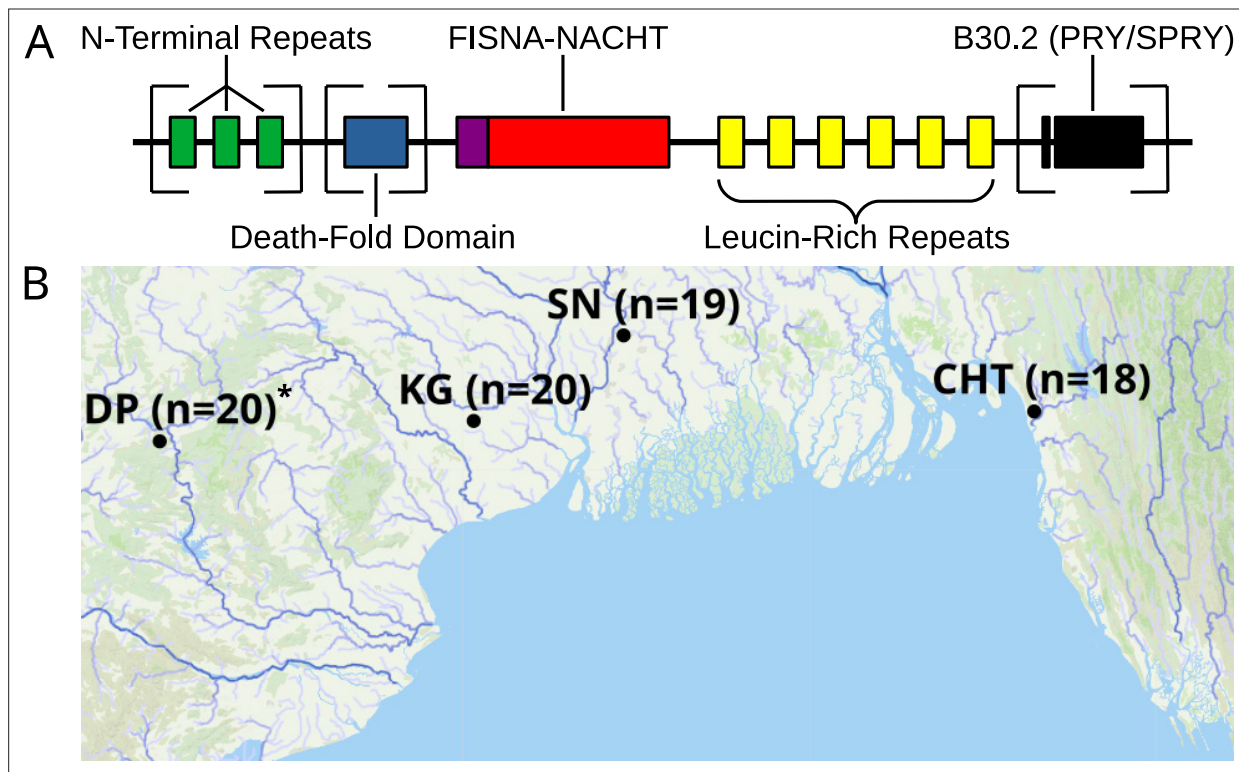


---

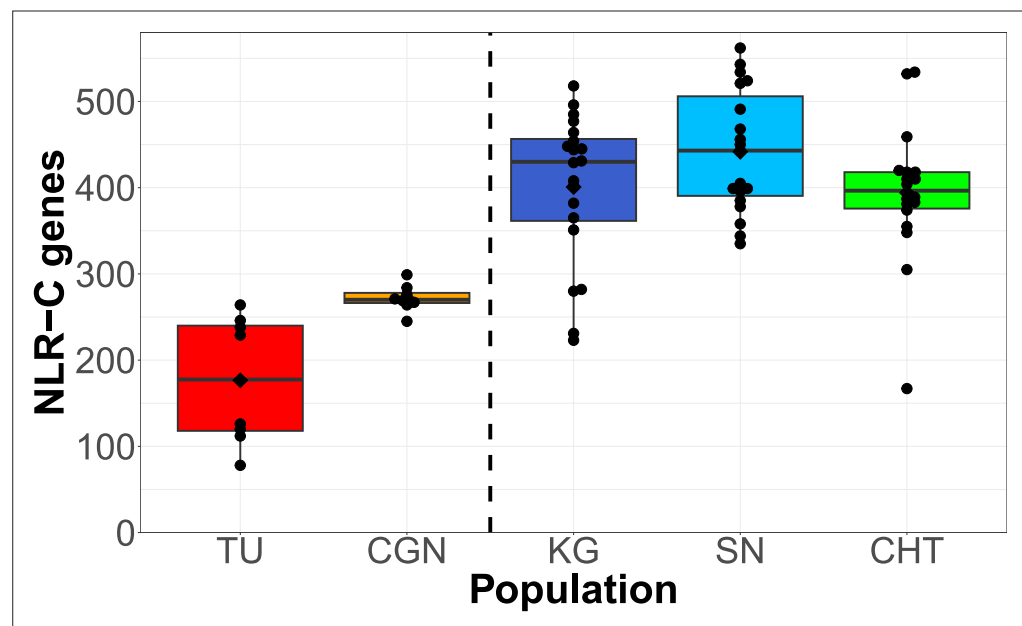
## Figures and figure supplements

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

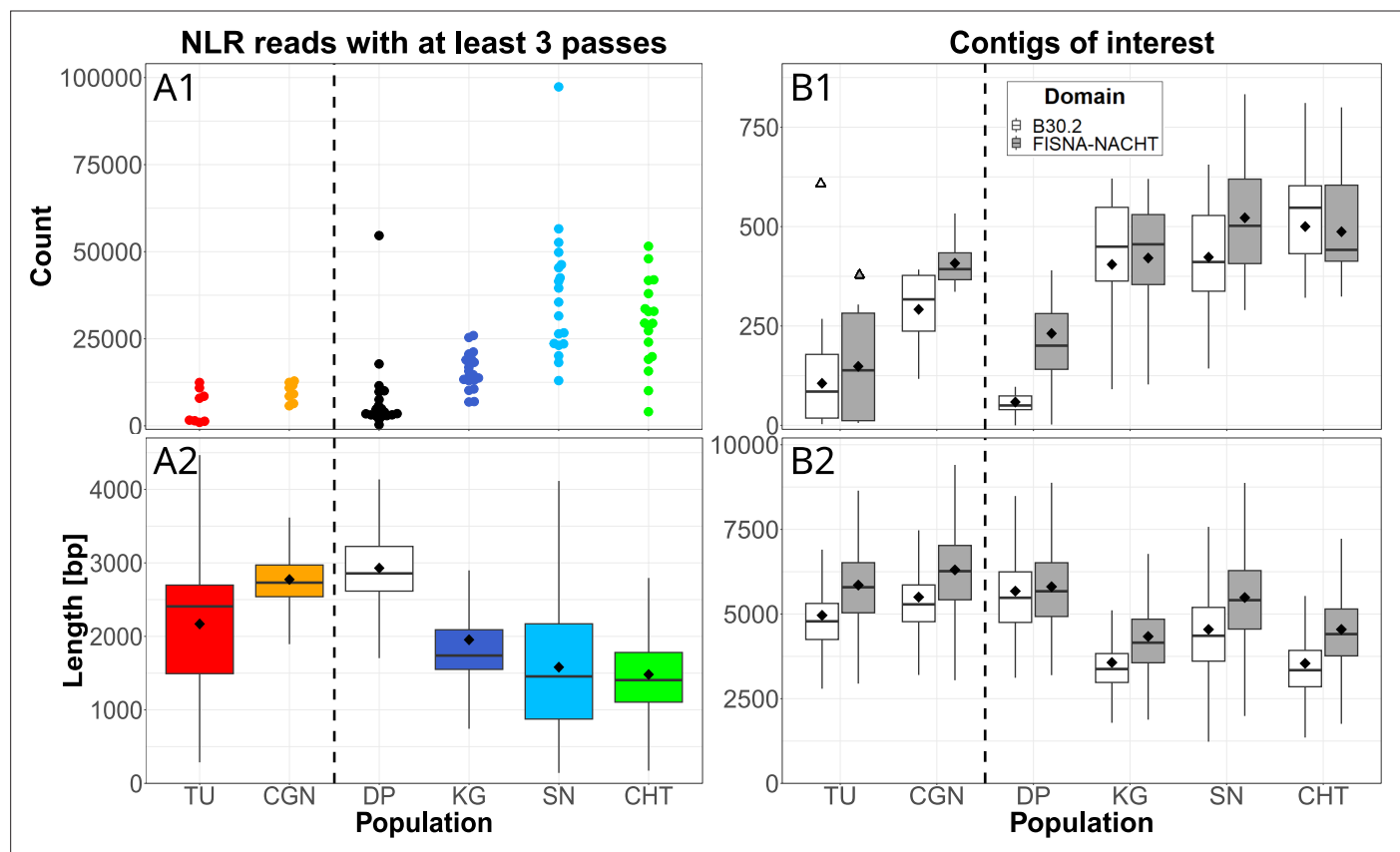
**Yannick Schäfer *et al.***



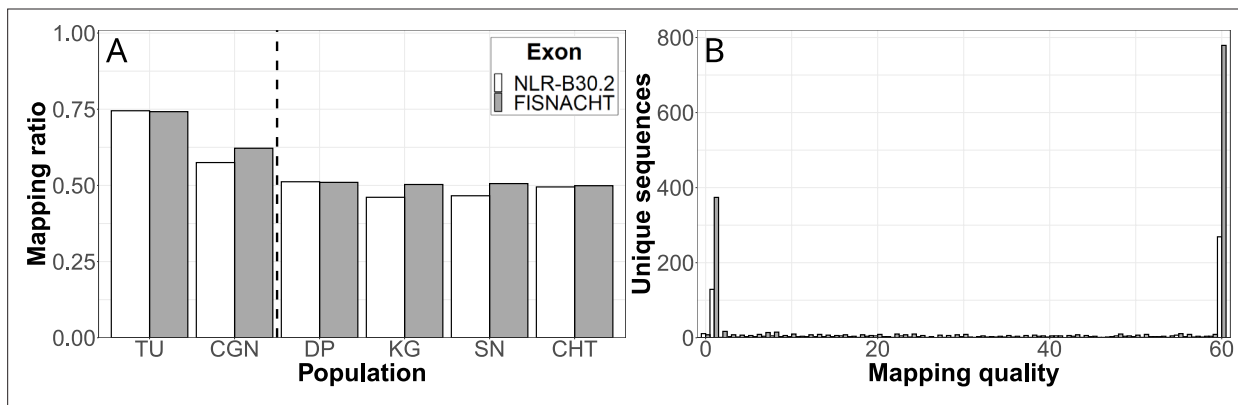
**Figure 1.** Structure of zebrafish NLRs and a map showing the origin of wild zebrafish samples. **(A)** Generalized, schematic representation of the domain architecture of an NLR-C protein. Each box represents a translated exon. The N-terminal repeats, the death-fold domain, as well as the B30.2 domain only occur in subsets of NLR-C genes. The number of N-terminal repeats and leucine-rich repeats can vary. Domains that can be either present or absent in different NLRs are surrounded by square brackets. **(B)** Sampling sites for wild zebrafish. All sites are located near the Bay of Bengal. Final sequenced sample sizes are indicated in parentheses. The map is based on geographic data collected and published by AQUASTAT from the Food and Agriculture Organization of the United Nations (**FAO, 2021**). The population DP is marked with an asterisk because its analysis and results are presented only in figure supplements.



