goody, Sullivan, and Kim 2014), and a good
98 understanding of its immune repertoire is necessary for the interpretation of experimental
99 results, for example in genetic screens or in drug screens. In a previous study we discovered
100 more than 200 NLR-protein encoding genes (Stein et al. 2007). The initial description and
101 subsequent analyses (Laing et al. 2008; van der Aa et al. 2009) have led to the following
102 conclusions: The zebrafish specific NLRs have a well-conserved NACHT domain
103 (PF05729), with a ~70 amino-acid extension upstream of the NACHT domain, the Fisna-
104 domain (PF14484, see this paper). This domain characterises this class of NLR proteins and
105 is found in all sequenced teleost fish genomes, but not outside the fishes (Stein et al. 2007).
106 The NLR proteins can be divided into four groups, each defined by sequence similarity in the
107 NACHT and Fisna domain, and these groups also differ in their N-terminal motifs. Groups 1
108 and 2 have death-fold domains; groups 2 to 4 contain repeats of a peptide motif that is only

109 found in this type of NLR protein (Fig. 1). In the initial description, all of the novel NLR
110 proteins ended with the Leucine-Rich-Repeats, but it was later found (van der Aa et al. 2009)
111 that several of them had an additional domain at the C-terminus, an
112 SPRY/B30.2 domain (PF00622). This domain also occurs in another multi-gene family
113 implicated in innate immunity, the fintrims (van der Aa et al. 2009).

114 The initial identification of the genes and subsequent analyses by others (Laing et al.
115 2008) suffered from the limitations of the then available Zv6 assembly and gene annotations
116 (published in 2006). In particular, there was a limited amount of data available for long-range
117 assembly arrangements and a lack of supporting evidence for gene models, such as well
118 annotated homologues from other species. In addition, the very high similarity of the NLR
119 genes, as well as their clustered arrangement in the genome further complicated the assembly.
120 As a result, many genomic regions were collapsed and many of the gene models were
121 incomplete. We have revisited this gene family to improve the genome assembly in the
122 regions of interest, have manually re-annotated and refined the gene structures and here
123 provide a full description and analysis. Our re-annotation and analysis of the NLR gene
124 family in zebrafish, and comparison with other species, shows a structured set of more than
125 400 genes. Different evolutionary pressures on the different parts of the proteins may have
126 created functional diversification between the groups.

127

128

129 **RESULTS**

130

131 *1. Identification of all NLR genes in the zebrafish genome*

132 To identify the entire set of fish-specific NLR encoding genes in the zebrafish genome,
133 we used various approaches to collect lists of candidate genes based on the Zv9 assembly
134 (GCA_000002035.2), (for details see Methods and Supplementary Methods). We identified
135 genomic regions containing domain motifs via hmmsearch
136 (hmmer.janelia.org/search/hmmsearch), electronic PCR (Schuler 1997), TBLASTN searches,
137 and by mining the existing annotations for keywords. This collection was purged of gene
138 models belonging to other known families, e.g. fintrims. We identified all overlapping
139 Ensembl and VEGA gene models for the remaining regions of interest. The VEGA gene
140 models were refined and extended through manual annotation and both gene sets merged,
141 resulting in 421 NLR gene models.

142 Beyond the seven conserved NLR genes, and nine other NLR genes (Table 1) that had
143 a different structure from those described previously and below, the zebrafish genome
144 contains 405 genes encoding NLR proteins that are members of the family we had previously
145 called 'novel fish NLR proteins' (Stein et al. 2007). Henceforth, we will refer to these as
146 NLR-B30.2 proteins (see below). From the 405 genes 37 were pseudogenes leaving 368
147 predicted protein-coding NLR-B30.2 genes (online supplementary material, Figshare:
148 <http://dx.doi.org/10.6084/m9.figshare.1473092>). The genome assembly components carrying
149 these gene models were checked for correct placement and relocated if necessary, thereby
150 contributing to improvements for the GRCz10 assembly (GCA_000002035.3).

151

152 2. Structure, conservation and divergence

153

154 *2a. Domain structure of the NLR family members*

155 The original set of 205 genes described in (Stein et al. 2007) was divided into four
156 groups based on sequence similarity in the Fisna and NACHT domains, and the sequence
157 elements upstream of the Fisna domain. The current dataset shows that all four groups share
158 the Fisna domain, the NACHT domain and the LRRs. Other domains are present only in
159 subsets of the genes; these include a death-fold domain, a B30.2 domain and various N-
160 terminal peptide repeats (Fig. 1 and Supplementary Fig. 1).

161

162 *Fisna and NACHT.* The sequence similarities in the Fisna and the NACHT domain in our
163 updated gene set confirm our previous subdivision of the genes into four main groups. A
164 defining motif for each group is the sequence of the Walker A motif, for which the consensus
165 over the whole set is G[IV]AG[IV]AGK[TS] with each group having its own, conserved
166 motif (Fig. 1, Supplementary Fig. 2). Some of the groups can be subdivided further. For
167 example, group 2 consists of two subgroups (Fig. 1, Supplementary Fig. 1), one large and
168 very homogeneous (group 2a), the other smaller, and also homogeneous, with all genes
169 located in a cluster on chromosome 22 (group 2b). Moreover, group 3 has several subgroups
170 that differ in their N-terminal peptides, or their LRRs, or their B30.2 domains (Fig 1.).

171 We previously found good matches for the Fisna-domain only in fish NLR proteins.
172 Thus, the presence of this domain in conjunction with the particular Walker A motifs is a
173 diagnostic property of the protein family, although short stretches of this domain resembled
174 peptides within mammalian Nod1 and Nod2 (Stein et al. 2007). We used our new collection

175 of Fisna sequences to build an HMM defining a Pfam family (deposited as PF14484) and to
176 search for homologies within mammalian proteins. This revealed alignments with high
177 significance to mammalian members of the NLR family (NLRP3 and NLRP12), with the
178 matching sequence located immediately upstream of the NACHT domain. Secondary
179 structure predictions based on two representatives from zebrafish and the rat using the
180 PSIPRED workbench suggest that the zebrafish Fisna domain may take on the same
181 conformation as the corresponding mammalian proteins (Fig. 2). Together, these findings
182 suggest that contrary to previous assumptions, the Fisna domain was present in the common
183 ancestor of mammalian and fish NLR proteins. Whether it forms a separate domain, or is
184 simply an N-terminal extension of the NACHT-domain, similar to the additional helices seen
185 in the structure of another NLR protein, Apaf1, remains to be determined by experimental
186 work. We did not find Fisna domains in non-vertebrate genomes.

187

188 *Death-fold domains and N-terminal repeats.* The four group-specific similarities
189 continue upstream of the Fisna-domain: Groups 1 and 2 both have a death-fold domain; a
190 Pyrin-domain (PYD) with an N-terminal peptide characteristic of BIR domains
191 (http://elm.eu.org/elmPages/LIG_BIR_II_1.html) in the former, and a PYD-like domain in
192 the latter.

193 The predicted proteins of groups 2, 3 and 4 contain several repeats of a ~30 amino acid
194 peptide motif. The repeats are not all identical, but occur in two main types. Surprisingly,
195 both of the main types of N-terminal repeat can associate with either group 2 or group 3
196 NACHT domains (Supplementary Fig. 1: group 2a (N-ter1) and group 2a (N-ter2)). ,
197 indicating extensive exon-shuffling between family members.

198

199 *LRRs.* The proteins of all groups contain several Leucine Rich Repeats (LRRs; Pfam
200 Clan: CL0022). The LRRs in groups 1, 2a and 3 are very similar to each other and occur in a
201 similar pattern; each protein contains between 2 and 7 repeats, with each repeat consisting of
202 two LRRs (Supplementary Fig. 3). There are two types of LRR repeat: the last LRR,
203 immediately upstream of the B30.2 domain, occurs in each gene exactly once, and barely
204 differs between the genes. The other type of LRR in groups 1, 2a and 3 can vary in number
205 between 0 and 6, and are more divergent in sequence. Thus, similar to the situation with the
206 LRRs in the lamprey VLR proteins (Rast and Buckley 2013), the C-terminal LRR seems to
207 be fixed, whereas the others vary more and are duplicated to varying degrees. Groups 2b and

208 4 do not show this arrangement, but they have yet another type of LRR, which can occur up
209 to more than 20 times.

210 *SPRY/B30.2 domain.* A B30.2 domain has so far been reported to be present in some
211 but not all fish NLR proteins (van der Aa et al. 2009). Our current set of sequences shows
212 that the presence of the B30.2 domain is restricted to groups 1, 2a and 3, and the domain is
213 missing in groups 2b and 4 (Fig. 1, Supplementary Fig.s 1,4). In view of the extreme
214 similarity between the N-terminal parts (NACHT and death-fold domain) of groups 2a and
215 2b, and the overall conservation of the gene structure throughout the whole family, it seems
216 most likely that the B30.2 domain was present in the common ancestor of the family, but lost
217 by groups 2b and 4, rather than independently gained by the other groups. We therefore refer
218 to the entire family as the NLR-B30.2 protein family.

219

220 *2b. Exon-intron structure*

221 All genes in this family have the same exon-intron structure (Supplementary Fig. 2).
222 The largest exon contains the NACHT domain with the N-terminal Fisna extension and the
223 winged helical and superhelical domains, as is also the case in NLRC3, for example. All
224 other domains (N-terminal peptides, LRRs, B30.2 domain) are each encoded on single
225 exons.

226

227 *2c. Sequence variation between the groups*

228 Visual inspection of the aligned amino acid sequences of all four groups showed that
229 some parts of the proteins were highly conserved, while others showed many amino acid
230 substitutions. The NACHT domains of groups 1, 2a and 3a show little sequence divergence
231 within each group, but strong divergence between the groups (Supplementary Fig. 2). In
232 contrast, there is no recognizable group-specific sequence pattern in the B30.2 domain.
233 Instead, a number of apparently sporadic substitutions are seen throughout the entire set (Fig.
234 1). This observation is supported by independent phylogenetic analyses of the domains that
235 show different tree topologies (Online supplementary material). The NACHT domains cluster
236 into monophyletic groups, as expected, since similarity in the NACHT domain was used to
237 define the group. In contrast, the tree for the B30.2 domain shows a more interleaved pattern.
238 The discrepancy between the trees suggests different evolutionary trajectories for the B30.2
239 and NACHT domains. Thus, on the one hand, proteins with different NACHT domains share

240 similar B30.2 domains, and on the other, proteins with nearly identical NACHT domains and
241 N-terminal motifs, such as those in groups 2a and 2b, have different C-termini.

242

243 *3. Evolutionary history*

244 *3a. Conservation and divergence within and between the four zebrafish NLR-B30.2 groups*

245 Both the shuffling of N-termini and the unequal divergence of the NACHT and B30.2
246 domains suggest a complex evolutionary history of the gene family. To analyse the
247 divergence, we calculated rates of non-synonymous and synonymous substitutions in the
248 NACHT and B30.2 domains (dN, dS, and dN/dS values). We studied only those groups that
249 show high intra-group conservation of the NACHT domains (groups 1, 2a, 3a and 3b). We
250 considered groups 3a and 3b separately because inspection of the protein alignment
251 suggested that although they belong into the same group by virtue of their NACHT domain,
252 their B30.2 domains had diverged. We omitted group 3c, as its NACHT domains are more
253 divergent, and may represent further groupings. Median values are in Table 2 and all data
254 displayed in Supplementary Fig. 5.

255 *Synonymous substitutions.* The median rate of synonymous sequence substitutions in
256 the NACHT domains was very low when comparing members within a group (0.01 - 0.05).
257 The values for comparisons between groups were 20- to 100-fold higher. The only case
258 where the between-group comparison also gave a low value was for group 3a versus 3b,
259 confirming their classification by NACHT domain as one group.

260 The B30.2 domains showed a different pattern. Similar to the NACHT domain, the
261 median dS values for comparisons within each group ranged between 0.02 and 0.04 for
262 groups 1, 2a and 3a. However, for the B30.2 domain, the values for the between-group
263 comparisons were low: they were minimally or not at all higher than for within-group
264 comparisons. The B30.2 domain sequences of the members of group 3c were more divergent,
265 both from each other those in groups 1, 2a and 3a (Table 2, Supplementary Fig. 5). The
266 patterns of synonymous divergence between groups were clearly different for NACHT and
267 B30.2 domains and confirmed the different behaviour of the two domains we had noted from
268 the alignment of the protein sequences. The observed patterns are most easily explained by
269 gene-conversion (see Discussion).

270 *Non-synonymous/synonymous substitutions.* Our calculations of the dN/dS values for
271 these comparisons reveal a second evolutionary mechanism acting in the B30.2 domain. We
272 find high median dN/dS values, indicating positive selection acting on the B30.2 domain

273 (Supplementary Fig. 5). Thus, we confirm and extend a previous conclusion for the B30.2
274 domain in the fintrim proteins (van der Aa et al. 2009), that positive selection might create
275 variation for pathogen recognition in this domain. We discuss below how the combination of
276 positive selection and gene conversion may have created variation of the B30.2 domain
277 throughout the entire family.

