Copy number variation and population-specific immune genes in the model vertebrate zebrafish

Yannick Schäfer[1], Katja Palitzsch[1], Maria Leptin[1], Andrew R. Whiteley[2], Thomas Wiehe[1], Jaanus Suurväli[1,3*]

[1*]Institute for Genetics, University of Cologne, Zülpicher Str. 47a, Cologne, 50674, NRW, Germany.
[2]W.A. Franke College of Forestry and Conservation, University of Montana, 32 Campus Drive Missoula, 59812, MT, USA.
[3]Department of Biological Sciences, University of Manitoba, 50 Sifton Rd, Winnipeg, R3T 2N2, MB, Canada.

*Corresponding author(s). E-mail(s): jaanus.suurvali@gmail.com;

Abstract

Different species often have varying numbers of NLR (nucleotide-binding domain leucine-rich repeat) immune genes. Fishes and plants can have hundreds, but not much is known about within-species variability in fish. To address that gap in knowledge we generated a species-wide inventory of NLRs in the model species zebrafish. To capture and analyse the diversity of NLRs in two laboratory strains and four wild populations from Southern Asia, we used targeted sequencing with PacBio. Our results revealed that most zebrafish NLRs are not present in the reference genome. Laboratory strains have fewer NLRs and less presence/absence variation than wild populations. Most wild zebrafish possess around 400 unique NLR genes out of over 1,500 that were seen. Finally, most fish NLRs exhibit very low sequence variation, other than copy number changes. Taken together, this study reveals that vertebrate immune gene families can have a considerable amount of gene copy number variability.

Keywords: copy number variation, gene duplication, immune family, NOD-Like Receptor, PacBio, RenSeq, zebrafish

1 Introduction

Every living being has a genetically encoded immune system, a highly plastic set of mechanisms used for defense against pathogens. Many of its features are lineage-specific (limited to specific groups of animals, plants, or other taxa), while others have defense roles in a wide range of species. Immune genes tend to evolve quickly and are often associated with a high degree of genetic variability, as well as strong selection elicited by the similarly fast-evolving pathogens [1]. Balancing selection is common, often maintaining multiple highly diverse haplotypes in a population [2, 3].

One particular protein family has been demonstrated to have an immune function in animals, plants, fungi, and even bacteria [4–7]. They have a simple structure, consisting of a large nucleotide-binding domain that is followed by a set of smaller repeat domains. In land plants and animals the most common of these small domains are Leucine-Rich Repeats (LRRs). Proteins in which a nucleotide-binding domain is followed by LRRs are generally abbreviated as NLRs.

The majority of genes that convey disease resistance in plants are in fact NLRs, typically with additional N-terminal effector domains. These genes show extensive haplotypic and copy number variation, with different individuals often possessing their own repertoires [8, 9]. This has given rise to a new term, 'pan-NLRome', which refers to all NLRs found across individuals and populations of a species [9]. Individual NLR repertoires are referred to as NLRomes.

Animal NLRs are only distantly related to plant NLRs [10], but are nevertheless essential components of the immune response. Some of them have highly conserved roles as pathogen sensors or transcription factors. Others are components or modulators of inflammasomes, large protein complexes that are assembled in response to biological or chemical signals of danger and activate other components of the immune system [4].

Although most knowledge on animal NLRs originates from studies of mammals (human and mouse), their NLR repertoires (20-30 genes) are smaller than those of many other species [11–16]. In mice, *Nlrp1* is associated with divergent haplotypes containing 1-3 copies of the gene [17], but this does not substantially change the overall NLR count.

Studies relying on short read-based sequencing typically report fish NLRomes in the range of 20-50 genes, similar to what is known from mammals [18–20]. However, short reads can underestimate copy numbers when the copied sequence itself is much longer. Two groups independently reported hundreds of NLRs in zebrafish (*Danio rerio*) soon after a highly contiguous genome became available [21, 22]. Later studies found that expansions and contractions of the NLR family are common among fishes, with many other species also having hundreds [15, 16, 23, 24].

The majority of NLRs found in fishes represent a fish-specific subtype that was originally labelled as NLR-C, based on phylogenetic analysis of the zebrafish NLRome [22]. This is not to be confused with NLRC, an abbreviation often used for NLRs that contain a Caspase Recruitment Domain (CARD). The zebrafish reference genome contains nearly 400 NLR genes, most of which are NLR-C, as in other fish [23]. Two thirds of them are located on the putative sex chromosome (Chromosome 4) [25], in a genomic region associated with extensive haplotypic variation [23, 26].

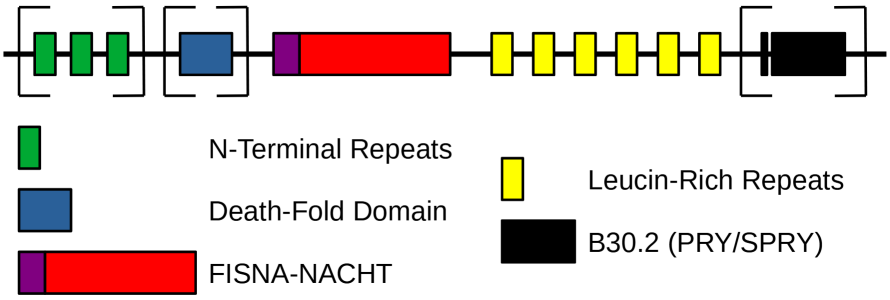

Fig. 1 Schematic structure of a typical zebrafish NLR-C gene. Each box represents an exon. Exons encoding for N-terminal repeats, death-fold domains and B30.2 domain occur only in subsets of NLR-C genes. The number of N-terminal repeats (green) and Leucine-Rich Repeats (yellow) can vary. Exons surrounded by square brackets are only found in subsets of the NLR-C genes.

A schematic structure of proteins encoded by zebrafish NLR-C genes is presented on Figure 1. They are characterized by a specific motif (FISNA) that precedes the nucleotide-binding domain (NACHT) and is encoded by the same large exon near the start of the protein [23]. Throughout the rest of this study, that exon is referred to as FISNACHT. FISNA and NACHT are in some cases preceded by the effector domain PYD, but this is encoded by a separate exon [23]. Additionally, many NLR-C proteins have a B30.2 domain (also known as PRY/SPRY) at the other end, separated from the others by Leucine-Rich Repeats (Figure 1) [23]. Neither B30.2 nor FISNA is associated with NLRs other than the fish-specific subtype NLR-C. While FISNA is specific to NLR-C, B30.2 is also found in other large immune gene families [16, 27]. Zebrafish NLR-C genes can be further divided into four groups based on sequence homology [23].

It is not known why fishes often have such large NLRomes, how they evolve, and how much within-species genetic variability occurs. However, the repeated expansions and contractions of the family [16] suggest that these genes have a high rate of birth and death. Numerous studies have shown that the expression of fish NLRs is induced in response to infections with pathogens ranging from viruses to bacteria (reviewed in [28, 29]). Some NLRs with PYD or CARD domains can even form inflammasomes similar to mammalian NLRs [30, 31]. A species-wide inventory of major NLR exons (pan-NLRome) in a model species such as zebrafish can provide valuable insights into the evolution and diversity of this understudied part of the fish immune system.

2 Results

2.1 Sequencing of wild and laboratory zebrafish

By adapting and modifying the RenSeq-PacBio protocol that combines RNA bait-based exon capture with PacBio SMRT technology [8], we generated Circular Consensus Sequence (CCS) data for targeted parts of the zebrafish immune repertoire.

In total we sequenced 77 zebrafish from four wild populations (Figure 2) and 16 from two laboratory strains (TU and CGN (KOLN), 8 fish each).

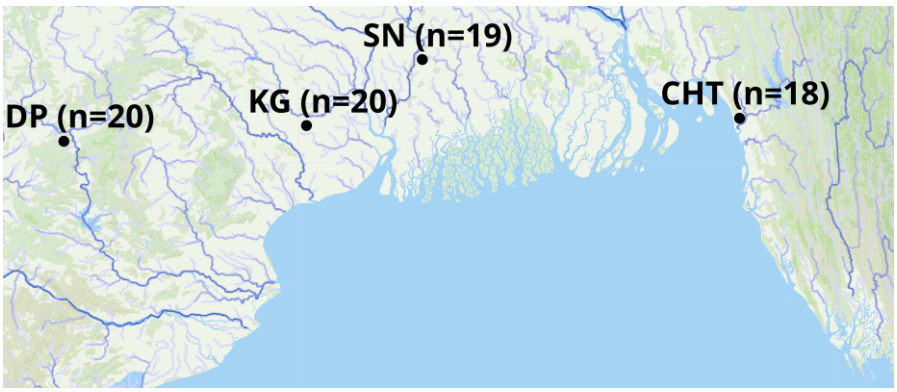

Fig. 2 Sampling sites for wild zebrafish. Sample sizes are indicated in parentheses. The map was created based on geographic data collected and published by AQUASTAT from the Food and Agriculture Organization of the United Nations [32].