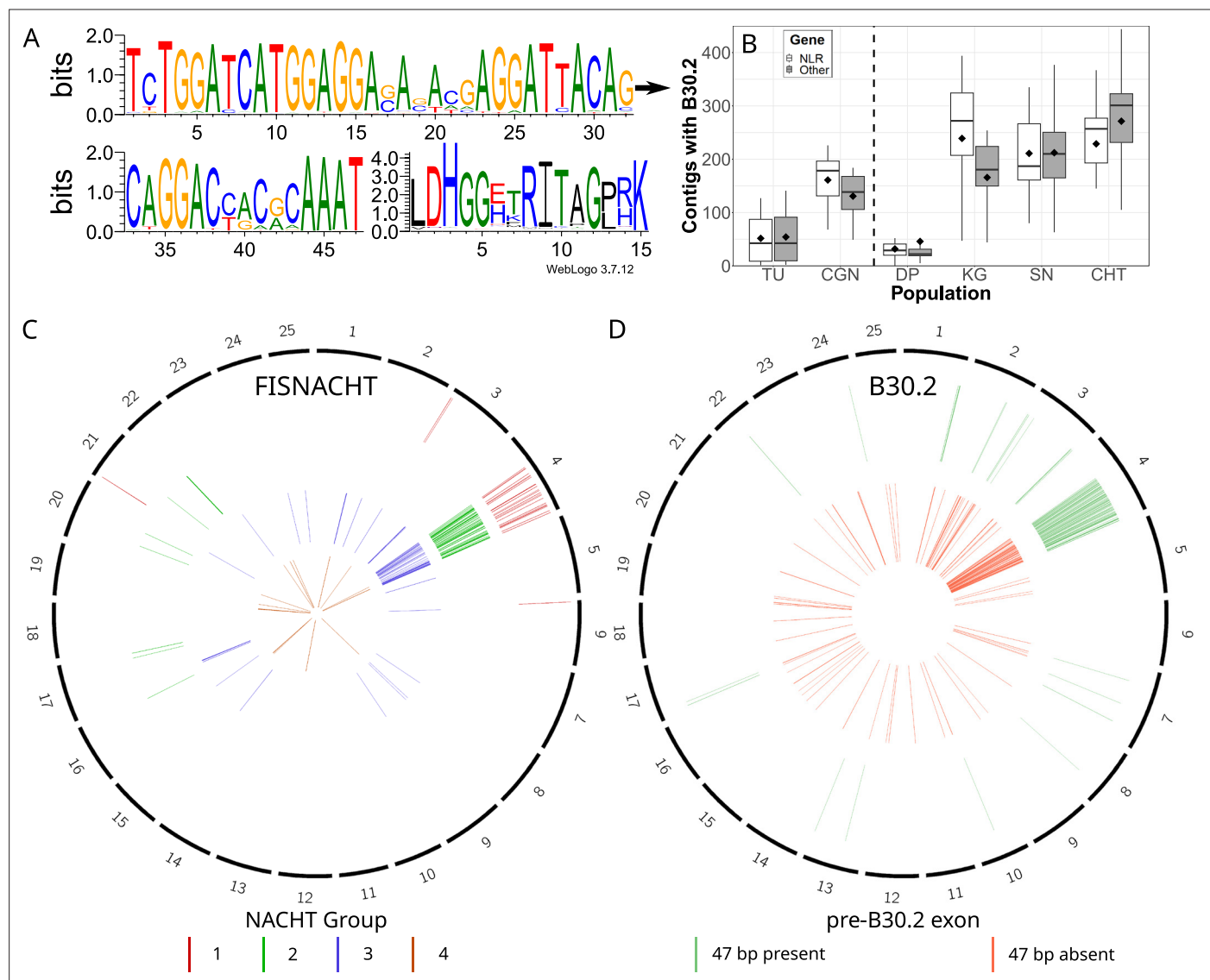
**Figure 2.** Total counts of NLRs found per individual, shown for each population. Black diamonds on the box plots denote means, horizontal lines denote medians. Left side: two laboratory strains; right side: three wild populations.