278

279 *3b. Origin of the NLR-B30.2 gene groups*

280 The degree of apparent gene conversion within the NLR-B30.2 gene family makes it
281 difficult or impossible to judge when the groups arose or when they expanded during fish
282 evolution. Moreover, the high divergence between the four groups suggests the split into
283 groups may be old. To explore whether the groups arose in the zebrafish or in an ancestral
284 species, we compared the NLR-B30.2 genes in the zebrafish with those in the closest relative
285 for which a whole genome sequence exists, the carp, as well as NLR-B30.2 genes in other
286 vertebrate genomes (see Materials and Methods).

287 A tree resulting from a recursive phylogenetic analysis indicates that the split into
288 groups occurred before the zebrafish-carp divergence (Fig. 3). Groups 1, 2, 3a/b and 4 have
289 orthologous relationships between carp and zebrafish: For example, group 2 from zebrafish is
290 most closely related to a group of genes in the carp that is distinct both from other carp NLR-
291 B30.2 genes and from the other zebrafish genes (see also a tree containing all available
292 zebrafish and carp genes in Supplementary Fig. 6).

293 Not unexpectedly, group 3c, which has a more heterogeneous set of sequences in the
294 zebrafish, shows a more complex evolutionary history. It falls into two groups, both of which
295 have an orthologous group in the carp.

296 *A. mexicanus* and *E. lucius* each have groups of genes that cluster with groups of the
297 zebrafish genes, rather than with each other, but not every group is represented in each of the
298 species. Nevertheless, this suggests that the split is even older than the carp-zebrafish split,
299 having perhaps occurred in basal teleosts. Finally, sequences from *L. oculatus*, *L. chalumnae*
300 and *C. milii* do not fall into these groups, suggesting that the split into groups must have
301 occurred in the Clupeocephala.

302 In summary, the groups did not diverge independently from duplicated ancestral genes
303 in each species, but already existed in a common precursor. By contrast, within a species, the
304 majority of genes arose by independent amplification of a founding family member.

305

306 *3c. Co-occurrence of the NACHT and B30.2 domains*

307 As first reported by van der Aa et al. 2009, the NLR-B30.2 domain fusion is found in
308 all teleost fish. Our collection of NLR-B30.2 genes showed that this domain fusion arose
309 prior to the common ancestor of teleosts, as it was also present in *L. oculatus* (for example
310 ENSLOCG0000000593), for which the genome sequence was not previously available. The
311 NLR-B30.2 proteins predicted in *L. oculatus* also contain an N-terminal Fisna extension,
312 though only distantly related to those in the teleosts, indicating that the ancestral gene
313 included this extension. This is consistent with the fact that the N-terminal extension of the
314 mammalian NLRP3 proteins has recognizable similarity to the Fisna domain, as described
315 above.

316 We do not find evidence of the NLR-B30.2 fusion in any of the tetrapod genomes, nor
317 in the *L. chalumnae* or *C. milii* genomes. These results indicate that the fusion occurred at
318 least in the common ancestor of the Neopterygii, prior to the third whole genome duplication
319 in the teleost lineage. Genome sequence data from sturgeon, paddlefish or bichir clades,
320 which are currently not available, would provide further information on the point of
321 emergence of the NLR-B30.2 fusion.

322

323 *3d. Expansion of the NLR-B30.2 family in fish*

324 The phylogeny of the NLR-B30.2 gene family also provides information on the timing
325 of the lineage expansion of the NLR-B30.2 genes observed in the zebrafish. The *C. carpio*
326 genome contains a similarly large family of NLR-B30.2 proteins, but due to the polyploidy of
327 the species some of the nearly identical sequences may constitute alleles rather than
328 paralogues. *A. mexicanus*, a direct outgroup to the zebrafish-carp clade, features the second
329 largest NLR-B30.2 gene family with ~100 members, showing that the lineage expansion
330 began prior to the zebrafish-carp split. Other fish species, including the spotted gar have
331 fewer than 10 NLR-B30.2 genes, while *E. lucius*, a Euteleost, has ~50. Thus, the initial gene
332 expansion either occurred in the basal branches of Teleostei with a subsequent loss in some
333 lineages, or independently in several lineages. Independent expansions and losses are a likely
334 scenario, given the expansions of NLR genes in many other species, such as sponges and sea
335 urchins (Yuen, Bayes, and Degnan 2014; Lapraz et al. 2006). The results on fish show that
336 expansions of the NLR-B30.2 family genes began as soon as the NLR-B30.2 fusion occurred,
337 with different dynamics in different lineages.

338

339 *3e. Age of the NLR-B30.2 family relative to conserved NLRs*

340 We compared the origin of the NLR-B30.2 gene family to the evolutionary history
341 of other NLR genes. There are many species-specific expansions of NLRs, such as the Nalp
342 proteins in mouse and human and the NLR-B30.2 genes in fish, as well as independent
343 inflations in *Amphioxus*, sea urchin, and sponge. However, there are also the seven NLR
344 proteins that are conserved in all vertebrates and show orthologous relationships. We
345 collected the available orthologues of the conserved NLRs from key metazoan species and
346 created an alignment. We also included all NLR genes from fish that did not belong to the
347 NLR-B30.2 group (listed above and in M&M).

348 We found that two of the conserved vertebrate NLR genes appear to be shared by all
349 animals (Fig. 5, Supplementary Fig. 6, and online supplementary material). The genes
350 for NWD1 (first described in zebrafish as NACHT-P1) and Apaf1, must have been present in
351 the last common ancestor of bilaterians and non-bilaterians, as they are found in sponges,
352 cnidarians, and all bilaterians analysed. We could not find any candidates in comb jellies
353 (ctenophores). The other five conserved NLR proteins - Nod1, Nod2, NLR3C, CIITA and
354 NLRX1 – arose later in evolution, at the base of the gnathostomes. An additional gene,
355 NLR3c-like, was present at this point, but appears to have been lost in the tetrapod lineage.

356 In summary, all of the conserved vertebrate NLRs are older than the NLR-B30.2
357 family. They are never duplicated and certainly not expanded to higher gene numbers in any
358 species.

359 Important events in the evolutionary history of the NLR-B30.2 family and conserved
360 NLR proteins are summarized in Fig. 4 and Supplementary Fig. 7. The oldest metazoan NLR
361 proteins are NWD1 and Apaf1, while Nod1, Nod2, NLR3C, CIITA and NLRX1 and
362 NLR3C-like were acquired in the lineage leading to gnathostomes. A fusion of the NLR and
363 B30.2 domains then occurred in the Neopterygian lineage and subsequently these genes
364 diversified into groups, within which they continued to expand in various fish lineages. The
365 high similarity of the NACHT domains within the groups appears to have been maintained by
366 gene conversion. Large-scale amplifications of NLR genes occurred independently in many
367 species.

368

369 *4. Genomic location*

370 The first survey of NLR genes on the zebrafish genome assembly Zv6 suggested that
371 they were located on 22 different chromosomes, with some enrichment on chromosome 4 (50

372 genes) and chromosome 14 (47 genes) (Stein et al. 2007). Since this analysis, the assembly of
373 the zebrafish genome has been significantly improved and the current Zv9 NLR gene set
374 shows a more restricted distribution (Fig. 5), with 159 (44%) of the genes located on the right
375 arm of chromosome 4. The remaining genes are distributed between 12 other chromosomes
376 (153 genes) and unplaced scaffolds (56 genes).

377 Additional sequencing and data gathering by the Genome Reference Consortium since
378 the release of Zv9 led to the rearrangement of multiple assembly components, including
379 relocation of sequence to different chromosomes. These placements are based on manual
380 curation by the Genome Reference Consortium, supported by genetic mapping data, clone
381 end sequence placements and optical mapping data (Howe and Wood 2015). The latest
382 assembly, GRCz10, reveals that the majority of the genes clustered on the long arm of
383 chromosome 4, where 75% of the NLR-B30.2 genes, including all group 1 and group 2a
384 genes, now reside (Fig. 5). Group 2b genes are now found exclusively in a cluster on
385 chromosome 22, which suggests that they arose via local duplications of a single precursor
386 gene that had lost its B30.2 domain. Similarly, group 3a genes are clustered together on
387 chromosome 4, with group 3b and 3c genes clustered on chromosomes 1 and 17,
388 respectively. Group 4 genes are found mostly on chromosome 15, some on chromosome 1,
389 with none found on chromosome 4. Both group 2 and group 3 have a few individual genes
390 dispersed over other chromosomes; careful inspection of the evidence on which this
391 allocation is based revealed no indications that it is incorrect. Some of the group 3 members
392 on other chromosomes are more divergent from the consensus for this group, suggesting they
393 may indeed have separated from the group early.

394 Within chromosome 4, no clear pattern can be detected in the distribution of the genes.
395 We are, however, aware of possible shortcomings in the assembly of the long arm of
396 chromosome 4; the highly repetitive nature of the sequence makes it difficult to exclude with
397 absolute certainty shuffling of gene locations. In addition to containing multiple copies of 5S
398 ribosomal DNA (Sola and Gornung 2001), 53% of all snRNAs and the majority of the NLR-
399 B30.2 genes, chromosome 4 also contains multiple copies of genes encoding a particular type
400 of Zn-finger protein, which we discuss below.

401 Finally, another striking feature of the genes' genomic location is that they tend to
402 accumulate near the ends of the chromosomes (Fig. 5b). With the exception of the cluster on
403 chromosome 22, and two single genes on Chromosomes 5 and 15, all other genes (81% of the

404 NLR genes outside chromosome 4) are located within 15% of chromosome ends. On
405 chromosome 4 we found 26% of the genes within 15% of the end.

406

407 5. Distribution of Fintrim and multiple Zn-finger encoding genes

408 We noticed that the NLR-B30.2 genes on chromosome 4 were often interspersed with
409 genes encoding multiple tandem Zn-finger proteins. In some cases, gene models had been
410 made that joined B30.2 domains with Zn-fingers, but our analyses showed that the B30.2-
411 encoding exons instead belonged to a neighbouring NLR gene, rather than the more distant
412 Zn-finger encoding exons. A possible explanation for the mis-annotation is that the
413 predictions created apparent Fintrim genes. Fintrim proteins are composed of multiple Zn-
414 fingers combined with a B30.2 domain and are assumed to act as sensors for immune stimuli
415 (van der Aa et al. 2012). They are often found in the vicinity of, or at least on the same
416 chromosome as, genes that are located in the MHC in mammals (MHC-associated genes are
417 found not in one cluster in the zebrafish, but on four different chromosomes [3, 8, 16, 19]).
418 We therefore analysed the distribution of the NLR-B30.2 genes relative to the location of
419 fintrims and multiple Zn-finger encoding genes.

420 To establish a list of fintrim genes from the current assembly, we collected and refined
421 candidate genes from the Zv9 genes set, resulting in 61 TRIM, 40 ftr and 18 btr genes (online
422 supplementary material). An alignment of 283 B30.2 domains from the NLR-B30.2 genes
423 and 117 from the fintrim and btr families shows that most of the B30.2 domains from the
424 NLR-B30.2 genes are more closely related to each other than to those of the TRIM families.
425 We found no close association in the genome between the fintrim and the NLR-B30.2 genes
426 (Fig. 6B), except for two cases where a single fintrim gene is found near an NLR-B30.2 gene
427 cluster (chr1 and chr15). If anything, fintrim genes are excluded from regions where NLR-
428 B30.2 genes are found and vice versa.

429 However, as noted above, genes encoding multiple Zn-fingers were interspersed among
430 the NLR-B30.2 genes. Unlike the TRIMs, which contain C3HC4 (RING) and Znf-B-box
431 domains (IPR001841 and IPR000315, respectively), the genes on chromosome 4 encoded yet
432 uncharacterised proteins consisting exclusively of tandem repeats of Zn-fingers of the
433 classical C2H2 type (IPR007087). To investigate this new gene family further, we collected
434 all gene models encoding Zn-fingers of this type. A total of 1259 gene models were found,
435 with the number of repeats per gene ranging from 1 to 36. Some of the cases of genes with
436 extremely large numbers of C2H2 domains (e.g. ENSDARP00000109314) may be mis-

437 annotations that combine adjacent, shorter genes, which we did not analyse manually in
438 further detail.