Our approach yielded sequence data for the large FISNACHT and B30.2 exons from all 93 zebrafish, as well as additional intronic sequence and smaller exons surrounding them. Most of the obtained NLR-associated reads were between 1 and 3.5 kb in length (Figure 3A2), contigs assembled *de novo* from these were 3-7 kb in length (Figure 3B2). Contig number per individual across both haplotypes ranged from fewer than 10 to around 800, with the lower range observed in samples with overall poor sequence coverage, mainly from the wild population DP (Figure 3A1). For most individuals this was sufficient to detect the presence of hundreds of NLRs, although the exact number detected varied (Figure 3B1).

2.2 Signature motif of B30.2 in NLRs

B30.2 and FISNACHT exons were generally not found on the same sequence reads and contigs. This complicates the study of B30.2 in particular as it is also found outside NLRs, most notably in the large family of TRIM immune genes [27]. However, by examining the existing NLR annotations in zebrafish reference genome, we found a

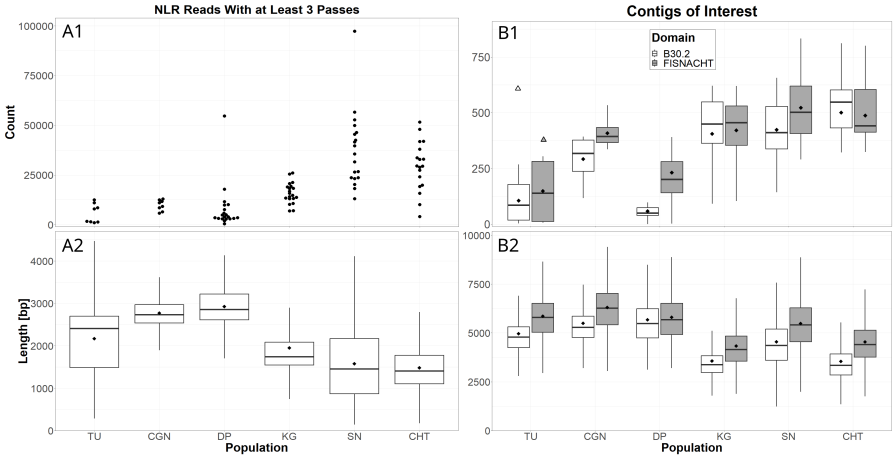

Fig. 3 Sequencing and assembly statistics, split by population. A1: Total CCS read counts that align to reference NLRs. Each black dot marks one zebrafish. A2: Length distribution of CCS sequences that map to NLRs. B1: Total counts of B30.2 and FISNACHT contigs assembled *de novo* from the data. Triangles above TU mark the number of NLR exons found in the reference genome. B2: Length distribution of *de novo* assembled NLR contigs. Mean values are marked with black diamonds, medians with horizontal black lines. Outliers are not shown on box plots.

highly conserved 47 bp exon that precedes B30.2 in most NLRs with a B30.2 (NLR-B30.2). This small exon is within 0.1-2 kb of the adjacent B30.2 and almost exclusively found in NLRs, with only a few exceptions (see Supplementary Table 1). Its sequence is highly conserved, although two pairs of amino acids and the nucleotides encoding them are more variable (Figure 4A).

We were able to differentiate between B30.2 from NLR genes and B30.2 from elsewhere in the genome by detecting the presence/absence of those 47 bp in proximity of B30.2 (Figure 4B). The ratio of NLR-B30.2 contigs and non-NLR-B30.2 contigs varied considerably between populations (Figure 4B). The KG population from West Bengal in particular had many more NLR-associated B30.2 domains than the others, while the number of B30.2 domains not associated with NLRs in KG was similar to what was observed for the geographically closest population SN (Figure 4B).

As expected, assembled contigs containing B30.2 with the preceding 47 bp motif generally aligned to the same regions in the reference genome as those containing FISNACHT (Figure 4C). In contrast, B30.2 exons without the motif were spread

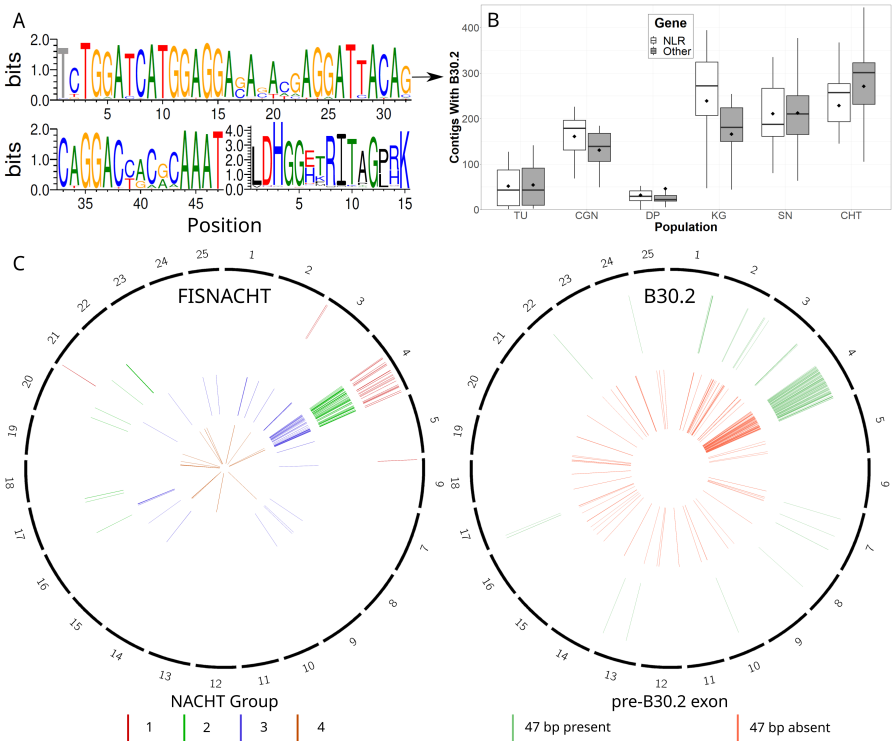

Fig. 4 Conserved exon connecting LRRs to B30.2. A: Sequence logos for nucleotides (on top, continued on bottom left of the panel) and amino acids (bottom right of the panel) of the 47-bp NLR-specific exon. The first nucleotide was ignored when translating to match the correct reading frame. The height of each base represents its information content in bits. B: B30.2 exon counts, separated in two categories by the presence or absence of the 47-bp NLR-specific exon. Means are marked by black diamonds, medians by black lines. C: Genomic distribution of FISNACHT belonging to each NLR (NACHT) group (left), and B30.2 that are with or without the 47-bp NLR-specific exon (right).

throughout most of the genome, with a higher concentration on Chromosome 4 (Figure 4C).

2.3 Zebrafish pan-NLRome

After identifying all NLR-specific sequences in our dataset, the next step was to assign those to individual genes. We first attempted to align all data to the reference genome (GRCz11), but most sequences aligned to it poorly. Moreover, careful examination revealed that often the only aligning part was the exon itself, with many divergent sequences aligning to the same NLR genes. This was observed both for CCS and the

de novo assembled contigs. We thus opted for a reference-free approach and used orthology clustering of the CCS to establish a pan-NLRome, defined as a set of FISNACHT and B30.2 that each had at least one matching sequence in one zebrafish. This resulted in 1,560 FISNACHT, 574 NLR-associated B30.2 (NLR-B30.2, identified by the preceding motif), and 743 B30.2 without the motif. Since FISNACHT is only found in NLRs, its count (1,560) can be considered equal to the total number of distinct NLR-C genes in the dataset.

By analysing the representative (consensus) sequences of each orthogroup we found 105 of the 1,560 FISNACHTs to be preceded by a PYD-encoding exon (see Figure 1). 100 of the PYDs were next to a group 1 FISNACHT [23]. Furthermore, among the representative sequences, 150 FISNACHT and 64 NLR-B30.2 had an early stop codon. Most of the truncated sequences had the best reference match on either Chromosome 4 or Chromosome 1 (59% and 17% of truncated FISNACHT, 64% and 23% of truncated B30.2).

2.4 Copy number variation of NLR genes

Now that we had a set of NLR sequences not dependent on the reference genome, we could align all our CCS reads against that. By doing this we found that for each sample, the fraction of CCS aligning to each individual NLR (relative depth) can vary greatly, from none to many. Furthermore, there are differences among genes, among individuals, and even among populations (Figure 5). Those with an increased relative depth have likely been duplicated at least once, and the copies align to the same match in the pan-NLRome. All populations showed distinct patterns of duplications, with laboratory strains differing from wild zebrafish (Figure 5).