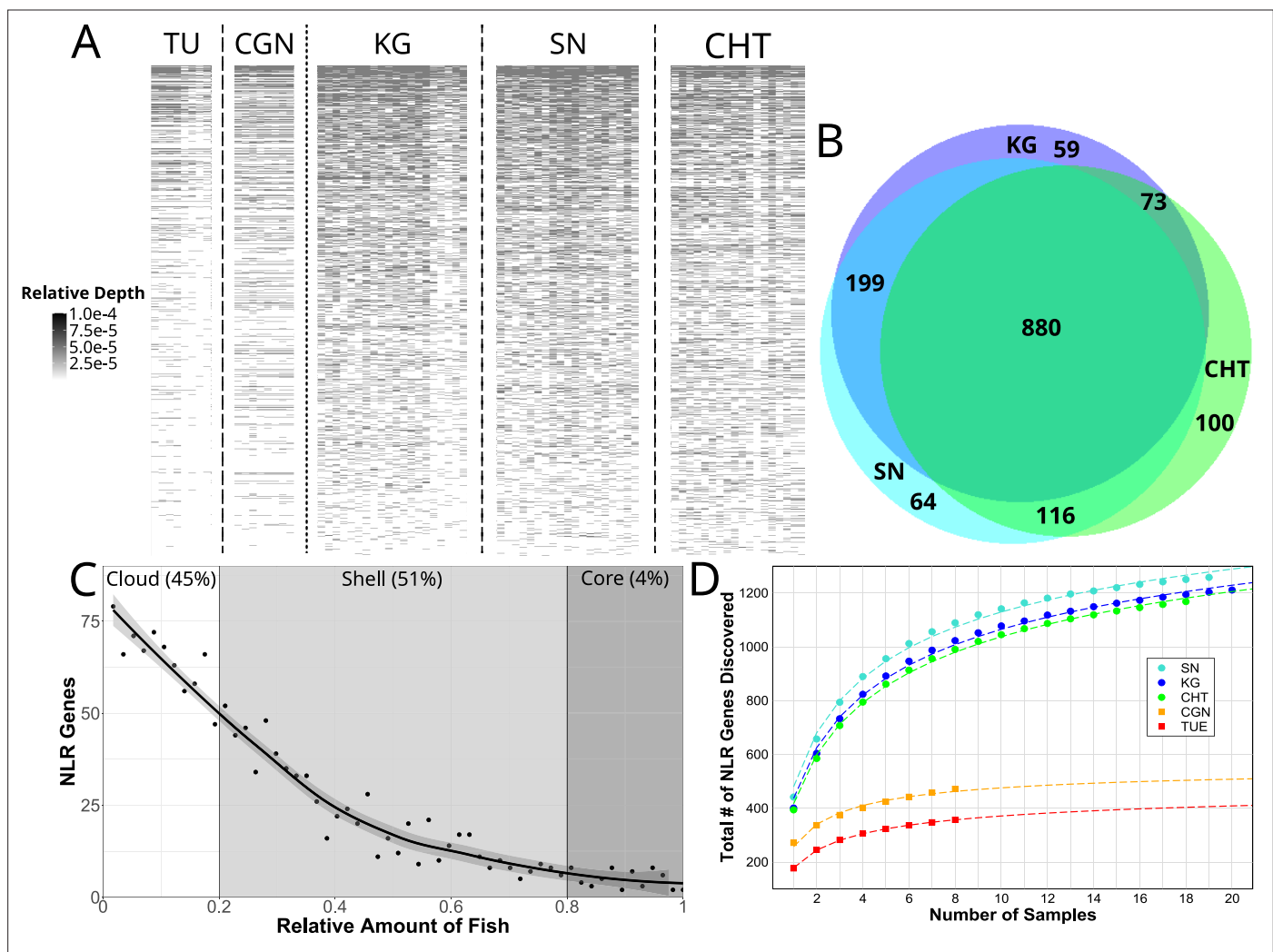
**Figure 2—figure supplement 1.** Sequencing and assembly statistics of circular consensus sequence (CCS) reads from NLR exons. **(A1)** Absolute numbers of CCS reads from NLR exons per sequenced individual. **(A2)** Lengths of the CCS reads that map to NLR genes. **(B1)** Absolute numbers of assembled contigs containing an NLR exon per sequenced individual. Triangles above TU mark the numbers of NLR exons found in the reference genome. **(B2)** Lengths of the individual assembled contigs that contain an NLR exon. Outliers not shown in the boxplots. The black diamonds on boxplots denote means, horizontal black lines denote medians.