439 The encoded proteins with small numbers of Zn-fingers included many known proteins,
440 including the Sna and Opa transcription factor families. These were broadly distributed in the
441 genome and largely excluded from chromosome 4 (Fig. 6). By contrast, those with larger
442 numbers of Zn-finger domains are progressively clustered in restricted regions of the
443 genome. For example the majority (66%) of genes with more than 10 C2H2 domains are
444 found on the right arm of chromosome 4, where they are interspersed in an irregular pattern
445 among the NLR-B30.2 genes. Outside chromosome 4, some multi-Zn-finger genes co-locate
446 with subsets of the TRIM genes, for example on chromosomes 3, 16 and 19, while others are
447 located in regions where neither NLR-B30.2 genes nor TRIMs are found. Similar to the
448 NLR-B30.2 genes, multi-Zn-finger genes outside of chromosome 4 tend to be close to
449 chromosomal ends (62% of genes within 15%). On chromosome 4, 8% of the multi-Zn-
450 finger genes are found within 15% from the chromosome ends.

451 In summary, the local duplications that may have led to the expansion of the NLR-
452 B30.2 genes on chromosome 4 may also have duplicated the multi-Zn-finger genes, which
453 have subsequently been transposed to other chromosomes.

454

455

456 **DISCUSSION**

457

458 *Phylogeny of vertebrate NLR proteins*

459 The family of NLR-B30.2 genes has been shaped by different genomic and genetic
460 mechanisms throughout evolution. These include repeated gene amplifications, shuffling of
461 exons and gene fusions, gene conversion and positive selection for diversity. We see low
462 rates of synonymous substitutions in the NACHT domain when comparing the members
463 within each group, and similarly low rates of synonymous substitutions for the B30.2
464 domains in comparisons across all groups. A low rate of synonymous sequence substitutions
465 can be interpreted as a sign of recent gene duplication. If we apply this interpretation using
466 the low divergence of the B30.2 domain, then we would have to conclude that the entire set
467 of genes in groups 1, 2a and 3a is the product of recent duplications. However, this is not
468 consistent with the significant divergence of the NACHT domains between the groups, the
469 different tree topologies of the two domains, nor with our finding that the split into NLR

470 families occurred before the divergence of the zebrafish and the carp. Therefore there must be
471 an alternative explanation. The pattern of synonymous divergence of the two domains
472 between groups is most parsimoniously explained by ongoing gene conversion in the
473 NACHT-B30.2 gene family, with conversion in the NACHT domain acting only within each
474 group, whereas the B30.2 domains show signs of gene conversion across the whole set of
475 genes of groups 1, 2a and 3a, i.e. also between members of different groups.

476 Gene conversion is not uncommon in gene families involved in immunity (see Pasquier
477 2006 for review). It can create diversity, for example in antibodies (Wysocki and Geftter
478 1989) or in the MHC (reviewed in Martinsohn et al., 1999), but it can also homogenize
479 genes, e.g. in the T cell receptor family (Jouvin-Marche, Heller, and Rudikoff 1986). In the
480 NLR-B30.2 both mechanisms may operate.

481 Gene conversion in the NACHT domain appears to be restricted to conversion within
482 groups, keeping the groups (a) homogeneous and (b) distinct from each other. In contrast,
483 gene conversion between B30.2 domains may have another effect, namely to create
484 additional variation. The high dN/dS values indicate positive selection for non-synonymous
485 variants in residues potentially involved in pathogen recognition. As noted by (van der Aa et
486 al. 2009) for fintrims, and also seen in the NLR-B30.2 proteins, the substitutions are
487 concentrated in regions of the B30.2 domain that are exposed on the surface and likely
488 involved in pathogen interactions. Once substitutions have been introduced in one of the
489 genes, gene conversion can then spread these throughout the family. If conversion is frequent
490 and the conversion tracts are short, i.e. cover only a small part of the B30.2 exon, then it
491 would recombine individual amino acid replacements occurring in different parts of the
492 domains and in different paralogues. The process would create novel combinations of amino
493 acid substitutions, and these variants would then again form the substrate for further
494 selection. At the same time, since gene conversion in the B30.2 domain acts across groups,
495 this mechanism also ensures that new recognition modules can spread beyond the group in
496 which they first arose. This can prevent the groups being locked in on a defined subset of
497 B30.2 domains. It is striking that the three groups of genes that show gene conversion in the
498 B30.2 domain are all localised on chromosome 4, whereas group 3b, which has diverged
499 from group 3a in its B30.2 domain, is located on chromosome 1.

500 The oldest NLR genes appear to be those encoding the ancestors of two conserved
501 NLRs, Apaf1 and NWD1, which we find in all animal lineages. These proteins have not been
502 reported to have immune functions. Apaf1, originally discovered as CED-4 in *C. elegans*, is

503 an ancient regulator of apoptosis. The so far only function reported for NWD1, first identified
504 in the zebrafish genome as NACHTP1 (Stein et al. 2007), is its involvement in androgen
505 signalling in the context of prostate cancer (Correa et al. 2014). It will be interesting to learn
506 whether this is a special case of a more general immune function yet to be discovered, or
507 whether, like Apaf1, this old gene does not have immune functions. The other conserved
508 genes first appear at the base of the gnathostomes, and all have roles in immunity or
509 inflammation - whether as transcription factors or as inflammasome components.

510 In parallel, NLR genes have duplicated and often undergone extensive species-specific
511 expansions throughout evolution. This is the case, for example, for the members of the
512 Nalp/NLRP family in the mouse and the NLR-B30.2 family we discuss here. The largest of
513 the known early expansions were in the sponge *A. queenslandica*, the sea urchin *S.*
514 *purpuratus* and the lancelet *B. floridae*, with about 120, 92, and 118 genes respectively
515 (Yuen, Bayes, and Degnan 2014; Huang et al. 2008). As more genomes are sequenced it is
516 likely that additional NLR expansions will be discovered. In vertebrates, the largest
517 expansions are those of the NLR-B30.2 family, although we also find other NLR gene
518 families, for example in the elephant shark *C. milii* (Supplementary Fig. 7, online
519 supplementary material).

520 The expansion of the families argues in favour of their involvement in immunity or
521 broader stress reactions, as seen in numerous other examples of expanded gene families.
522 Expansions can increase the amount of gene product, for example to adapt to stressful
523 environmental conditions (Kondrashov et al. 2002; Kondrashov 2012), as in the cold
524 adaptation in several gene families expressed in Antarctic notothenioid fishes (Chen et al.
525 2008). Expansions can also allow the creation of the variety of sequences that is needed for
526 immune recognition, as in the case of antibodies and T-cell receptors, or the more recent
527 example of the VLR genes in lampreys and hagfish (Li et al. 2013).

528 One possible scenario for the creation of the current NLR-B30.2 gene family in the
529 zebrafish is as follows. Early in the fish lineage, after their initial creation through the fusion
530 of the NLR and B30.2 components, the NLR-B30.2 genes underwent duplications, similar to
531 many other NLR genes in other lineages (Hamada et al. 2012; Yuen, Bayes, and Degnan
532 2014). At this point, the available data do not allow us to trace these earliest duplications.

533 In the Clupeocephala, the paralogues then diversified into groups. Whether the
534 common ancestor had four genes (or a similarly small number), or whether each of the four
535 genes had already begun to be duplicated to form small families is not clear. At this point,

536 gene conversion may already have been occurring and if the early prototypes had already
537 amplified into gene families in the common ancestor, then gene conversion may have acted
538 within each group. What seems clear is that if gene conversion ever acted across the whole
539 gene, conversion of the NACHT domain between groups must have stopped when the groups
540 became too divergent. It cannot have acted between the NACHT domain encoding exons of
541 different groups, as the differences between the groups have been maintained and are still
542 visible now. Since not all currently extant fish have representatives of all four groups, it may
543 be that either whole sets of these genes can be easily lost, or else that the common precursor
544 had only one gene from each group, and that not all lineages inherited all four prototypes.
545 The near-identity of some of the genes we find in the zebrafish (difference between
546 paralogues lower than rate of polymorphism) shows that duplications continue to occur.

547 It is worth speculating about the functional and selective forces that prevent sequence
548 homogenization between the NACHT domains of different groups. If the proteins form large
549 multimeric complexes, as the known inflammasome NLRs do, then their efficient functioning
550 might require that only proteins from the same group can multimerise, for instance to elicit
551 distinct down-stream signaling events. This is supported by the groups featuring different N-
552 terminal domains. A mixed multimer may not be able to assemble a functional N-terminal
553 effector complex.

554 The C-terminal domains - LRRs and B30.2 - do not show the same clear subdivision
555 into families as the N-terminal and the NACHT domains, and homogenizing gene conversion
556 must therefore have affected only part of each gene, or affected different parts differently.
557 This is not without precedent, since gene conversion often proceeds across DNA segments of
558 limited length (see Innan 2009 for review) and parts of a gene can escape sequence
559 homogenization (Innan 2009; Teshima 2004).

560 Both LRRs and B30.2 domains have been implicated in recognition of pathogen- or
561 danger-associated molecular patterns. The B30.2 domain of Trim5a binds to HIV-1 and is
562 involved in blocking HIV-1 proliferation in monkeys (Stremlau et al. 2004). In Nod1 and
563 Nod2, the LRRs recognise components of pathogens (Inohara et al. 2005), including flagellin
564 monomers (Akira, Uematsu, and Takeuchi 2006), while in Toll-like receptor proteins they
565 recognise viral dsRNA (Kawai and Akira 2009).

566 The sequences of the LRRs in the NLR-B30.2 genes are not particularly variable, and it
567 therefore seems unlikely that they have a role in specific ligand-recognition. The B30.2
568 domains however show significant amino acid variation between the members. It may

569 therefore have the same function as the related B30.2 domain in the fintrim genes, which has
570 been suggested to be under positive selection to allow variation in specificity for pathogen
571 recognition. It is conceivable that the acquisition of the B30.2 domain and the option to use it
572 for specific recognition of a wide range of pathogens drove the amplification of these genes.

573 While we lack sufficient genomic data from salt-water fishes, we are tempted to
574 speculate that the massive inflation of the NLR-B30.2 group may have occurred along with
575 the adaptation to fresh water environments when ancestors of the zebrafish encountered a
576 new pathogen fauna. Alternatively, the NLR-B30.2 system may functionally complement the
577 innate immune system during the first few weeks of life of the zebrafish larva: the larva is
578 exposed to the outside world and starts eating after two days of development, but a functional
579 adaptive immune systems arises only after three to five weeks (Lam et al. 2004). We have
580 not investigated whether the presence of NLR-B30.2 expansions in a fish species correlates
581 with the time of development of the adaptive immune system in that species.

582

583 *Shuffling between genes and creation of new genes*

584 A mechanism involved in the initial creation of the NLR-B30.2 family appears to have
585 been exon shuffling, both within the family and between the NLR genes and other gene
586 families. For example, the N-terminal peptide repeats occur in several variants, but a given
587 variant is not strictly associated with any particular group: at least two of the variants are
588 found both in association with group 2a and group 3.

589 We also find evidence for recombination with other immune genes. The B30.2 domain
590 of the NLR-B30.2 proteins most closely resembles that of the fintrim proteins, a fish-specific
591 gene family for which the origin of the fusion between the Zn-fingers with the B30.2 domain
592 is not known (van der Aa et al. 2009), suggesting that exon shuffling occurred during the
593 generation of the ancestral genes of the NLR-B30.2 and the fintrim gene families.

594 Apart from this possible case of exon-exchange, the relationship between the three
595 large and partially related families – the NLR-B30.2 genes, the fintrim genes and the multi-
596 Zn-finger genes we describe here – are unclear. While it is striking that the fintrims share the
597 B30.2 domain with the NLR-B30.2 genes and the Zn-fingers with the multi-Zn-finger
598 proteins, they do not preferentially map to the same regions of the genome, and the Zn-finger
599 is of a different type. By contrast, the multi-Zn-finger genes are mostly found on
600 chromosome 4, interspersed between the NLR-B30.2 genes. We have not attempted to trace
601 the relative evolutionary histories of these three families.

602 A further gene that may have arisen from domain shuffling between these gene families
603 is the human gene encoding pyrin (marenostriin/MEFV). Pyrin is a protein that is composed
604 of an N-terminal PYD domain, for which the best match in the zebrafish is the PYD domain
605 in the group 1 NLR-proteins. The C-terminal part of pyrin contains a Zinc-finger and a B30.2
606 domain, which resembles the zebrafish fintrim proteins of the btr family. The most likely
607 interpretation for the origin of this gene, which must have arisen at the base of the tetrapods,
608 is therefore a recombination between an NLR gene and a neighbouring fintrim gene.

609

610 *Chromosome 4*

611 The zebrafish chromosome 4 has unusual properties. Its long arm is entirely
612 heterochromatic, replicates late and shows a reduced recombination rate. It contains an
613 accumulation of 5S rRNA, snRNA, tRNA and mir-430 clusters (Anderson et al. 2012; Howe
614 et al. 2013), as well as the expanded protein coding gene here.