Of the 1,560 FISNACHT and 574 NLR-B30.2, nearly half were detected in at least one individual from each of the four wild populations. This shared core set of NLR exons consisted of 714 FISNACHT (46%) and 229 NLR-B30.2 (40%). There were also

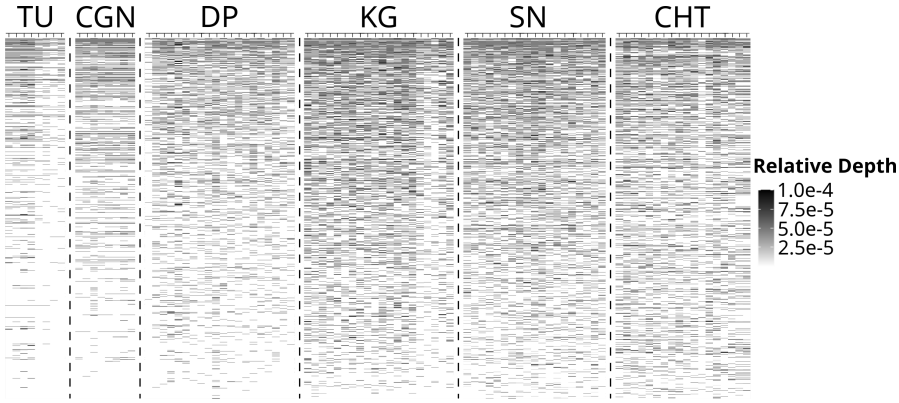

Fig. 5 Presence/Absence variation of NLR genes. Columns correspond to individual zebrafish. Rows correspond to the representative FISNACHT sequences. NLRs with less than ten matching CCS across the entire dataset are not shown. The TU, KG and CHT samples with lower depths overall (lighter colors) all had less sequencing depth than other samples of the corresponding populations.

NLR sequences shared between just two or three wild populations, as well as private NLRs only found in a single wild population (Figure 6A).

There was a considerable amount of variability in absolute NLR counts (Figure 6B). Some of this can be attributed to technical differences, as four individuals from each population were sequenced separately from the others (detailed in Materials and Methods). However, all analysed populations exhibited copy number variation even among samples sequenced in the same batch, with the highest NLR counts observed for the wild populations KG, SN and CHT (Figure 6B). In DP, the wild population with low sequence yields, the number of FISNACHT exons appeared close to laboratory strains. Same patterns of copy number differences were observed for NLR-B30.2. We observed more FISNACHT than NLR-B30.2 in the data, consistent with previous observations of B30.2 only being present in some subsets of zebrafish NLRs [23].

Modelling the process of adding NLRs from new individuals until no new copies are discovered allowed us to estimate their total numbers in each population without requiring additional samples (Figure 6C1, C2). We estimated a total of 520 and 570 NLRs in the laboratory strains TU and CGN, respectively. Wild populations have four times as many, with 2,283 in KG, 1,896 in SN and 2,452 in CHT for KG, SN

and CHT. For DP the estimated pan-NLRome size was 25,284, but this is likely an artifact caused by the low sequencing coverage.

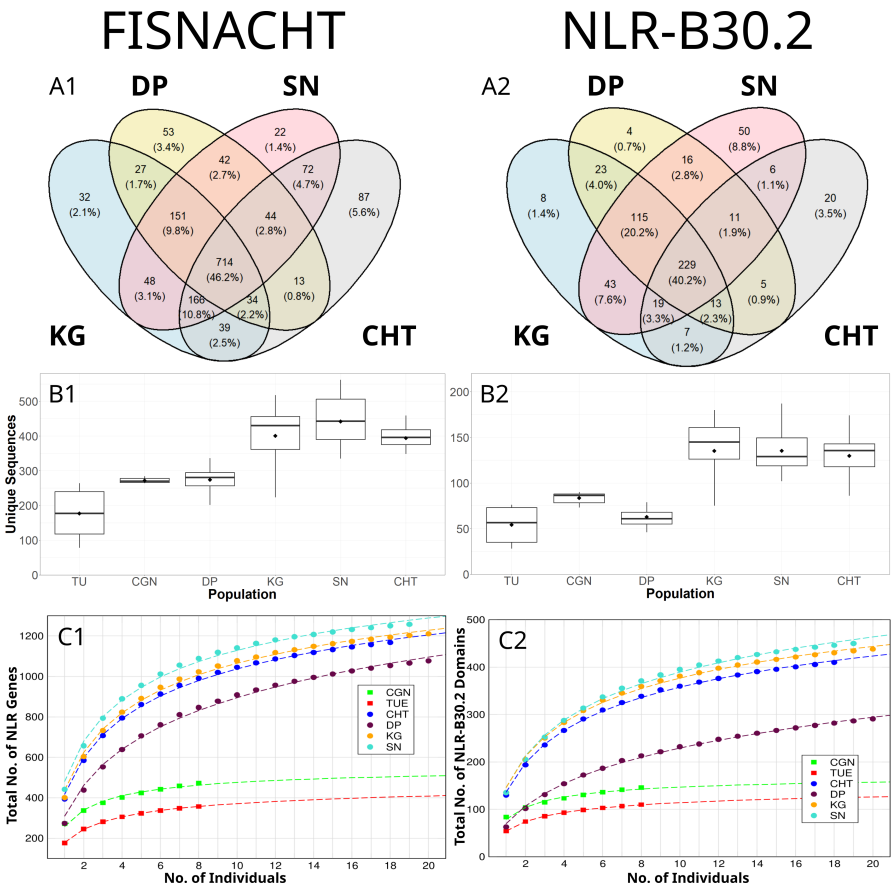

Fig. 6 Copy number variation of NLR genes. A1, A2: Numbers of private and shared NLR sequences in wild populations. B1, B2: Numbers of unique NLR sequences (each with one or more copies per individual) found in fish of the sequenced strains. Black diamonds on the box plots denote means, horizontal lines denote medians. C1, C2: Population-specific pan-NLRomes and sets of NLR-B30.2 domains. Data points (filled circles and squares) show the average number of additionally discovered NLR genes (as identified via their FISNACHT domain) when adding data from the x-th individual to a previously examined set of $x-1$ individuals. The dotted lines represent the result of non-linear curve fitting to the function $y = \alpha \sum_{k=1}^{x}(1/k^{\beta})$, the generalized x-th harmonic number (detailed in Materials and Methods).

2.5 Comparison of pan-NLRome to the reference genome

To find out how much overlap the pan-NLRome has with NLRs in the reference genome (GRCz11), all representative sequences in it were mapped to GRCz11. The reference genome is based on the TU strain [33], but only 75% of NLRs from our TU fish aligned to it well (Figure 7A). In CGN, the other laboratory strain, this number was even less, and only 60-65% of NLRs were successfully mapped. In wild populations the number dropped to around 50%. Nearly all matches to reference aligned either very poorly or extremely well (Figure 7B). 9 FISNACHT and 10 NLR-B30.2 did not map anywhere in the reference. In summary, the reference genome is not representative of the full diversity of NLRs in its origin strain and even less so for other populations.

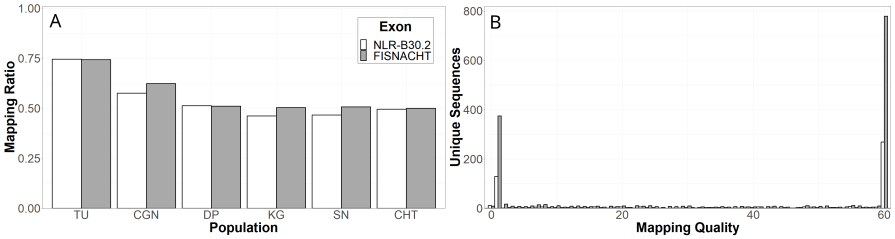

Fig. 7 Statistics from mapping the unique NLR sequences to the reference genome GRCz11. A: Proportions of unique FISNACHT and NLR-B30.2 sequences that were successfully mapped to the reference genome GRCz11 with a mapping quality of 60, by population. B: Most representative NLR sequences from the pan-NLRome aligned to GRCz11 with either very high (60) or very low quality.

2.6 Purifying selection on single nucleotide variants

To learn more about the evolutionary trajectory of NLR genes, we calculated the genetic diversity parameters θ_π (nucleotide diversity), θ_w (Watterson's estimator) and their ratio θ_π/θ_w. This ratio was calculated as opposed to Tajima's D because it is a unit-less measurement and shows fold differences in the two parameters more clearly. In a previous study of overall background genetic variation in zebrafish, laboratory strains were found to have a θ_π of 0.001-0.004 and θ_w of 0.001-0.003. The same estimates for wild populations were 0.004-0.008 for θ_π and 0.003-0.009 for θ_w [34].

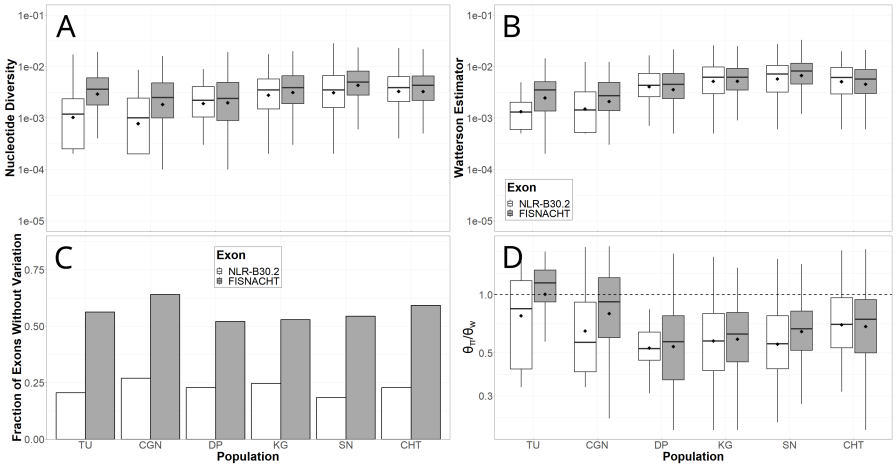

Fig. 8 Sequence variation of FISNACHT and B30.2 exons of the NLRome, shown by population. A: Nucleotide diversity (θ_π). B: Watterson estimator (θ_w). C: Fraction of invariant exons. D: Ratio of θ_π/θ_w. Invariant exons are not shown. The black, dotted line marks a ratio of 1. The black diamonds on box plots denote means, horizontal lines denote medians.

The NLR exons were largely on the lower end of sequence diversity (as expected of protein-coding sequence), with values up to 0.0012 for θ_π and for up to 0.0019 for θ_w. The estimates for B30.2 and NACHT exons were both similarly low in most populations (Figure 8A, B), and in many exons there were no genetic variants at all in a given population (Figure 8C). The least variable exons were B30.2 in laboratory zebrafish, which had fewer single nucleotide variants than the same exons in wild populations and also fewer than FISNACHT exons in general (Figure 8A, B).

Overall, θ_w tended to be higher than θ_π and the majority of variable NLR exons from wild zebrafish had θ_π/θ_w ratios less than 1 (Figure 8D). This observation implies an excess of rare alleles which can be interpreted as a signal of purifying selection. In laboratory zebrafish the ratios were closer to 1, with the exception of NLR-B30.2 in the KOLN strain (CGN) (Figure 8D).

3 Discussion

We sequenced and assembled the exons of hundreds of NLRs from 93 zebrafish, thus capturing the diversity of this gene family in four wild populations and two laboratory

strains. We found evidence that each genome from a wild individual contains only a fraction of more than 1,500 identified NLR copies. The number of unique NLRs per individual, each with 1 or more copies, ranged from around 100 to 550. While some of the lower numbers were likely underestimated due to low sequencing depths in specific samples and the population DP, the observed slow increase in newfound NLR gene copies per sample in all populations other than DP indicates that not many NLRs were missed.

One major challenge in obtaining our data was the physical distance separating the two sets of targeted exons, FISNA-NACHT (FISNACHT) and B30.2. Our protocol used PCR with primers targeting ligated adapters to amplify the below-nanogram amounts of genomic DNA obtained from exon capture. This limited our fragment sizes to the lengths of what the polymerase was able to amplify. Zebrafish NLRs can have their exons spread out across tens of kilobases separated by large introns, so that B30.2 and FISNACHT exons are generally unlinked. We also did not design any baits for the short repeat exons (LRRs) separating B30.2 and FISNACHT, as LRRs are commonly found in many genes across the genome. As a result, we cannot determine which of the retrieved exons belong to the same gene copies. We therefore analysed the two exon categories separately throughout the study. If in future experiments enough genetic material can be collected to skip amplification (more than a fin clip), whole NLR genes could be sequenced and assembled. However, we note that even when long reads with very high precision will be available, the construction of a perfect pan-NLRome across many individuals might still not be easily achievable. This is due to difficulties in distinguishing copies that were affected by recent duplication, unequal recombination or ectopic gene conversion between different gene copies. Such complex genetic signals lead to difficulties in assembly and clustering which in turn result in anomalies in the data in ways we also observe: some unique pan-NLRome sequences that result from orthology clustering had to be discarded in a subsequent filtering step

as they were not supported by CCS reads. In each of those cases, the reads which were used to (mis-)assemble the respective NLR-containing contigs mapped better to other sequence of the pan-NLRome.

With a shift of focus towards pan-genomic studies (see [35–37]), future research on large multi-copy gene families should perhaps aim to generate gene databases similar to our pan-NLRome as starting points for deeper analyses. As shown for zebrafish NLRs, the availability of a single high quality reference genome is not sufficient to estimate the true diversity of large gene families.

The overall zebrafish pan-NLRome resulting from our experiment consists of 1,560 unique FISNACHT-containing sequences. Since the combination of FISNA and NACHT is only found in NLR-C genes and is one of their defining features, this number can also be used to estimate the number of NLR-C in zebrafish. Mathematical modelling based on a harmonic distribution suggests that the investigated laboratory strains possess roughly 500-600 NLRs while wild populations possess around 2,000. The pan-NLRomes are closed because sequencing more samples adds only few new NLRs to the total collection. However, to sequence all NLR genes, it would be necessary to sequence hundreds of individuals from each population. Besides, there are more, phylogeographically different populations - for example in Nepal and the Western Ghats - which likely have tens if not hundreds of NLR copies not found in the populations we sequenced [38].

Our predicted number of NLRs in a population is on the same scale as the 2,127 putative NLR genes found in the NLRome of *Arabidopsis thaliana* [9]. Structural similarities of NLRs in plants and animals are thought to be the result of convergent evolution [39], but it appears that the similarities extend beyond domain structure. While the zebrafish and *A. thaliana* face different environments and pathogenic challenges, their NLRs appear to have undergone similar evolutionary trajectories, such as

the formation of isolated gene clusters [40]. Copy numbers vary greatly between *Arabidopsis* accessions [40], which is something we also observe in zebrafish. The sizes of NLR repertoires differ between zebrafish individuals in at least three of the wild populations. Whether the low number of NLR genes in the fourth wild population, DP, is an artifact caused purely by low coverage or also by evolutionary factors remains unclear at this time.

We have previously demonstrated a substantial reduction in single nucleotide variation in zebrafish laboratory strains compared to wild populations [34]. Here we showed that the copy numbers of the NLRome are similarly reduced. In the same line, we also find laboratory strains to have less presence/absence variation than wild populations. In addition, the reduced amount of pathogenic challenges in laboratory environment could lead to a rapid loss of unnecessary genes, especially if unequal recombination is frequent [41].

In *Arabidopsis thaliana*, 464 conserved, high confidence orthogroups were previously identified, 106-143 of which were defined as the core NLRome because they were found in a large subset of accessions [9]. In wild zebrafish, we found a core NLRome consisting of 714 NLR genes, each of which was detected in at least one individual from all wild populations. Only half of those (52.5%, 375 in total) could be confidently mapped to the reference genome. We postulate that as immune genes, many of the core NLR genes are likely conserved between populations because they provide a fitness advantage in the defense against pathogens, although some may be retained because of neutral processes such as genetic drift. The additional NLRs shared between only some of the wild populations, and the population-specific NLRs may represent local adaptations to ecological niches. Additionally, there could be functional redundancy within the NLRome, so that different individuals have different NLRs with the same functional role. This idea is further supported by the fact that most wild fish had only around 400 NLRs and therefore only a subset of the core NLRome.

We found a large collection of B30.2 exons, but only a subset of these are associated with NLRs. This was expected, because the B30.2 is also found in other protein families. We were able to distinguish B30.2 encoding exons that belong to NLRs rather than other families because we identified a small exon that is almost exclusively found preceding the B30.2 of NLRs and rarely present in other B30.2-containing genes (Supplementary Table 1). It consists of 47 highly conserved nucleotides encoding an amino acid sequence that connects the B30.2 domains of NLRs with their preceding Leucine-Rich Repeats. The function and evolutionary history of this exon may be further characterized in future studies, but for the purpose of this study it was simply used to identify B30.2 sequences that originate from NLR genes.

By analysing the NLR-specific B30.2 (NLR-B30.2) exons, we found them to exhibit patterns of copy number variation akin to those described above for FISNACHT. Furthermore, a core set in wild populations similar to what was found for FISNACHT was identified here as well, representing nearly half of all unique B30.2 sequences found across populations - again similar to what was observed for FISNACHT. B30.2 exons not associated with NLR genes looked to be copy number variant as well, with differences in both absolute copy numbers and in population-specific ratios of NLR-associated B30.2 to other B30.2.

The fact that hundreds of NLR gene copies are maintained in the zebrafish, together with hints of purifying selection in the sequences of their FISNACHT and B30.2 exons suggest that the evolution of these genes is driven by significant selective pressures. Although the expression of fish NLRs is often induced by pathogen exposure (reviewed in [28, 29]), the functions of zebrafish NLR-C genes are poorly understood. While many human NLRs contain death fold effector domains such as PYRIN or CARD, only a small subset of zebrafish NLRs include them. Not much is known about the molecular function of NLRs that do not contain N-terminal effector domains, and the majority of zebrafish NLRs do not possess these. It is possible that B30.2 takes up the role

of the effector in some cases but NLR-C genes of group 4 contain neither B30.2 nor any other known effectors [23]. B30.2 domains in other proteins have been reported to interact with immune repertoire components such as β2-microglobulin, interferon-regulatory factors or caspases [42–44]. Alternatively, the B30.2 domain may have a role in pathogen recognition by binding to specific pathogen-associated molecular patterns. In any case, the consistent presence of B30.2 exons in NLR-C genes as well as their excess of rare alleles point to a selective benefit of the B30.2 exon for the NLR gene of the individual.