615 Chromosome 4 was recently shown to function as the sex chromosome in wild
616 zebrafish ZW/WW sex determination, with the sex determining signal being located near the
617 right end chromosome 4 (Wilson et al. 2014). The sex determination region in the grass carp
618 may also be associated with NACHT domain encoding genes (Y. Wang et al. 2015). This
619 was concluded from the comparison of the genome sequences of one male and one female
620 carp, where those regions present in the male and absent in the female were interpreted as
621 sex-determining. In addition to the NACHT domain genes, this region also included other
622 immunity genes, such as the immunoglobulin V-set, ABC transporters, and proteasome
623 subunits. While the co-location between sex determination and immune signalling molecules
624 we describe here may support this conclusion, it is of course equally possible that the finding
625 in the grass carp is simply caused by allelic diversity in these highly variable genes between
626 the two individuals. It is nevertheless intriguing that two fast evolving genetic systems are
627 located in such close proximity in zebrafish. Perhaps, after an initial round of NLR gene
628 duplications, a run-away evolutionary process of further amplification to created the present
629 chromosome 4, which is now a hotspot for rapid evolutionary processes.

630

631 **METHODS**

632

633 *Re-annotation of NLR genes in the zebrafish genome*

634 To establish a complete list of all genes encoding NLR proteins in the zebrafish
635 genome, we first conducted a search of the Zv9 genome assembly for sequences that encoded
636 the characteristic protein domains, using a combination of approaches. We constructed a
637 hidden Markov model (HMM) for the Fisna domain and used this together with the HMM for
638 the NACHT domain obtained from PFAM to search the Zv9 assembly with hmmsearch
639 (hmmer.janelia.org/search/hmmsearch), resulting in 297 Fisna and 328 NACHT locations
640 (see online supplementary material). As an alternative way to identify NACHT domains
641 specific for the novel NLRs, we ran electronic PCRs (PMID: 9149949) with primer sets for a
642 segment stretching from the C-domain into the winged helix domain that we had used for
643 experimental analysis of the genes (unpublished work). Each set of primers was specific for
644 one of the NLR groups (Supplementary Methods). This resulted in 321 hits. To find regions
645 in the genome encoding B30.2 domains we conducted a TBLASTN search, which yielded
646 503 hits. As B30.2 domains also occur in other large, immune-related protein families (see
647 below), such a high number of domains was consistent with expectations.

648 Secondly, we collected all Ensembl genes overlapping the above motifs (487 predicted
649 genes) and also all manually annotated genes (vega.sanger.ac.uk) that had been marked as
650 NLR or as containing a NACHT domain during manual annotations in the past (307
651 predicted genes).

652 The collection was purged of gene models that did not match the criteria for being
653 novel NLRs, excluding e.g. the B30.2 domain-containing fintrim genes. Sixteen NACHT
654 domain proteins in the combined list do not belong to the group of novel fish NLRs since
655 they do not contain the Fisna-domain, and the sequence surrounding their WalkerA motifs
656 does not match the one typical for the novel NLRs. They include the seven conserved NLRs
657 that are orthologous across all vertebrate species (Nod1, Nod2, Nlrc3, Nlr1, CIITA, Apaf1,
658 NWD1/NachtP1), and nine further proteins with an NLR structure. (Table 1 and online
659 supplementary material)

660 Comparison of the purged gene sets with the genomic regions that encoded parts of
661 NLR proteins showed that many genes in this family had been annotated incorrectly, and for
662 others there were no predictions at all, probably due to the repetitive nature of this gene
663 family and the limited availability of supporting evidence in the form of cDNAs.

664 The regions containing the sequences identified in our searches were therefore re-
665 annotated manually, correcting and adding gene models to create full-length genes. This re-
666 annotation had to be restricted to regions located on finished sequence, since whole genome
667 shotgun (WGS) contigs in Zv9 were not accessible to manual annotation. For these contigs,
668 the automated Ensembl gene models were retained in their original form, recognizable in our
669 final list by their “ENSDARG” identifier (Supplementary Excel Table1). The resulting
670 protein sequences were then aligned using Clustal-Omega (Sievers et al. 2011) or Muscle
671 (Edgar 2004) and compared. Sequences that appeared truncated were analysed further by
672 searching for additional exons to complete them, until, in an iterative process, we had
673 optimised them. Some sequences remained incomplete, either because they were located next
674 to sequence gaps, or because no additional exons could be detected. In these cases it is not
675 known whether the truncation of the gene is a true biological event caused by recent
676 recombination, or whether it is due to a mis-assembly of the genome sequence.

677 The optimised gene set was combined with the gene predictions in Ensembl (Methods,
678 hand-filtered alignments and location checks of the remaining genes to identify accordance).
679 The final list of novel zebrafish NLR proteins contains 368 members (Supplementary Excel
680 Table1). A further 36 predictions for NLR genes had been annotated as pseudogenes and
681 were therefore not retrieved for this list (online supplementary material). The refined genes
682 have since been integrated into the VEGA and Ensembl gene sets, however since the
683 annotation was performed on pre-GRCz10 paths, the latest GRCz10 gene set (Ensembl80)
684 might differ marginally from the described results.

685

686 Conserved NLR genes across metazoa

687 We used the zebrafish gene identifiers for the conserved NLRs in zebrafish to query the
688 Orthoinspector 2.0 database (Linard et al. 2014) at <http://lbg.iqbmc.fr/orthoinspector> for
689 orthologs in published genomes and downloaded the corresponding sequence. We then
690 queried a custom Blast database of the *Cyprinus carpio* proteome, as well as the NCBI nr
691 database for selected fish species using BLASTP. After removing redundant hits we
692 calculated alignments employing Clustal-Omega v.1.2 (Sievers et al. 2011) and subsequently
693 removed sequences of poor quality. In a second inference we also used trimal (Capella-
694 Gutiérrez, Silla-Martínez, and Gabaldón 2009) to reduce the alignment to the conserved
695 residues. We employed protpst v.3.2 (Darriba et al. 2011) to infer the best fitting
696 evolutionary model and found that the LG model with Gamma optimisation performed best

697 under the AKIKE information criterion. We then ran RAxML v7.7.2 (Stamatakis 2006) on
698 both alignments on the Cologne University CHEOPS super computer and calculated
699 bootstrap values. Phylogenetic trees were visualised and edited in Dendroscope v.3.2.5
700 (Huson et al. 2007).

701

702 *figmop and tblastn screen for NLR-B30.2 candidates in other fish genomes*

703 Expanded gene families are not well annotated in most genomes. Rather than relying
704 on gene predictions for identifying NLR-B30.2 genes, we therefore directly searched the
705 genome sequences of 6 species: *Latimeria chalumnae*, *Lepisosteus oculatus*, *Callorhinchus*
706 *milii*, *Esox lucius*, *Astyanax mexicanus*, *Cyprinus carpio*. We downloaded genome data either
707 from NCBI servers or the genome project websites. We then used the Figmop (Curran,
708 Gilleard, and Wasmuth 2014) pipeline to find contigs and scaffolds in the genomes with
709 NLR-B30.2 candidates on them. The Figmop pipeline builds a profile of conserved motifs
710 from a starting set of sequences and uses these to search a target database with the MEME
711 software suite (Bailey et al. 2009).

712 We used zebrafish NLR-B30.2 sequences from all four groups to create a set of 15
713 motifs to search the above genomes. The resulting contigs were then subjected to the
714 Augustus (v3.0.3) gene prediction pipeline (Stanke and Waack 2003) to predict genes *de*
715 *novo*, setting zebrafish as the "species". We complemented this approach by TBLASTN
716 searches using the NACHT as well as the B30.2 domains as queries in individual searches
717 and then kept those predictions in which the domains occurred in the proper order (thereby
718 excluding spurious cases caused by mis-assembly or incomplete genes).

719

720 *Phylogenetic analyses of the NLR-B30.2 groups*

721 We used a recursive approach for identifying genes for the phylogeny that were
722 representative of the overall sequence divergence in the gene family. We selected only those
723 that had both the NACHT and B30.2 domains. We then recursively performed the following:
724 (1) constructed a sequence alignment of ~500 residues (starting with the NACHT domain) in
725 the dataset using Clustal-Omega, (2) constructed a phylogeny using a Bayesian approach
726 using MrBayes with mcmc=1,000,000, sump burnin=1000 and sumt burnin=1000 and (3)
727 removed monophyletic paralogs from the dataset. The recursive analysis was halted when
728 no instances of paralogous sister-sequences remained (Fig. 3), with the exception that at least
729 one zebrafish and carp sequence from each of the major groups was retained. Once the final

730 dataset of sequences was determined we removed gap-containing and highly variable
731 columns from the alignment and re-ran MrBayes with mcmc=2,000,000, sump burnin=2000
732 and sumt burnin=2000 and re-confirmed our inferred tree with maximum likelihood in
733 RAxML.

734 We also used RAxML to infer a phylogeny of all currently available *D. rerio* and *C.*
735 *arpio* NLR-B30.2 genes. As described for the conserved NLR genes above we based our
736 phylogeny on an alignment calculated with Clustal-Omega v.1.2, reduced to conserved
737 regions with trimal, and model testing with proptest (JTT+G+F model found to be optimal).

738

739 Divergence analysis

740 For zebrafish genes in groups 1, 2a, 3a and 3b we calculated all pairwise dNdS values
741 for NACHT domain containing exon and the B30.2 domain independently using the KaKs
742 calculator (D. Wang et al. 2010): We extracted the respective regions from our protein
743 alignment, then used tranalign (Rice, Longden, and Bleasby 2000) to create DNA alignments
744 from these proteins and cds. We then calculated all pairwise comparisons and used paraAT
745 (Zhang et al. 2012) and submitted the resulting alignments to the KaKs calculator
746 independently estimating under the MYN model and the model averaging option with aid of
747 the gnu-parallel tool (Tange 2011). We then used our own iPython (Pérez and Granger 2007)
748 script to sort data and calculated means, medians, and errors in the R statistic software (R-
749 Core-Team 2015).

750 We also used the tranaligned regions to calculate independent phylogenies for the
751 NACHT and B30.2 exons with RAxML. We loaded the inferred trees into Dendroscope and
752 employed this software to visualise connection between branches belonging to the same gene
753 in both trees.

754

755 **ACKNOWLEDGEMENTS**

756

757 We thank Robert Remy and Giuliano Crispantu for some early work on this project and
758 Jonathan Wood for verifying genomic locations of assembly components. Financial support
759 was provided by EMBO and the DFG SFB 670 "Zellautonome Immunität" to ML, DFG SFB
760 680 "Molecular basis of evolutionary innovation" to TW, the HHMI International Early
761 Career Scientist Program (55007424), MINECO (Sev-2012-0208), AGAUR program (2014
762 SGR 0974), and an ERC Starting Grant (335980_EinME) to FK, the European Molecular

763 Biology Laboratory to JM, the Wellcome Trust to KH (zebrafish genome sequencing project)
764 and the National Human Genome Research Institute (NHGRI) grant HG002659 to GKL
765 (gene annotation), and a grant from the Volkswagen Foundation to PHS. We thank the
766 CHEOPS support team and the Bundesland Nordrhein Westfalen for making HPC
767 applications freely available at the University of Cologne.

768 **Figure legends**

769

770 **Fig. 1:** Structure of the fish-specific NLR-B30.2 protein family members.

771 A. Alignment overview of 288 proteins for which full length predictions are available.
772 Regions with long insertions in a few of the genes that had resulted in the introduction
773 of long gaps in the alignment were deleted for the sake of clarity and simplicity. Gaps
774 were manually introduced to highlight intron/exon boundaries, except between the C-
775 terminal extensively duplicated LRRs in group 2b and the extensively duplicated N-
776 terminal repeats in groups 2b and 3b. The colour coding is a random assignment of
777 colours to amino acid created in Jalview.

778 A gap was inserted between the LRRs 6 and 7 of groups 2b and 3b to allow the
779 conserved C-terminal LRR and the B30.2 domains of groups 1, 2a and 3a to be
780 positioned immediately after the 1- 6 LRRs in these groups. An alignment of the full
781 set of predictions is provided in the online supplementary material.

782 B. A schematic representation of the protein domains in each group, on the same scale as
783 in the alignment above. Each box represents an exon. Please refer to Supplementary
784 Fig. 2 and the alignment file (online supplementary material) for the many details not
785 captured in this simplified Fig..
786

787

788

788 **Fig. 2:** The Fisna domain and its homologs in mammalian proteins.

789 A. Alignment of sequences identified in mammalian genomes using an HMM search
790 with PF14484 and examples of zebrafish Fisna domains.

791 B. Secondary structure predictions for selected examples.
792

793

793 **Fig. 3:** Relationships between NLR-B30.2 genes in different fishes.