There are at least two explanations for the extensive copy number variation seen among zebrafish populations. First, it could be attributed to a high degree of haplotypic variation. Accordingly, large DNA fragments would contain different sets of genes and gene copies, similar to the zebrafish MHC loci [3]. Supporting this theory is the fact that extensive haplotypic variation occurs on the long arm of Chromosome 4, the location containing over two thirds of all NLRs in zebrafish [26]. Such segregating haplotype blocks would explain the existence of the core NLRome, but not the frequent presence of genes that occur only in a single individual.

Alternatively, the evolution of NLR-C and possibly other large gene families could be primarily driven by duplication events [45, 46] and secondarily by gene conversion [22]. Gene duplications are caused by unequal recombination, transposon activity (retrotransposons, helitrons and others) or whole genome or chromosome duplications. The arrangement of NLR-B30.2 genes in clusters on the long arm of Chromosome 4 suggests that tandem duplication via unequal crossing-over (see [41]) may have played an important role in the expansion. Since there are many transposable elements (TEs) on the long arm of Chromosome 4 [33], it is also reasonable to assume that at least some of them have assisted in the local expansion and in transferring NLRs and their exons to chromosomes other than Chromosome 4.

It is tempting to speculate that Chromosome 4 could be a source of NLRs which continuously generates new copies in a process of "runaway" evolution [46]. However, to maintain a stable genome size, gene gains would have to be balanced by an equal amount of gene loss. The function of NLR-C genes may be lost via mutations and pseudogenization, but the pseudogenes would still add to the genome size. Alternatively, the entire sequence could be lost through deletions and unequal recombination. This would allow only NLR genes contributing to the functionality of the immune system to be kept, while others would disappear.

In plants, tandem duplication is thought to be the primary driver of NLR gene expansion [47]. Increased association of TEs with duplicated gene families led to the hypothesis that transposition played an essential role in the creation of the gene families [48]. Although the NLR families of plants and animals arose independently [39], their evolutionary paths may have much in common. If there is a common mechanism driving the diversity of unrelated NLR genes in such distantly related species, then the same mechanism might also act on other similarly large gene families. Examples of this could include NLRs of other phylogenetic clades or even non-NLR families, for example TRIM [16, 27] and NITR [49] immune genes or odorant receptors [50]. This mechanism would likely involve a combination of haplotypic variation, transposon activity and unequal recombination, combined with mechanisms that control gene retention.To clarify the extent to which each of these mechanisms contributed to past and recent duplications of NLR-B30.2 genes in the genome of the zebrafish and related teleosts, investigations on the abundance and types of transposons in close proximity to them as well as on rates of unequal crossing-over events in specific genomic regions will be necessary.

To reach a better understanding of the fish-specific expansion of the NLR-B30.2 gene family specifically but also of genomic adaptations of the innate immune system in general, it will be necessary to thoroughly examine how the mechanisms that drive

and limit gene family expansions work on a molecular level. We discovered that there are large differences in the copy number and composition of the NLRomes of individual fish even within recently diverged wild populations. This finding shows that the mechanisms which diversify and restrict gene families are at full play in the zebrafish NLR gene family.

4 Methods

4.1 Sample collection

Wild zebrafish had originally been sampled from four sites in their native habitat, in collaboration with local researches. Three sites were located in India (DP, KG, SN) and one in Bangladesh (CHT) (Table 1, Figure 2). Samples were collected in the frame of other projects as previously described for CHT [38] and SN [51]. KG and DP samples were collected around the same time as SN, with the same set of collaborators and by using similar methods. We used a fin clip from 20 fish per wild population as starting material. In addition, 8 laboratory zebrafish from each of the strains Tübingen (TU) and KOLN (CGN) were kindly provided by Dr. Cornelia Stein (Leptin and Hammerschmidt labs, University of Cologne). All samples were preserved in 95% ethanol until use.

Table 1 Zebrafish samples sequenced in this study. For populations in which not all fish yielded data, the final number of samples is indicated in parentheses

Name	Origin	Coordinates	Fish
Tübingen (TU)	Unknown (pet store)	-	8
KOLN (CGN)	Unknown (pet store)	-	8
Dandiapalli (DP)	Odisha, India	N22.22155 E8479430	20
Leturakhal (KG)	West Bengal, India	N22.26189 E87.27881	20
Santoshpur (SN)	West Bengal, India	N22.93765 E88.55311	20 (19)
Chittagong (CHT)	Chattogram, Bangladesh	N22.47400 E91.78300	20 (18)

4.2 DNA extraction

High Molecular Weight (HMW) DNA from laboratory zebrafish was extracted from caudal fin clips using the QIAGEN MagAttract HMW DNA extraction kit. HMW DNA from wild zebrafish was extracted from caudal fin clips using the NucleoSpin Tissue Kit from MACHEREY-NAGEL with the following adjustments. Tissues other than muscle were removed before DNA extraction with forceps. The incubation time of the Proteinase K treatment was changed from 1-3 hours to 10-15 minutes. An RNAse A treatment step was included by incubating with 400 µg RNAse A (Sigma-Aldrich) for 2 minutes at room temperature. All DNA samples were quantified and quality checked with Qubit 3.0 (ThermoFisher Scientific), 0.8% agarose gels and the 4200 TapeStation Electrophoresis System (Agilent Technologies). DNA extraction failed for one of the 20 CHT samples, but was succesful for the other 95 fin clips.

4.3 Shearing and barcoding

High molecular weight (HMW) DNA was sheared into 1.5 to 6 kb fragments with the red miniTUBEs of the Covaris M220 ultrasonicator. Quality control after shearing was performed using the 4200 TapeStation Electrophoresis System (Agilent Technologies). The obtained DNA fragments were size selected with 0.4x AMPure XP beads (Beckmann Coulter Inc) to exclude fragments smaller than 1.5 kb.

For wild zebrafish samples, a DNA damage repair step was included in order to repair any possible DNA damage resulting from long periods of storage (particularly important for the older CHT samples). The repair step was carried out with PreCR Repair Mix (New England Biolabs).

The DNA fragments were then barcoded with the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs) and NEBNext Multiplex Oligos for Illumina, Index Primers Set 1 (New England Biolabs). The manufacturer's standard protocol was followed until the amplification step for the enrichment of barcode-ligated fragments.

At this stage, the recommended amplification protocol (PCR program) was modified to suit large DNA fragments (Table 2) and the high fidelity Kapa polymerase (Kapa HiFi Hotstart Readymix, Kapa Biosystems) was used. 12 PCR cycles were conducted with the program in table 2.

Table 2 PCR program for barcoding and library amplification.

Step	Temperature	Duration	
Initialization	98°C	4 minutes	
Denaturation	98°C	30 seconds	× 12
Annealing	65°C	30 seconds	
Elongation	72°C	12 minutes	
Final Elongation	72°C	20 minutes	
Storage	4°C	∞	

The resulting barcoded DNA was purified and size-selected two more times, first with 0.5x AMPure XP beads and then with 0.4x AMPure XP beads. The amount of DNA was quantified with Qubit and quality checked with gel electrophoresis on a 0.8% agarose gel. The samples were then pooled with each pooled sample containing barcoded DNA of either 4 fish (CGN, TU, first library of each wild population) or 8 fish (the remaining libraries of the wild populations).

4.4 NLR capture with hybridization baits

Target enrichment was carried out using the MYbaits v3 customized target enrichment kit for Next Generation Sequencing (MYcroarray, now part of Daicel Arbor Biosciences). The bait set contained nearly 20,000 unique 120-bp biotinylated RNA molecules in equimolar amounts. Most of the baits were designed to specifically bind to the FISNACHT and B30.2 exons in the genome, but we also targeted other genes of interest.

Bait hybridization and target enrichment for each pooled sample was performed according to the MYbaits manual version 3.02, with half the amount of baits and reagents used for the 4-fish pools than for the 8-fish pools. Following an overnight incubation of the pooled DNA samples with RNA baits, bait-bound DNA fragments were extracted from the solution with DB MyOne Streptavidin C1 beads (ThermoFisher Scientific). The enriched libraries were subsequently amplified with P5 and P7 primers (Illumina) by running 26 cycles of the program described in Table 2. If the DNA yield was less than 1,000 ng afterwards (measured by Qubit), five more PCR cycles were added.

Enrichment success was evaluated by qPCR, using 5x HOT FIREPol EvaGreen qPCR Mix Plus (ROX) (Solis BioDyne) and primers specific for the FISNACHT exons from each of the four groups of NLRs (Table 3, Table 4). The gene IL-1 was used as a single copy control. All primers were custom-ordered from biomers.net GmbH. The qPCR experiment was deemed successful if a strong enrichment could be seen for all NLR groups, weaker enrichment for il-1, and no enrichment for the random intron. After this, the sample was selected for subsequent PacBio library construction and purified with 0.7x Ampure PB Beads (Pacific BioSciences).

Table 3 qPCR program for the evaluation of enrichment efficiency.

Step	Temperature	Duration	
Initialization	95°C	12 minutes	
Denaturation	95°C	15 seconds	× 12
Annealing	65°C	20 seconds	
Elongation	72°C	20 minutes	

4.5 Library construction and sequencing

The final libraries were prepared with the SMRTbell Template Prep Kit 1.0-SPv3 (Pacific BioSciences). At the ligation step, the recommended amount of PacBio

Table 4 Sequences of the qPCR primers used for evaluation of target enrichment.

Gene	Direction	Sequence
il1	+	5'-tgg-tga-acg-tca-tca-tcg-cc-3'
il1	-	5'-tcc-agc-acc-tct-ttt-tct-cca-a-3'
foxo6 intron	+	5'-agt-tct-gtg-tgg-gaa-cag-gg-3'
foxo6 intron	-	5'-gtg-cat-ctt-tag-cgt-tgg-ct-3'
NLR group 1	+	5'-cct-gac-aca-ggt-caa-caa-aac-a-3'
NLR group 1	-	5'-gat-tgt-ctt-ttc-ctt-cag-ccc-ag-3'
NLR group 2	+	5'-tgg-att-ggg-ctg-aag-gga-aa-3'
NLR group 2	-	5'-agg-ttc-agt-cct-tta-gtc-tct-gg-3'
NLR group 3	+	5'-ctg-ctg-gag-gtg-aaa-gat-cag-ac-3'
NLR group 3	-	5'-gat-tgt-tga-gca-gtg-agc-agg-a-3'
NLR group 4	+	5'-tac-ctg-gac-aag-aca-aag-cca-3'
NLR group 4	-	5'-ctc-ctt-ctc-ttc-agc-cca-gtc-3'

adapters was increased from 1 to 5 µl per 40 µl total reaction volume and the reaction was incubated over night at room temperature. For the SN and CHT libraries in pools of eight (see table 5), barcoded PacBio adapters were used instead of regular ones. The product codes for barcodes were BC1001 and BC1002 for CHT, BC 1003 and BC1004 for SN.

The first libraries (TU, CGN, 4 DP and 4 KG samples) were size selected to 2-8 kb with the BluePippin pulsed field electrophoresis system (Sage Science). The following libraries were size selected to 1.5-8kb.

All sequencing was done at the Max Planck Genome Centre, located at the Max Planck Institute for Plant Breeding in Cologne. All TU, CGN, DP and KG zebrafish, as well as 4 CHT and 4 SN samples were sequenced with 1M v2 SMRT Cells of the Sequel instrument (Pacific BioSciences). The rest of the samples (all with barcoded adapters) were multiplexed together and sequenced with an 8M SMRT Cell of the much higher throughput Sequel 2 instrument (Pacific BioSciences). One of the already sequenced SN samples (SN24) was also resequenced in this run, as it yielded no data in the first one. Furthermore, Pacific BioSciences upgraded their kits with a superior polymerase after we had sequenced TU, CGN and the first 4 samples of each wild population; all samples other than those were sequenced with their LR (Long Run) polymerase.

An overview of the sequencing scheme is presented in (Table 5).

Table 5 Sequencing scheme for the zebrafish samples. Library sequenced after the introduction of an improved (Long Run) sequencing chemistry are marked with LR. Samples that yielded no data after sequencing are marked with asterisks

Individuals	Library	Sequencing Technology
TU01, TU02, TU03, TU06	TU L1	PacBio Sequel (SMRT Cell 1M v2)
TU08, TU10, TU12, TU14	TU L2	PacBio Sequel (SMRT Cell 1M v2)
CGN1, CGN2, CGN3, CGN4	CGN L1	PacBio Sequel (SMRT Cell 1M v2)
CGN5, CGN6, CGN7, CGN8	CGN L2	PacBio Sequel (SMRT Cell 1M v2)
DP07, DP09, DP10, DP12	DP L1	PacBio Sequel (SMRT Cell 1M v2)
DP15, DP20, DP23, DP24, DP25, DP28, DP31, DP34	DP L2	PacBio Sequel (SMRT Cell 1M v2, LR)
DP03, DP05, DP13, DP16, DP21, DP29, DP31, DP33	DP L3	PacBio Sequel (SMRT Cell 1M v2, LR)
KG35, KG41, KG42, KG43	KG L1	PacBio Sequel (SMRT Cell 1M v2)
KG03, KG05, KG07, KG12, KG14, KG15, KG18, KG19	KG L2	PacBio Sequel (SMRT Cell 1M v2, LR)
KG20, KG22, KG24, KG26, KG29, KG32, KG33, KG44	KG L3	PacBio Sequel (SMRT Cell 1M v2, LR)
SN21, SN23, (SN24*), SN26	SN L1	PacBio Sequel (SMRT Cell 1M v2)
SN03, SN04, SN08, SN09, SN10, SN11, SN12, SN24	SN L2	PacBio Sequel II (SMRT Cell 8M, LR)
SN13, SN14, SN15, SN16, SN17, SN18, SN19, SN20	SN L3	PacBio Sequel II (SMRT Cell 8M, LR)
CHT19, CHT23, CHT26, CHT28	CHT L1	PacBio Sequel (SMRT Cell 1M v2)
CHT01 - CHT07, (CHT13*)	CHT L2	PacBio Sequel II (SMRT Cell 8M, LR)
CHT08, CHT10 - CHT12, CHT14 - CHT16, (SN25*)	CHT L3	PacBio Sequel II (SMRT Cell 8M, LR)

4.6 Read processing and assembly

Raw data were de-multiplexed and stripped of primer/adapter sequences with `lima` from Pacific Biosciences. For the samples sequenced with the PacBio Sequel I, the parameters `--enforce-first-barcode --split-bam-named --W 100` were used with `lima` v1.0.0 for the runs without the LR polymerase. For Sequel runs with the LR polymerase, lima v1.8.0 and v1.9.0 were used with the same parameters. To

remove PacBio barcodes from the data produced on Sequel II, `lima` v1.11.0 was used with parameters `--split-bam-named --peek-guess` and for the subsequent removal of NEBNext barcodes, the parameters were changed to `--enforce-first-barcode --split-bam-named --peek-guess`. Consensus sequences of all DNA fragments with a minimum of three passes (henceforth referred to as CCS reads) were calculated using `ccs` (v4.2.0, Pacific Biosciences) with default parameters. PCR duplicates were identified and flagged with `pbmarkdup` (v1.0.0, Pacific Biosciences) with default parameters, then excluded from downstream analyses. Any chimeric reads containing a primer sequence in the middle were identified with `blastn` (v2.11.0+) [52] and removed. The filtered CCS reads were assembled into contigs for each fish separately using `hifiasm` (v0.15.4-r347) [53] with default parameters.

4.7 NLR identification

In addition to the known NLR genes extracted from Ensembl (Release 107, [54]), the reference genome was translated in all frames using `transeq` (from EMBOSS:6.6.0.0) [55] and searched for further NLRs using `hmmsearch` from hmmer (v3.2.1) [56] without bias correction and the Hidden Markov Model (HMM) profiles for zf_FISNA-NACHT and zf_B30.2 from [24]. Each position in which the zf_FISNACHT model found a hit with a maximum i-Evalue of $1e-200$ and a minimum alignment length of 500 aa was considered a FISNACHT exon. The filtering thresholds for B30.2 exons were an i-Evalue of $1e-5$ and a minimum alignment length of 150 aa.

To distinguish CCS reads of NLR genes from other CCS reads, the CCS reads of each fish were mapped against the reference genome GRCz11 using `pbmm2` (v1.3.0) [57] with preset ccs. CCS reads which mapped within a known NLR gene or one found with our HMM-based approach were considered NLR reads.

Contigs containing NLR exon sequences were identified by translating all contigs of each fish in all frames with `transeq` (from EMBOSS:6.6.0.0) and subsequently

searching for FISNACHT and B30.2 domains using `hmmsearch` from hmmer (v3.2.1) without bias correction and the Hidden Markov Models (HMMs) zf_FISNA-NACHT and zf_B30.2 again. The HMM-based approach was chosen for the contigs in particular because we assumed that there would be NLR sequences in the data which are absent in the reference genome and therefore might not be mapped. The approach enabled us to include all FISNACHT and B30.2 exon data found by [24] in our searches, optimizing the search sensitivity.

The 47 bp exon preceding B30.2 was first found via manual inspection and blasted against the reference genome. The genes found with this approach were characterized based on the domains found in the reference annotation from Ensembl (see Supplementary Table 1).

B30.2 exons from NLRs were distinguished from B30.2 elsewhere in the genome by generating a Hidden Markov Model for the 47-bp exon based on the blast hits and searching the contigs for matches to this model with `hmmsearch` from hmmer (v3.2.1). The model was created from all sequences of this exon in the reference genome (GRCz11) found in biomart, that had been aligned by `mafft` (v7.505 [58] with automatically set parameters. The HMM for the motif was created with `hmmbuild` from hmmer (v3.2.1). The alignment was translated with Jalview (v2.11.2.5) [59].