794 Phylogenetic tree resulting from a recursive phylogenetic analysis. Inflations of groups in
795 zebrafish are indicated in yellow with numbers of genes per group displayed. Bootstrap
796 values are given above each branch where higher than 50% and pp values (from MrBayes)
797 below each branch, where they are higher than 50% and the topology is congruent.
798

799

799 **Fig. 4:** Evolutionary history of NLR genes

800 A reduced dendrogram of Metazoa based on the NCBI taxonomy database displaying key
801 events in the evolution of NLR genes as described in the main text. See Supplementary Fig. 7
802 and the supplementary online material for phylograms.
803

804

805

805 **Fig. 5:** Location of the NLR genes in the genome in assemblies Zv9 and GRCz10.

806 A. The chromosomes containing NLR-B30.2 genes are shown in the outer circle (note
807 that corrections of the genome between Zv9 and the GRCz10 have changed the
808 lengths of some of the chromosomes. The genes were annotated on Zv9 and lifted
809 over to the GRCz10 path where possible as the GRCz10 gene set did not become
810 available until May 2015. The members of the four groups of NLR genes are shown
811 as radial lines within the circles with group 1 in the outermost and group 4 in the inner
812 ring. Each gene is connected by a black line to its most closely related paralog, based
813 on the number of amino acid substitutions per site calculated in MEGA5 (Poisson
814 correction model). Most genes are most closely related to a near neighbour. The
815 changes in the assembly have have lead to many genes that were closely related but
816 resided on different chromosomes in Zv9 being located in closer proximity in
817 GRCz10.

818 B. Normalised location of NLR-B30.2 genes on chromosomes. Each chromosome is
819 shown as a horizontal line of 100% length, and the NLR genes are plotted at their
820 relative positions along the chromosome. Apart from the genes on chromosome 4
821 (marked in blue), all other genes are found within the first or last quarter of the
822 chromosome.

823

824 **Fig. 6:** Genomic positions of genes for multi-Zn-finger proteins, Fintrims and NLR-B30.2
825 proteins.

826 A. Frequency of genes encoding proteins with the indicated number of Zn-finger
827 domains. Total number of Zn-finger encoding gene predictions: 1102.

828 B. The genomic locations of Zn-finger encoding genes is plotted on a circos diagram
829 representing the *Zv9* assembly for five subgroups defined by the number of Zn-finger
830 domains. Inner circles (light to dark blue): 1-3, 4-5, 6-7, 8-9, 10-12 and more than 12
831 domains; Outer circle (red): NLR genes

832 C. Spatial distribution of NLR genes, TRIM genes and genes with at least 10 Zn-finger
833 domains in the GRCz10 assembly (for gene liftover see Fig. 5).

834

835

836 **Supplementary Fig. Legends**

837

838 **SFig 1:** Overview of the entire set of 368 predicted NLR-B30.2 proteins in the zebrafish,
839 based on a Clustal Omega alignment. The original alignment (online supplementary material)
840 was edited by hand to improve the alignments of the N-terminal repeats and the LRR. The
841 colour code for the amino acids was assigned at random. Gaps were introduced in the
842 alignment at the positions of introns (marked by a grey arrowhead below the alignment).
843 Further large gaps are created because some positions are prone to variable and often long
844 insertions or internal duplications (marked by red asterisks below the alignment).
845 The domains are marked below the sequences; domain boundaries within the NACHT
846 domain are entered according to reference (Proell et al. 2008).
847 Positions of introns are marked by grey arrowheads below the sequences.

848

849 **SFig 2:** Comparison of FISNA and NACHT domains in groups 1 – 4 in the NLR-B30.2
850 genes.

851 Logos generated in Jalview from the alignment in the online supplementary material.

852 Differences in the Walker A motif are diagnostic for groups.

853

854 **SFig 3:** Comparison of LRRs in groups 1 – 3 and mouse NLRP3 LRRs.

855 Logos generated in Jalview from the alignment in online supplementary material.

856

857 **SFig 4:** Differences in the B30.2 domain between the indicated groups.

858 Logos generated in Jalview from the alignment in online supplementary material.

859

860 **SFig 5:** Boxplots of dN/dS values for the Fisna and NACHT encoding exons and the exons
861 coding for the B30.3 domain. Median values are indicated by a black bar, means are
862 displayed as a diamond. The y-axis is log-scaled.

863

864 **SFig 6:** Phylogenetic maximum likelihood tree of the NLR-B30.2 proteins in *D. rerio* and *C.*
865 *carpio*.

866

867 **SFig 7:** Phylogenetic maximum likelihood tree of the conserved NLR proteins in Metazoa.

References

- Akira, Shizuo, Satoshi Uematsu, and Osamu Takeuchi. 2006. "Pathogen Recognition and Innate Immunity." *Cell* 124 (4): 783–801. doi:10.1016/j.cell.2006.02.015.
- Anderson, Jennifer L, Adriana Rodríguez Marí, Ingo Braasch, Angel Amores, Paul Hohenlohe, Peter Batzel, and John H Postlethwait. 2012. "Multiple Sex-Associated Regions and a Putative Sex Chromosome in Zebrafish Revealed by RAD Mapping and Population Genomics." Edited by Laszlo Orban. *PLoS ONE* 7 (7): e40701. doi:10.1371/journal.pone.0040701.
- Bailey, Timothy L, Mikael Boden, Fabian A Buske, Martin Frith, Charles E Grant, Luca Clementi, Jingyuan Ren, Wilfred W Li, and William S Noble. 2009. "MEME Suite: Tools for Motif Discovery and Searching." *Nucleic Acids Research* 37: gkp335–W208. doi:10.1093/nar/gkp335.
- Bonardi, Vera, Karen Cherkis, Marc T Nishimura, and Jeffery L Dangl. 2012. "A New Eye on NLR Proteins: Focused on Clarity or Diffused by Complexity?." *Current Opinion in Immunology* 24 (1): 41–50. doi:10.1016/j.coi.2011.12.006.
- Capella-Gutiérrez, Salvador, José M Silla-Martínez, and Toni Gabaldón. 2009. "trimAl: a Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses." *Bioinformatics* 25 (15): 1972–1973. doi:10.1093/bioinformatics/btp348.
- Chen, Zuo Zhou, C-H Christina Cheng, Junfang Zhang, Lixue Cao, Lei Chen, Longhai Zhou, Yudong Jin, et al. 2008. "Transcriptomic and Genomic Evolution Under Constant Cold in Antarctic Notothenioid Fish." *Proceedings of the National Academy of Sciences* 105 (35) (September 2): 12944–12949. doi:10.1073/pnas.0802432105.
- Correa, Ricardo G, Maryla Krajewska, Carl F Ware, Motti Gerlic, and John C Reed. 2014. "The NLR-Related Protein NWD1 Is Associated with Prostate Cancer and Modulates Androgen Receptor Signaling." *Oncotarget* 5 (6): 1666–1682.
- Curran, David M, John S Gilleard, and James D Wasmuth. 2014. "Figmap: a Profile HMM to Identify Genes and Bypass Troublesome Gene Models in Draft Genomes." *Bioinformatics* 30 (22) (November 15): 3266–3267. doi:10.1093/bioinformatics/btu544.
- Darriba, Diego, Guillermo L Taboada, Ramón Doallo, and David Posada. 2011. "ProtTest 3: Fast Selection of Best-Fit Models of Protein Evolution." *Bioinformatics* 27 (8): 1164–1165. doi:10.1093/bioinformatics/btr088.
- Edgar, Robert C. 2004. "MUSCLE: a Multiple Sequence Alignment Method with Reduced Time and Space Complexity." *BMC Bioinformatics* 5 (1): 113. doi:10.1186/1471-2105-5-113.
- Goody, Michelle F, Con Sullivan, and Carol H Kim. 2014. "Studying the Immune Response to Human Viral Infections Using Zebrafish." *Developmental and Comparative Immunology* 46 (1): 84–95. doi:10.1016/j.dci.2014.03.025.
- Hamada, Mayuko, Eiichi Shoguchi, Chuya Shinzato, Takeshi Kawashima, David J Miller, and Nori Satoh. 2012. "The Complex NOD-Like Receptor Repertoire of the Coral *Acropora Digitifera* Includes Novel Domain Combinations." *Molecular Biology and Evolution* 30 (1) (August 30): mss213–176. doi:10.1093/molbev/mss213.
- Howe, Kerstin, and Jonathan MD Wood. 2015. "Using Optical Mapping Data for the Improvement of Vertebrate Genome Assemblies." *GigaScience* 4 (1): 10. doi:10.1186/s13742-015-0052-y.
- Howe, Kerstin, Matthew D Clark, Carlos F Torroja, James Torrance, Camille Berthelot, Matthieu Muffato, John E Collins, et al. 2013. "The Zebrafish Reference Genome Sequence and Its Relationship to the Human Genome." *Nature* 496 (7446): 498–503. doi:10.1038/nature12111.
- Huang, Shengfeng, Shaochun Yuan, Lei Guo, Yanhong Yu, Jun Li, Tao Wu, Tong Liu, et al.

2008. “Genomic Analysis of the Immune Gene Repertoire of *Amphioxus* Reveals Extraordinary Innate Complexity and Diversity.” *Genome Research* 18 (7): 1112–1126. doi:10.1101/gr.069674.107.
- Huson, Daniel H, Daniel C Richter, Christian Rausch, Tobias DeZulian, Markus Franz, and Regula Rupp. 2007. “Dendroscope: an Interactive Viewer for Large Phylogenetic Trees.” *BMC Bioinformatics* 8 (1): 460. doi:10.1186/1471-2105-8-460.
- Inohara, Naohiro, Mathias Chamaillard, Christine McDonald, and Gabriel Nunez. 2005. “NOD-LRR PROTEINS: Role in Host-Microbial Interactions and Inflammatory Disease.” *Annual Reviews in Biochemistry* 74 (1): 355–383. doi:10.1146/annurev.biochem.74.082803.133347.
- Jouvin-Marche, E, M Heller, and S Rudikoff. 1986. “Gene Correction in the Evolution of the T Cell Receptor Beta Chain.” *The Journal of Experimental Medicine* 164 (6) (December 1): 2083–2088.
- Kawai, Taro, and Shizuo Akira. 2009. “The Roles of TLRs, RLRs and NLRs in Pathogen Recognition.” *International Immunology* 21 (4): 317–337. doi:10.1093/intimm/dxp017.
- Kondrashov, Fyodor A. 2012. “Gene Duplication as a Mechanism of Genomic Adaptation to a Changing Environment.” *Proceedings of the Royal Society B: Biological Sciences* 279 (1749): 5048–5057. doi:10.1098/rspb.2012.1108.
- Kondrashov, Fyodor A, Igor B Rogozin, Yuri I Wolf, and Eugene V Koonin. 2002. “Selection in the Evolution of Gene Duplications.” *Genome Biology* 3 (2):research0008.1–0008.9. doi:10.1093/molbev/msp129
- Koonin, E V, and L Aravind. 2000. “The NACHT Family - a New Group of Predicted NTPases Implicated in Apoptosis and MHC Transcription Activation.” *Trends in Biochemical Sciences* 25 (5): 223–224.
- Kufer, Thomas A, and Philippe J Sansonetti. 2011. “NLR Functions Beyond Pathogen Recognition.” *Nature Immunology* 12 (2): 121–128. doi:10.1038/ni.1985.
- Laing, Kerry J, Maureen K Purcell, James R Winton, and John D Hansen. 2008. “A Genomic View of the NOD-Like Receptor Family in Teleost Fish: Identification of a Novel NLR Subfamily in Zebrafish.” *BMC Evolutionary Biology* 8 (1): 42. doi:10.1186/1471-2148-8-42.
- Lam, S H, Chua, H L, Gong, Z, Lam, T J, & Sin, Y M. 2004. “Development and maturation of the immune system in zebrafish, *Danio rerio*: a gene expression profiling, in situ hybridization and immunological study.” *Developmental and comparative immunology*, 28(1): 9–28. doi:10.1016/S0145-305X(03)00103-4
- Lapraz, François, Eric Röttinger, Véronique Duboc, Ryan Range, Louise Duloquin, Katherine Walton, Shu-Yu Wu, et al. 2006. “RTK and TGF- β Signaling Pathways Genes in the Sea Urchin Genome.” *Developmental Biology* 300 (1): 132–152. doi:10.1016/j.ydbio.2006.08.048.
- Li, Jianxu, Sabyasachi Das, Brantley R Herrin, Masayuki Hirano, and Max D Cooper. 2013. “Definition of a Third VLR Gene in Hagfish.” *Proceedings of the National Academy of Sciences* 110 (37): 15013–15018. doi:10.1073/pnas.1314540110.
- Linard, Benjamin, Alexis Allot, Raphaël Schneider, Can Morel, Raymond Ripp, Marc Bigler, Julie D Thompson, Olivier Poch, and Odile Lecompte. 2014. “OrthoInspector 2.0: Software and Database Updates.” *Bioinformatics*: btu642. doi:10.1093/bioinformatics/btu642.
- Liu, Ying, Yi-Bing Zhang, Ting-Kai Liu, and Jian-Fang Gui. 2013. “Lineage-Specific Expansion of IFIT Gene Family: an Insight Into Coevolution with IFN Gene Family.” Edited by Nicolas Salamin. *PLoS ONE* 8 (6): e66859–14. doi:10.1371/journal.pone.0066859.
- Martinsohn, J T, A B Sousa, L A Guethlein, and J C Howard. 1999. “The Gene Conversion

- Hypothesis of MHC Evolution: a Review.” *Immunogenetics* 50 (3-4) (November): 168–200.
- Pasquier, Louis Du. 2006. “Germline and Somatic Diversification of Immune Recognition Elements in Metazoa.” *Immunology Letters* 104 (1-2): 2–17.
doi:10.1016/j.imlet.2005.11.022.
- Pérez, F, and B E Granger. 2007. “IPython: a System for Interactive Scientific Computing.” *Computing in Science & Engineering* 9 (3): 21–29. doi:10.1109/MCSE.2007.53.
- Proell, Martina, Stefan J Riedl, Jörg H Fritz, Ana M Rojas, and Robert Schwarzenbacher. 2008. “The Nod-Like Receptor (NLR) Family: a Tale of Similarities and Differences.” Edited by Nick Gay. *PLoS ONE* 3 (4): e2119. doi:10.1371/journal.pone.0002119.
- Rast, Jonathan P, and Katherine M Buckley. 2013. “Lamprey Immunity Is Far From Primitive.” *Proceedings of the National Academy of Sciences* 110 (15): 5746–5747. doi:10.1073/pnas.1303541110.
- R Core Team. (2015). R: A Language and Environment for Statistical Computing. Vienna, Austria.
- Rice, P, I Longden, and A Bleasby. 2000. “EMBOSS: the European Molecular Biology Open Software Suite.” *Trends in Genetics* 16 (6): 276–277.
- Rowe, Hannah M, Jeffrey H Withey, and Melody N Neely. 2014. “Zebrafish as a Model for Zoonotic Aquatic Pathogens.” *Developmental and Comparative Immunology* 46 (1): 96–107. doi:10.1016/j.dci.2014.02.014.
- Schuler, G. 1997. “Sequence Mapping by Electronic PCR.” *Genome Research* 5: 541–550.
- Sievers, Fabian, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, et al. 2011. “Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega.” *Molecular Systems Biology* 7: 1–6. doi:10.1038/msb.2011.75.
- Sola, L, and E Gornung. 2001. “Classical and Molecular Cytogenetics of the Zebrafish, *Danio rerio* (Cyprinidae, Cypriniformes): an Overview.” *Genetica* 111 (1-3): 397–412.
- Stamatakis, Alexandros. 2006. “RAxML-VI-HPC: Maximum Likelihood-Based Phylogenetic Analyses with Thousands of Taxa and Mixed Models.” *Bioinformatics* 22 (21): 2688–2690. doi:10.1093/bioinformatics/btl446.
- Stanke, M, and S Waack. 2003. “Gene Prediction with a Hidden Markov Model and a New Intron Submodel.” *Bioinformatics* 19: ii215–ii225. doi:10.1093/bioinformatics/btg1080.
- Stein, Cornelia, Mario Caccamo, Gavin Laird, and Maria Leptin. 2007. “Conservation and Divergence of Gene Families Encoding Components of Innate Immune Response Systems in Zebrafish.” *Genome Biology* 8 (11): R251. doi:10.1186/gb-2007-8-11-r251.
- Stremlau, Matthew, Christopher M Owens, Michel J Perron, Michael Kiessling, Patrick Autissier, and Joseph Sodroski. 2004. “The Cytoplasmic Body Component TRIM5alpha Restricts HIV-1 Infection in Old World Monkeys.” *Nature* 427 (6977): 848–853. doi:10.1038/nature02343.
- Tange, O. 2011. *Gnu Parallel—the Command-Line Power Tool*. The USENIX Magazine.
- Ting, Jenny P Y, Ruth C Lovering, Emad S Alnemri, John Bertin, Jeremy M Boss, Beckley K Davis, Richard A Flavell, et al. 2008. “The NLR Gene Family: a Standard Nomenclature.” *Immunity* 28 (3): 285–287. doi:10.1016/j.immuni.2008.02.005.
- van der Aa, Lieke M, Jean-Pierre Levraud, Malika Yahmi, Emilie Lauret, Valérie Briolat, Philippe Herbomel, Abdenour Benmansour, and Pierre Boudinot. 2009. “A Large New Subset of TRIM Genes Highly Diversified by Duplication and Positive Selection in Teleost Fish.” *BMC Biology* 7 (1): 7. doi:10.1186/1741-7007-7-7.
- van der Aa, Lieke M, Luc Jouneau, Emmanuel Laplantine, Olivier Bouchez, Lidy Van Kemenade, and Pierre Boudinot. 2012. “FinTRIMs, Fish Virus-Inducible Proteins with E3 Ubiquitin Ligase Activity.” *Developmental and Comparative Immunology* 36 (2):

- 433–441. doi:10.1016/j.dci.2011.08.010.
- Wang, Dapeng, Yubin Zhang, Zhang Zhang, Jiang Zhu, and Jun Yu. 2010. “KaKs_Calculator 2.0: a Toolkit Incorporating Gamma-Series Methods and Sliding Window Strategies.” *Genomics, Proteomics & Bioinformatics* 8 (1): 77–80. doi:10.1016/S1672-0229(10)60008-3.
- Wang, Yaping, Ying Lu, Yong Zhang, Zemin Ning, Yan Li, Qiang Zhao, Hengyun Lu, et al. 2015. “The Draft Genome of the Grass Carp (*Ctenopharyngodon Idellus*) Provides Insights Into Its Evolution and Vegetarian Adaptation.” *Nature Genetics* 47 (6): 625–631. doi:10.1038/ng.3280.
- Wilson, Catherine A, Samantha K High, Braedan M McCluskey, Angel Amores, Yi-lin Yan, Tom A Titus, Jennifer L Anderson, et al. 2014. “Wild Sex in Zebrafish: Loss of the Natural Sex Determinant in Domesticated Strains.” *Genetics* 198 (3): 1291–1308. doi:10.1534/genetics.114.169284.
- Wysocki, L J, and M L Geftter. 1989. “Gene Conversion and the Generation of Antibody Diversity.” *Annual Review of Biochemistry* 58 (1): 509–531. doi:10.1146/annurev.bi.58.070189.002453.
- Yuen, Benedict, Joanne M Bayes, and Sandie M Degnan. 2014. “The Characterization of Sponge NLRs Provides Insight Into the Origin and Evolution of This Innate Immune Gene Family in Animals.” *Molecular Biology and Evolution* 31 (1): 106–120. doi:10.1093/molbev/mst174.
- Zhang, Zhang, Jingfa Xiao, Jiayan Wu, Haiyan Zhang, Guiming Liu, Xumin Wang, and Lin Dai. 2012. “ParaAT: a Parallel Tool for Constructing Multiple Protein-Coding DNA Alignments.” *Biochemical and Biophysical Research Communications* 419 (4): 779–781. doi:10.1016/j.bbrc.2012.02.101.

Table 1:

All NLR proteins in the zebrafish genome.

Zf identifier	Protein name
1. NLRs with orthologs in mammals	
ENSDARP00000052748	nod1
ENSDARP00000124380	nod2
ENSDARP00000102939	nlr3 / nod3
ENSDARP00000101928	nlr1
ENSDARP00000099546	NWD1/NACHTP1
ENSDARP00000105957	CIITA
ENSDARP00000105810	Apaf1
2. Other NLRs	
ENSDARP00000118135	NLRP6 (4 of 5)
ENSDARP00000105086	NLRP6 (1 of 5)
ENSDARP00000107209	NLRP6 (3 of 5) nlrb5
ENSDARP00000126513	NLRP6 (2 of 5)
ENSDARP00000104483	NLRP6 (5 of 5)
ENSDARP00000126444	NLRC5
OTTDARP00000028005	-
ENSDARG00000088041 is UniProtKB R4GEV1_DANRE	NLRC3-like
ENSDARG00000087736	-
3. Fish specific NLR multigene family	
Online supplementary material	-

Table 2:

Median dN (a), dS (b), and dN/dS (c) values calculated from all pairwise comparisons in the exons coding for the NACHT domain and the B30.2 domain. For dS values below 0.2 are highlighted and for dN/dS values higher than 1 are highlighted.

Table 2a dN	Fisna-NACHT exon				B30.2 exon			
	Group1	Group2a	Group3a	Group3b	Group1	Group2a	Group3a	Group3b
Group1	0.058	0.44	0.497	0.482	0.097	0.094	0.1	0.398
Group2a		0.029	0.411	0.398		0.009	0.096	0.398
Group3a			0.038	0.11			0.107	0.4
Group3b				0.042				0.226

Table 2b dS	Fisna-NACHT exon				B30.2 exon			
	Group1	Group2a	Group3a	Group3b	Group1	Group2a	Group3a	Group3b
Group1	0.116	1.799	1.684	1.660	0.067	0.061	0.066	1.119
Group2a		0.043	1.525	1.491		0.053	0.058	1.091
Group3a			0.038	0.216			0.065	1.069
Group3b				0.094				0.445

Table 2c dN/dS	Fisna-NACHT exon				B30.2 exon			
	Group1	Group2a	Group3a	Group3b	Group1	Group2a	Group3a	Group3b
Group1	0.52	0.245	0.292	0.291	1.444	1.513	1.569	0.352
Group2a		0.639	0.267	0.267		1.589	1.667	0.361
Group3a			0.942	0.506			1.707	0.371
Group3b				0.677				0.497