4.8 Pan-NLRome clustering

The FISNACHT and B30.2-containing contigs were each sorted into orthoclusters using `get_homologues` (build x86 64-20220516) [60] with the orthoMCL algorithm running under Perl (v5.36.0). Whenever an orthocluster contained more than one contig, a consensus sequence for the cluster was created from all those contigs with cons from EMBOSS:6.6.0.0. These consensus sequences and contig sequences of singleton clusters made up the representative orthocluster sequences. Some representative sequences were reversed with `revseq` from EMBOSS:6.6.0.0 so that all exons were in the same

orientation. The representative sequences were then blasted against each other using `blastn` (v2.12.0+) with default parameters. In cases in which at least 3 kb of a representative sequence matched another with at least 98% identity, the two clusters they represented were fused into a new cluster by combining their contigs and generating a new consensus sequence from them. This process was conducted twice and reduced the number of FISNACHT clusters from the initial 4,743 to 2,008 and B30.2 clusters from 14,879 to 2,635.

Next, the filtered CCS reads of each fish were mapped to the orthocluster consensus sequences using `pbmm2` (v1.3.0) with a minimum sequence identity of 95% in the alignment. `samtools` (v1.7) [61] was used to analyse and further filter the bam files. If the representative sequence had at least one primary alignment with length ¿ 1 kb, mapping quality 60 and no more than 9 soft-clipping bases at both ends of the mapped read, the orthocluster was considered to be present in the respective fish. Orthoclusters for which the consensus did not occur in any fish were excluded from further analyses.

FISNACHT exons were assigned to the four NLR groups based on the sequence of their Walker A motifs as described before in [23] (Table 6). For genomic localization, orthocluster representative sequences were mapped to the reference genome GRCz11 with pbmm2 (commit 1.3.0) and a mapping quality threshold of 60.

Table 6 Characteristic sequences of the Walker A motif for different zebrafish NLR-C groups.

NACHT Group	Characteristic Sequence
Group 1	GIAGVGKT
Group 2	GVAGIGKS
Group 3	GIAGIGKT
Group 4	GVAGIGKT

4.9 Modelling the size of zebrafish NLRomes

To estimate the actual sizes of the population-specific pan-NLRomes, we modelled the increase in the total number of identified FISNACHT domains (as a proxy for the number of NLR genes) when adding another individual to a set of already analysed individuals (as shown in [62]). Briefly, given a sample of n individuals, there are

$$w_n(x) = \frac{n!}{(x-1)!(n-x)!} = x\binom{n}{x} \tag{1}$$

ways to choose $x-1$ individuals from the entire sample and add another - not yet chosen - one. For each x we calculated the average increase (filled squares and circles in Figure 6C1 and C2) in the number of identified FISNACHT domains over all possible choices of individuals. Based on these numbers we fitted the non-linear function

$$y = \alpha H(x, \beta)\,, \tag{2}$$

where $H(x, \beta)$ is the generalized harmonic number with parameter β, i.e.

$$H(x, \beta) = \sum_{k=1}^{x} \frac{1}{k^{\beta}} \tag{3}$$

in which $\beta = 1$ is equivalent to the expected number of segregating sites in a standard coalescent (Equation 4, Table 7). Here, we may reinterpret lineage splits as duplication events and a parameter $\beta > 1$ reflecting an increase of the duplication rate towards the present. Importantly, if $\beta > 1$, the series in equation (3) converges, and may be interpreted as the limiting number of NLR genes in a closed NLRome. The NLRome would be open, if $\beta \leq 1$.

4.10 Variant calling

Variants in each fish were called on the bam files containing the alignments of NLR-filtered CCS reads to the orthocluster consensus sequences using `deepvariant` (r1.0)

Table 7 Values of fitted parameters and saturation limits for FISNACHT and NLR-B30.2 domains by population.

Population	FISNACHT			NLR-B30.2		
-	α	β	**Limit**	α	β	**Limit**
TU	178.274	1.43356	519.548	53.8579	1.40774	164.73
CGN	257.207	1.62786	569.367	78.7156	1.61283	177.246
DP	309.14	1.01231	25284	69.3609	0.87454	∞
KG	436.761	1.2152	2288.41	145.715	1.1418	1113.23
SN	479.892	1.26093	2152.12	145.548	1.10183	1514.35
CHT	416.712	1.18893	2451.81	135.677	1.11911	1218.54

[63] with the PacBio model. Joint genotyping of the individual samples was done with `glnexus` (v1.2.7-0-g0e74fc4) [64] with its deepvariant-specific setting. θ_π of the NLR exons was calculated by filtering the VCF files for variants within the exon regions and employing `vcftools --site-pi` (v0.1.16) [65]. θ_w was calculated via

$$\theta_w = \frac{S}{h_{n\text{-}1} l} \tag{4}$$

where S is the number of segregating sites in an exon, $h_{n\text{-}1}$ is the n - 1st harmonic number of the sum of haplotypes and l is the length of the exon (1761 was used for FISNACHT and 540 for B30.2).

4.11 Data visualization

Circular genome plots were created with `circos` (v 0.69-8) [66] running on Perl 5.036000. DNA and protein sequence logos were created with WebLogo (v3.7.12) [67]. Other plots, heat maps and Venn diagrams were created in RStudio (v2022.07.2) with R (v4.2.1) using `ggplot2` (v3.3.6) and `ggvenn` (v0.1.9) or `xmgrace` (v5.1.25; https://plasma-gate.weizmann.ac.il/Grace/). For the heat map in Figure 5, only the 1,253 FISNACHT clusters in which at least one individual had 10 reads mapped to the representative sequence were shown for better visibility. Final processing of the images was done in Inkscape (v1.1.2; https://inkscape.org/).

4.12 Data availability

NLR reads are available in the NCBI Sequence Read Archive (SRA), BioProject PRJNA966920.

4.13 Code availability

Scripts are available on GitHub (https://github.com/YSchaefer/pacbio_zebrafish).

Acknowledgments

The authors are thankful to Emilia Martins and Anuradha Bhat for their contributions to wild sample collection, and Cornelia Stein and the lab of Mathias Hammerschmidt for laboratory zebrafish. We acknowledge the contributions of Philipp Schiffer (helping with writing the initial project proposal), Lisa Vogelsang (assistance with lab work), Robert Fürst and Anna Rottmann (management of computational infrastructure). The authors are grateful to Bruno Hüttel, who provided advice and support during library construction and sequenced all samples at the Max Planck Genome Centre. This work was funded by a grant to TW and ML in the frame of the priority program SPP1819 of the German Research Council (DFG). JS was additionally funded by an NSERC grant to Colin Garroway, University of Manitoba, MB, Canada.

Author contributions

JS co-designed and coordinated the study, prepared sequence libraries from laboratory zebrafish, contributed to computational analyses and co-wrote the manuscript. YS did the majority of computational analyses and co-wrote the manuscript. KP extracted DNA and prepared sequence libraries from wild zebrafish. ML co-designed the study. TW co-designed the study, did the modelling and provided constructive feedback throughout the computational analyses. ARW obtained wild zebrafish samples. All authors contributed to the final manuscript with their commentaries and feedback.

Competing interest declaration

The authors declare no competing interest.

Supplementary information

Supplementary Table 1

References

[1] Bergelson, J., Kreitman, M., Stahl, E. A. & Tian, D. Evolutionary dynamics of plant R-genes. *Science* **292**, 2281–2285 (2001).

[2] Stahl, E. A., Dwyer, G., Mauricio, R., Kreitman, M. & Bergelson, J. Dynamics of disease resistance polymorphism at the Rpm1 locus of Arabidopsis. *Nature* **400**, 667–671 (1999).

[3] McConnell, S. C., Restaino, A. C. & de Jong, J. L. Multiple divergent haplotypes express completely distinct sets of class I MHC genes in zebrafish. *Immunogenetics* **66**, 199–213 (2014).

[4] Almeida-da Silva, C. L. C., Savio, L. E. B., Coutinho-Silva, R. & Ojcius, D. M. The role of NOD-like receptors in innate immunity. *Frontiers in Immunology* **14**, 1122586 (2023).

[5] Jones, J. D. & Dangl, J. L. The plant immune system. *Nature* **444**, 323–329 (2006).

[6] Uehling, J., Deveau, A. & Paoletti, M. Do fungi have an innate immune response? An NLR-based comparison to plant and animal immune systems. *PLOS Pathogens* **13**, e1006578 (2017).

[7] Gao, L. A. *et al.* Prokaryotic innate immunity through pattern recognition of conserved viral proteins. *Science* **377**, eabm4096 (2022).

[8] Witek, K. *et al.* Accelerated cloning of a potato late blight–resistance gene using renseq and SMRT sequencing. *Nature Biotechnology* **34**, 656–660 (2016).

[9] Van de Weyer, A.-L. *et al.* A species-wide inventory of NLR genes and alleles in Arabidopsis thaliana. *Cell* **178**, 1260–1272 (2019).

[10] Urbach, J. M. & Ausubel, F. M. The NBS-LRR architectures of plant R-proteins and metazoan NLRs evolved in independent events. *Proceedings of the National Academy of Sciences* **114**, 1063–1068 (2017).

[11] Huang, S. *et al.* Genomic analysis of the immune gene repertoire of amphioxus reveals extraordinary innate complexity and diversity. *Genome Research* **18**, 1112–1126 (2008).

[12] Hibino, T. *et al.* The immune gene repertoire encoded in the purple sea urchin genome. *Developmental Biology* **300**, 349–365 (2006).