Figure 1

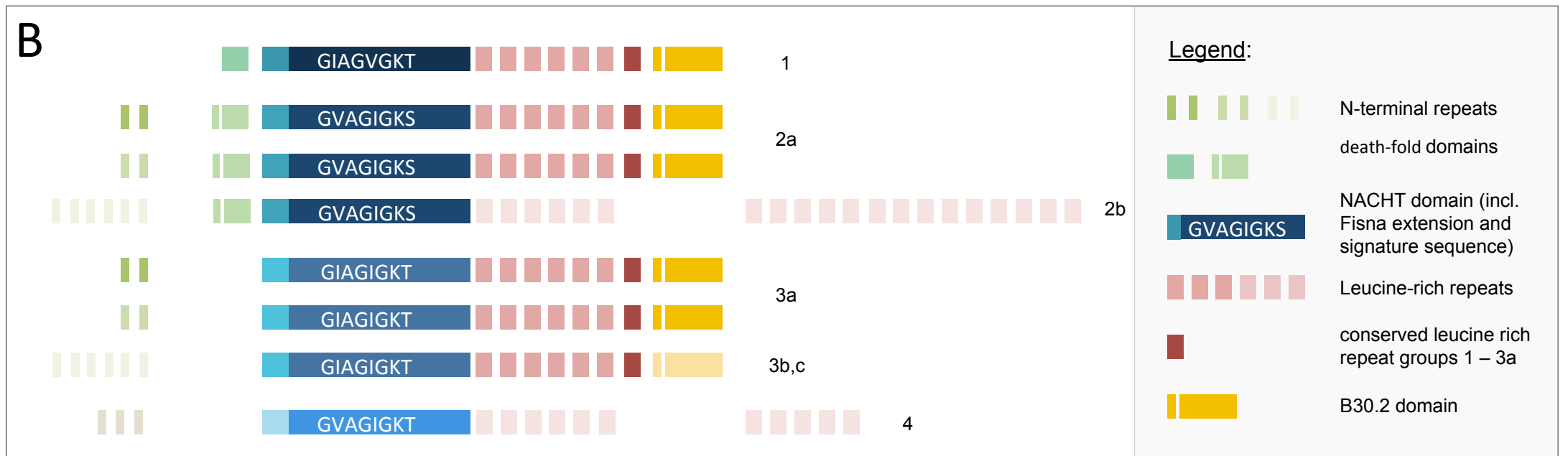


Figure 2

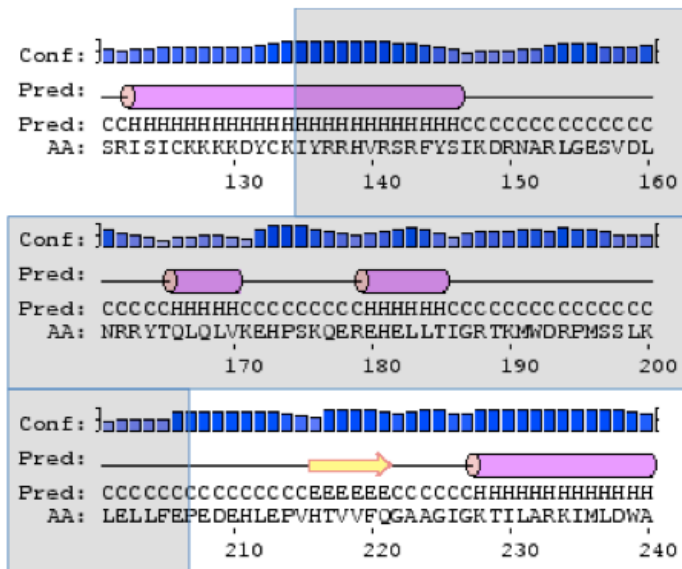
A

Rat	D3ZH10.1	135	206	IYRRHVRSRFYSIKDRNARLGESVDLNRRYTQLQLVKEHPSKQEREHELLTIGRT..KMwDRPMSSLKLELLFE
Panda	D2I5A6.1	132	205	KYKKHVRSRFQCIKDRNARLGESVNLNKRYTRLRLVKEHQSQQEREHELLAIGRTSAKMLDSPVSSLNMELLFE
Mouse	Q3TZ22.1	128	201	TYKDYVRRKFQLMEDRNARLGECVNLNRYTRLRLVKEHSNPtWTQKQFVDVEWERSRTRRHQTSPIQMETLFE
Baboon	A9L902.1	128	201	AYRDYVRRKFRLMEDRNARLGECVNLNRYTRLRLVKEHSNPMQAQQQLLDIGRGHTRTVGHQASPIKIETLFE
Zebrafish	E7F6Q7.1	19	91	QLKCSLKKKQSVFEGIAQQGGSTLLNNIYTDLYITQGASEQVNTHEHEIR.QIEAASRRRHQSLEIQVECKLLFE
	F1QLB9.1	5	77	KLKFNLLKKKHQSVSEGIKQGGSKLLNDVYTDLYMTQCASKQVNTHEHEVR.QIQAAARRNETQERPIECKDMFE
	E7FBD8.1	244	317	ELKLSLKRKYEMVYEGLSQQGQVPVFESVYTDLYITDGINASVNSEHEFRSKIEQLEETGKVNRTPIAREEIFF
	E9QFT1.1	152	224	RFKSNIKQKHQYIFEGFAKQRNPTLLNEIYIDLYITEGGNGEIRNEHEYK.RIEKSLKTAATATIPIKCRDIF1
	E7F5D9.1	209	281	NLKSSLKNRSQRVFEKIAKHGNPVQLNKIFTELYITEGGSGEVNNEHEVR.QIETQCWRPETHETPIRCNDIFK
	E7F2M7.1	221	293	ELKLNLLRKFQCLSEGTATQGDPTLLNEIYTELYITESDSGEINNEHEVR.QIETQSRRATEDTISIQCRDIF1
	F1QCD2.1	175	247	QHQVNLKKKFQCLYEGMVTQDNPAALLNEIYTELYITESSEGEISNEHEVR.QIETQSRRSTTEDTAIKCRDIF1



B

D3ZH10_RAT



F1QY47_DANRE

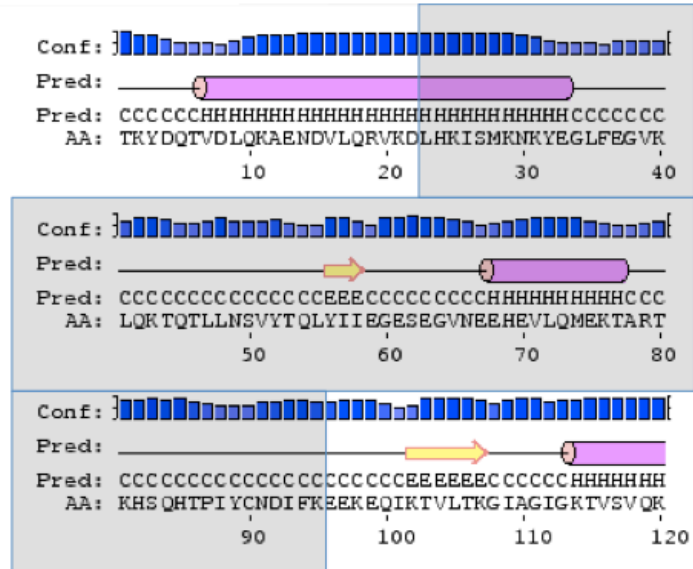


Figure 3

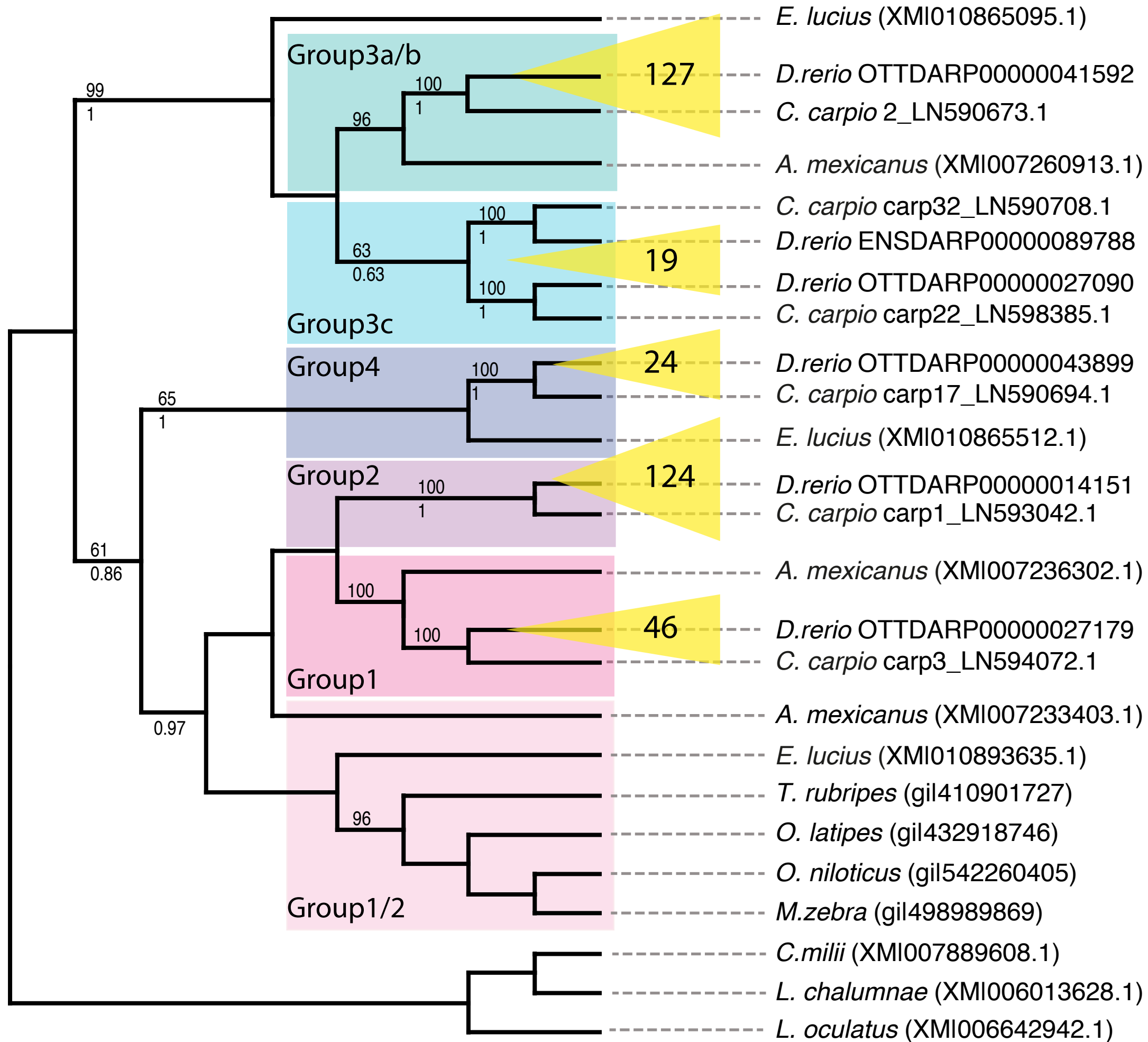


Figure 4

bioRxiv preprint doi: <https://doi.org/10.1101/027131>; this version posted September 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