[13] Noel, B. *et al.* Pervasive tandem duplications and convergent evolution shape coral genomes. *Genome Biology* **24**, 1–38 (2023).

[14] Yuen, B., Bayes, J. M. & Degnan, S. M. 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**, 106–120 (2014).

[15] Tørresen, O. K. *et al.* Genomic architecture of haddock (Melanogrammus aeglefinus) shows expansions of innate immune genes and short tandem repeats. *BMC Genomics* **19**, 1–17 (2018).

[16] Suurväli, J., Garroway, C. J. & Boudinot, P. Recurrent expansions of B30.2-associated immune receptor families in fish. *Immunogenetics* **74**, 129–147 (2022).

[17] Lilue, J. *et al.* Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci. *Nature Genetics* **50**, 1574–1583 (2018).

[18] Rajendran, K. *et al.* Pathogen recognition receptors in channel catfish: I. identification, phylogeny and expression of NOD-like receptors. *Developmental & Comparative Immunology* **37**, 77–86 (2012).

[19] Li, J., Chu, Q. & Xu, T. A genome-wide survey of expansive NLR-C subfamily in miiuy croaker and characterization of the NLR-B30.2 genes. *Developmental & Comparative Immunology* **61**, 116–125 (2016).

[20] Chen, Z., Xu, X., Wang, J., Zhou, Q. & Chen, S. A genome-wide survey of NOD-like receptors in chinese tongue sole (Cynoglossus semilaevis): Identification, characterization and expression analysis in response to bacterial infection. *Journal of Fish Biology* **99**, 1786–1797 (2021).

[21] Stein, C., Caccamo, M., Laird, G. & Leptin, M. Conservation and divergence of gene families encoding components of innate immune response systems in zebrafish. *Genome Biology* **8**, 1–23 (2007).

[22] Laing, K. J., Purcell, M. K., Winton, J. R. & Hansen, J. D. A genomic view of the NOD-like receptor family in teleost fish: identification of a novel NLR subfamily in zebrafish. *BMC Evolutionary Biology* **8**, 1–15 (2008).

[23] Howe, K. *et al.* Structure and evolutionary history of a large family of NLR proteins in the zebrafish. *Open Biology* **6**, 160009 (2016).

[24] Adrian-Kalchhauser, I. *et al.* The round goby genome provides insights into mechanisms that may facilitate biological invasions. *BMC Biology* **18**, 1–33 (2020).

[25] Anderson, J. L. *et al.* Multiple sex-associated regions and a putative sex chromosome in zebrafish revealed by RAD mapping and population genomics. *PLOS ONE* **7**, e40701 (2012).

[26] McConnell, S. C., Hernandez, K. M., Andrade, J. & de Jong, J. L. Immune gene variation associated with chromosome-scale differences among individual zebrafish genomes. *Scientific Reports* **13**, 7777 (2023).

[27] Van Der Aa, L. M. *et al.* A large new subset of TRIM genes highly diversified by duplication and positive selection in teleost fish. *BMC Biology* **7**, 1–23 (2009).

[28] Chang, M. X., Xiong, F., Wu, X. M. & Hu, Y. W. The expanding and function of NLRC3 or NLRC3-like in teleost fish: Recent advances and novel insights. *Developmental & Comparative Immunology* **114**, 103859 (2021).

[29] Chuphal, B., Rai, U. & Roy, B. Teleost NOD-like receptors and their downstream signaling pathways: A brief review. *Fish and Shellfish Immunology Reports* 100056 (2022).

[30] Kuri, P. *et al.* Dynamics of in vivo ASC speck formation. *Journal of Cell Biology* **216**, 2891–2909 (2017).

[31] Li, J.-y. *et al.* Characterization of an NLRP1 inflammasome from zebrafish reveals a unique sequential activation mechanism underlying inflammatory caspases in ancient vertebrates. *The Journal of Immunology* **201**, 1946–1966 (2018).

[32] FAO. AQUASTATS: Rivers of South and East Asia. URL https://data.apps.fao.org/catalog/dataset/dc2a5121-0b32-482b-bd9b-64f7a414fa0d. Last accessed 05 April 2023, License: CC BY-NC-SA 3.0 IGO.

[33] Howe, K. *et al.* The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496**, 498–503 (2013).

[34] Suurväli, J. *et al.* The laboratory domestication of zebrafish: from diverse populations to inbred substrains. *Molecular Biology and Evolution* **37**, 1056–1069 (2020).

[35] Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J. & Edwards, D. Plant pangenomes are the new reference. *Nature Plants* **6**, 914–920 (2020).

[36] Sherman, R. M. & Salzberg, S. L. Pan-genomics in the human genome era. *Nature Reviews Genetics* **21**, 243–254 (2020).

[37] Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023).

[38] Whiteley, A. R. *et al.* Population genomics of wild and laboratory zebrafish (danio rerio). *Molecular Ecology* **20**, 4259–4276 (2011).

[39] Yue, J.-X., Meyers, B. C., Chen, J.-Q., Tian, D. & Yang, S. Tracing the origin and evolutionary history of plant nucleotide-binding site–leucine-rich repeat (NBS-LRR) genes. *New Phytologist* **193**, 1049–1063 (2012).

[40] Lee, R. R. & Chae, E. Variation patterns of NLR clusters in Arabidopsis thaliana genomes. *Plant Communications* **1**, 100089 (2020).

[41] Otto, M., Zheng, Y. & Wiehe, T. Recombination, selection, and the evolution of tandem gene arrays. *Genetics* **221**, iyac052 (2022).

[42] Samukawa, S., Yoshimi, R., Kirino, Y. & Nakajima, H. The PRY/SPRY domain of pyrin/TRIM20 interacts with β2-microglobulin to promote inflammasome formation. *Scientific Reports* **11**, 23613 (2021).

[43] Higgs, R. *et al.* The E3 ubiquitin ligase Ro52 negatively regulates IFN-β production post-pathogen recognition by polyubiquitin-mediated degradation of IRF3. *The Journal of Immunology* **181**, 1780–1786 (2008).

[44] Chae, J. J. *et al.* The B30.2 domain of pyrin, the familial mediterranean fever protein, interacts directly with caspase-1 to modulate IL-1β production. *Proceedings of the National Academy of Sciences* **103**, 9982–9987 (2006).

[45] Magadum, S., Banerjee, U., Murugan, P., Gangapur, D. & Ravikesavan, R. Gene duplication as a major force in evolution. *Journal of Genetics* **92**, 155–161 (2013).

[46] Schiffer, P. H., Gravemeyer, J., Rauscher, M. & Wiehe, T. Ultra large gene families: a matter of adaptation or genomic parasites? *Life* **6**, 32 (2016).

[47] Cannon, S. B., Mitra, A., Baumgarten, A., Young, N. D. & May, G. The roles of segmental and tandem gene duplication in the evolution of large gene families in Arabidopsis thaliana. *BMC Plant Biology* **4**, 1–21 (2004).

[48] Hughes, A. L., Friedman, R., Ekollu, V. & Rose, J. R. Non-random association of transposable elements with duplicated genomic blocks in Arabidopsis thaliana. *Molecular Phylogenetics and Evolution* **29**, 410–416 (2003).

[49] Rodríguez-Nunez, I., Wcisel, D. J., Litman, G. W. & Yoder, J. A. Multigene families of immunoglobulin domain-containing innate immune receptors in zebrafish: deciphering the differences. *Developmental & Comparative Immunology* **46**, 24–34 (2014).

[50] Mombaerts, P. Molecular biology of odorant receptors in vertebrates. *Annual Review of Neuroscience* **22**, 487–509 (1999).

[51] Shelton, D. S. *et al.* Collective behavior in wild zebrafish. *Zebrafish* **17**, 243–252 (2020).

[52] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410 (1990).

[53] Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* **18**, 170–175 (2021).

[54] Cunningham, F. *et al.* Ensembl 2022. *Nucleic Acids Research* **50**, D988–D995 (2022).

[55] Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European molecular biology open software suite. *Trends in Genetics* **16**, 276–277 (2000).

[56] Wheeler, T. J. & Eddy, S. R. nhmmer: DNA homology search with profile HMMs. *Bioinformatics* **29**, 2487–2489 (2013).

[57] Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

[58] Katoh, K., Misawa, K., Kuma, K.-i. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Research* **30**, 3059–3066 (2002).

[59] Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M. & Barton, G. J. Jalview version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).

[60] Contreras-Moreira, B. & Vinuesa, P. Get_homologues, a versatile software package for scalable and robust microbial pangenome analysis. *Applied and Environmental Microbiology* **79**, 7696–7701 (2013).

[61] Li, H. *et al.* The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

[62] Medini, D., Donati, C., Rappuoli, R. & Tettelin, H. The pangenome: a data-driven discovery in biology. *The pangenome: Diversity, dynamics and evolution of genomes* 3–20 (2020).

[63] Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology* **36**, 983–987 (2018).

[64] Yun, T. *et al.* Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics* **36**, 5582–5589 (2020).

[65] Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

[66] Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Research* **19**, 1639–1645 (2009).

[67] Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Research* **14**, 1188–1190 (2004).