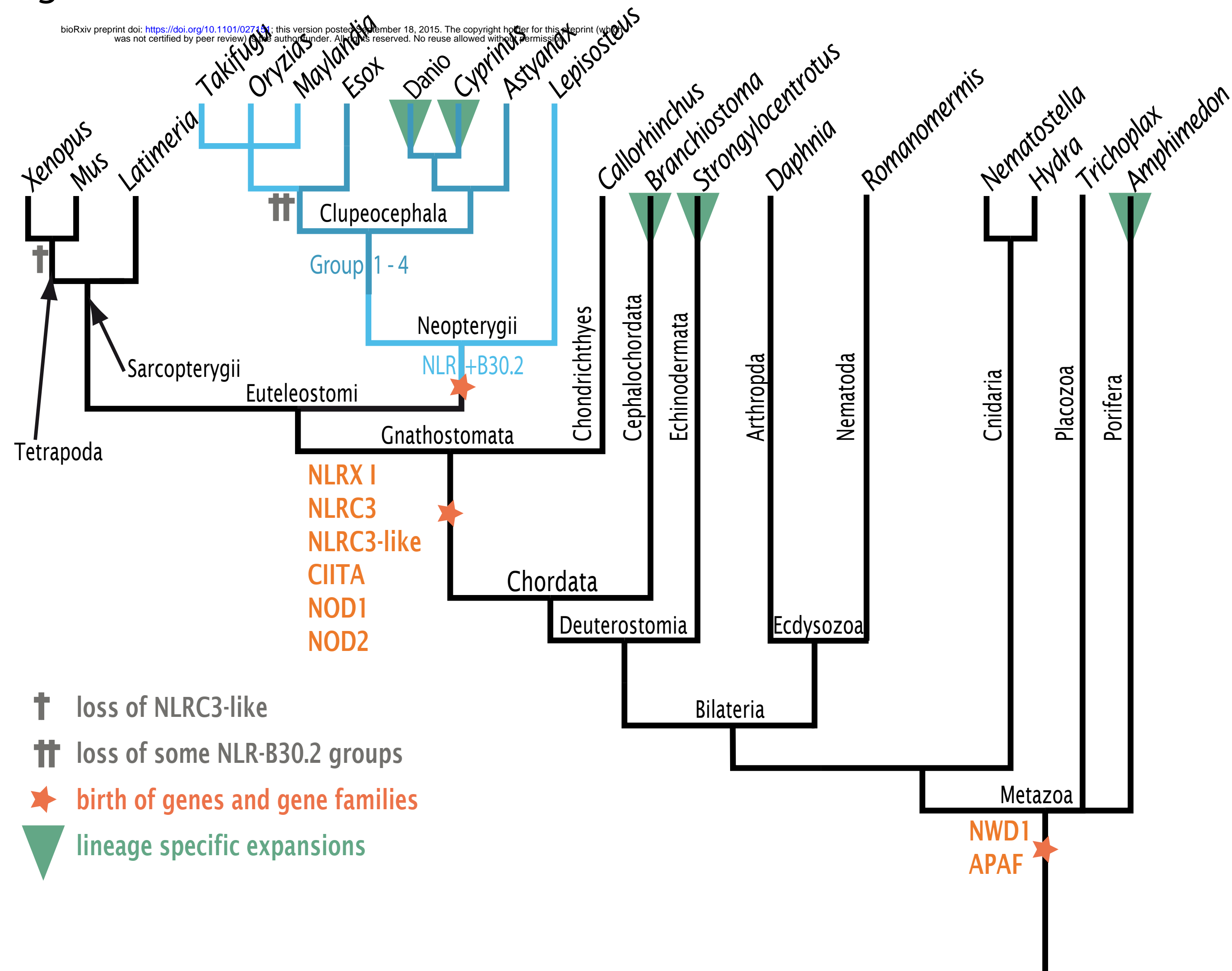


Figure 5

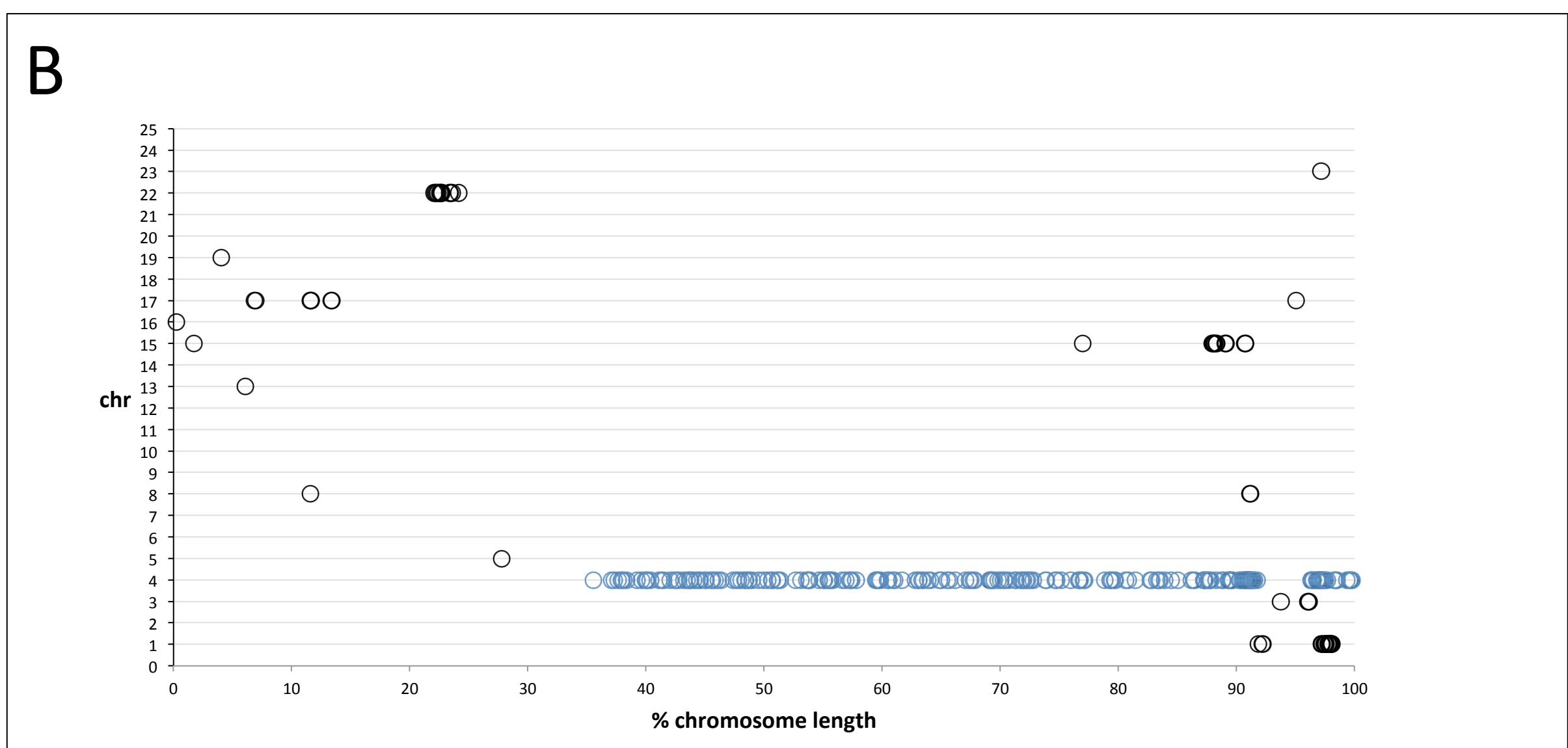
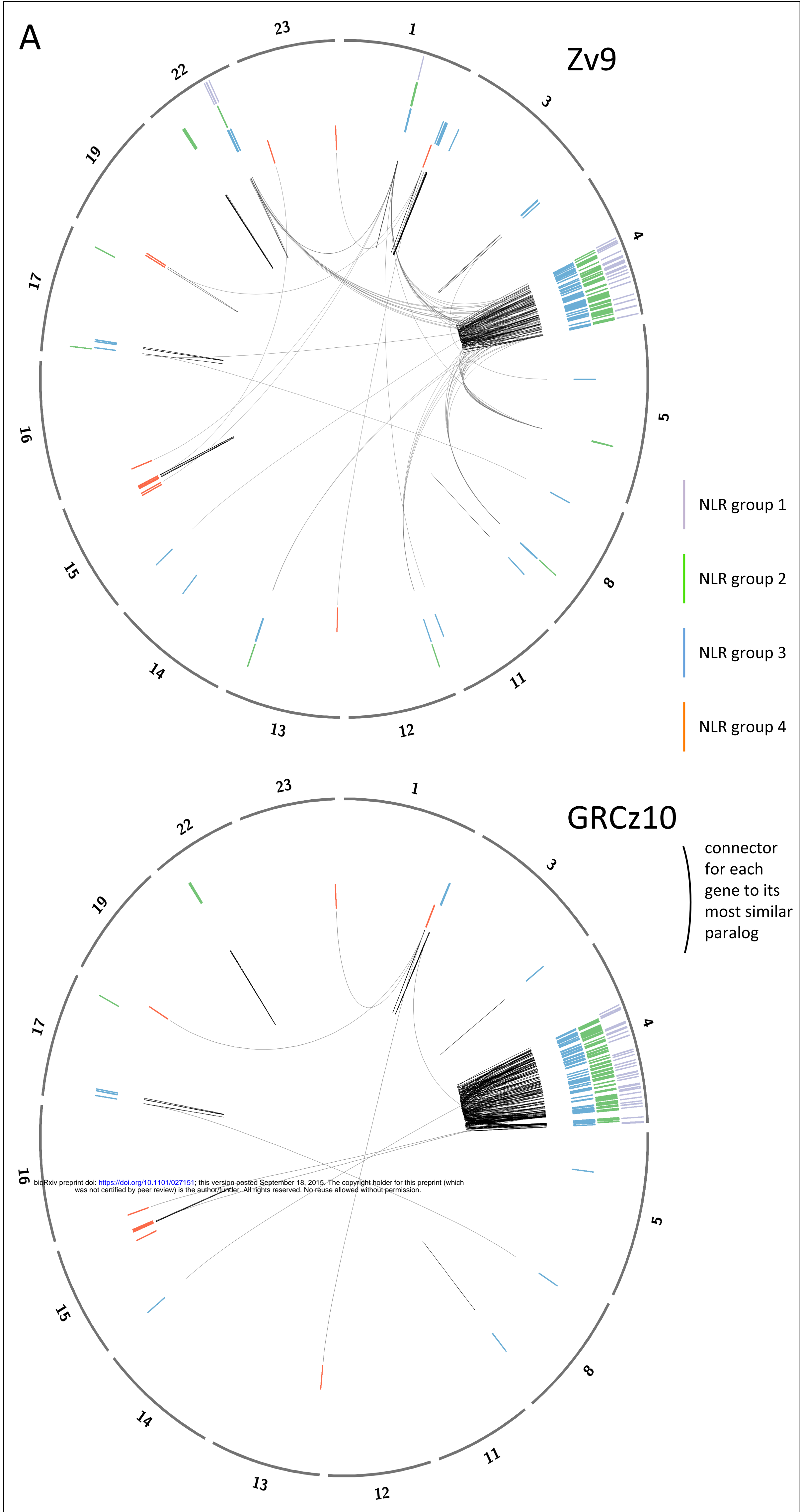
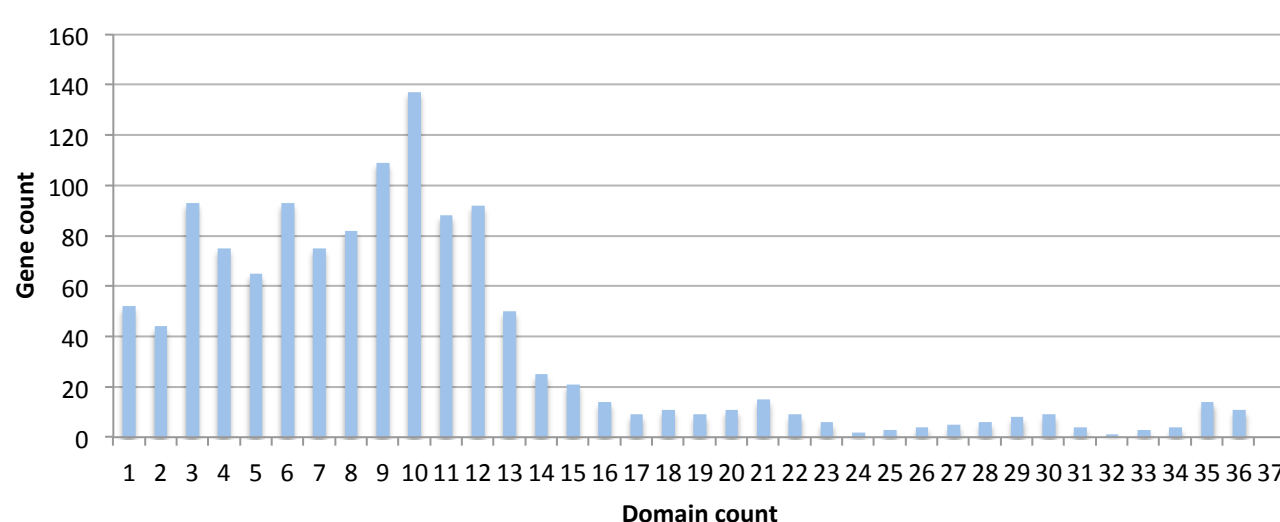
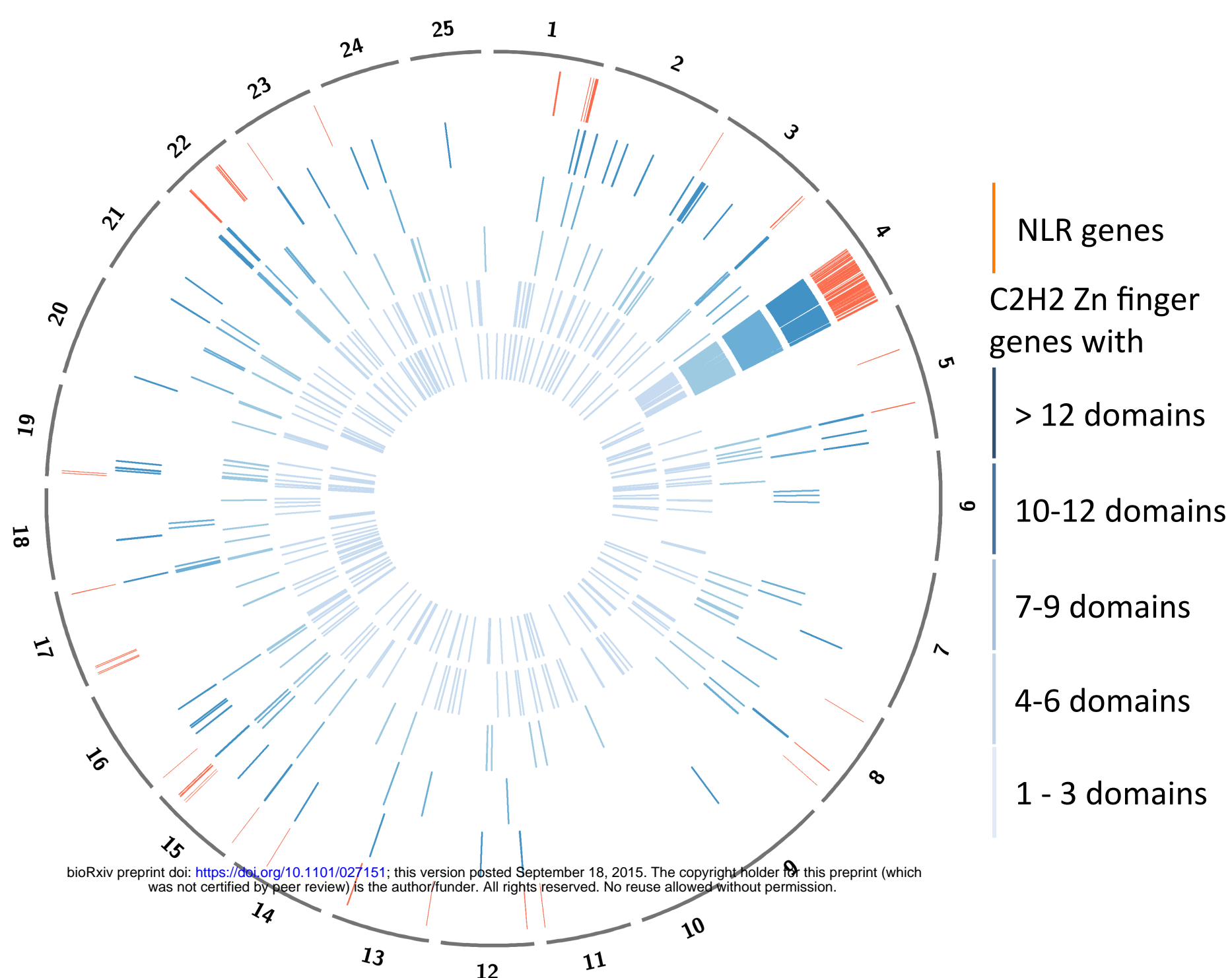


Figure 6

A Number of repeats in Zn-finger genes



B Zn-finger genes vs. NLR genes (Zv9)



C All gene groups (GRCz10)