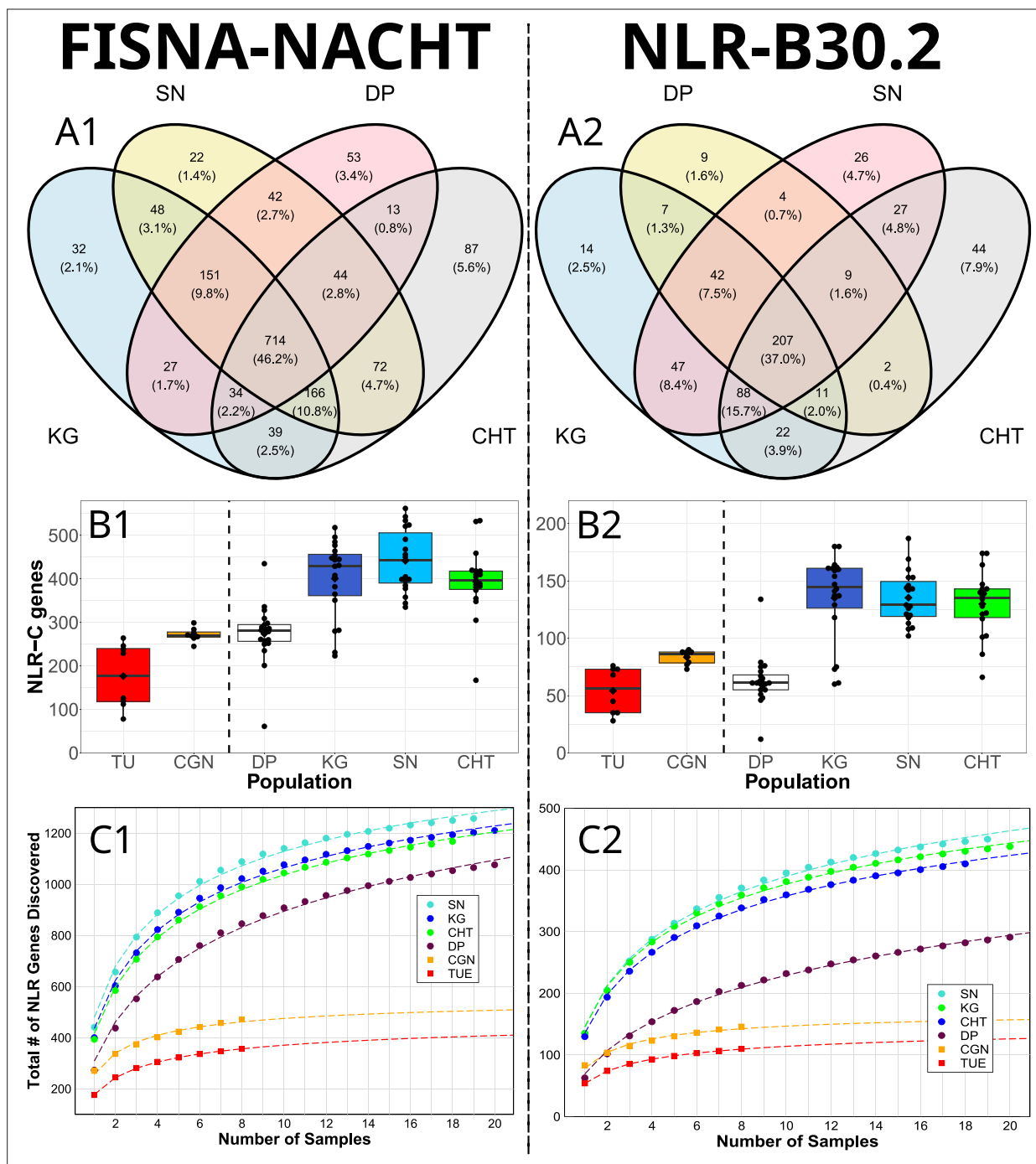
**Figure 2—figure supplement 2.** Assembled NLRs in the reference genome GRCz11. **(A)** Proportions of unique FISNA-NACHT and NLR-B30.2 sequences that were successfully mapped to the reference genome GRCz11 with a mapping quality of 60, by population. **(B)** Distribution of mapping qualities for all unique NLR sequences that aligned to GRCz11, showing that most map either with very high (60) or very low quality.



**Figure 2—figure supplement 3.** Identification of B30.2 domains associated with zebrafish NLRs. **(A)** Nucleotide sequence logo (on top, continued on bottom left) and amino acid logo (on bottom right) for a small, highly conserved 47 bp exon that precedes B30.2 in zebrafish NLRs and not in other genes. The first nucleotide of the exon was removed to generate the correct amino acid translation. Logos were created with Weblogo, the height of each base represents its information content in bits (Crooks et al., 2004). **(B)** Absolute numbers of contigs containing a B30.2 exon per sequenced individual, split by presence/absence of the NLR-specific exon. The black diamonds on boxplots mark the means. **(C)** Genomic distribution of FISNA-NACHT domains in the GRCz11 reference genome. **(D)** Genomic distribution of B30.2 domains in the GRCz11 reference genome.

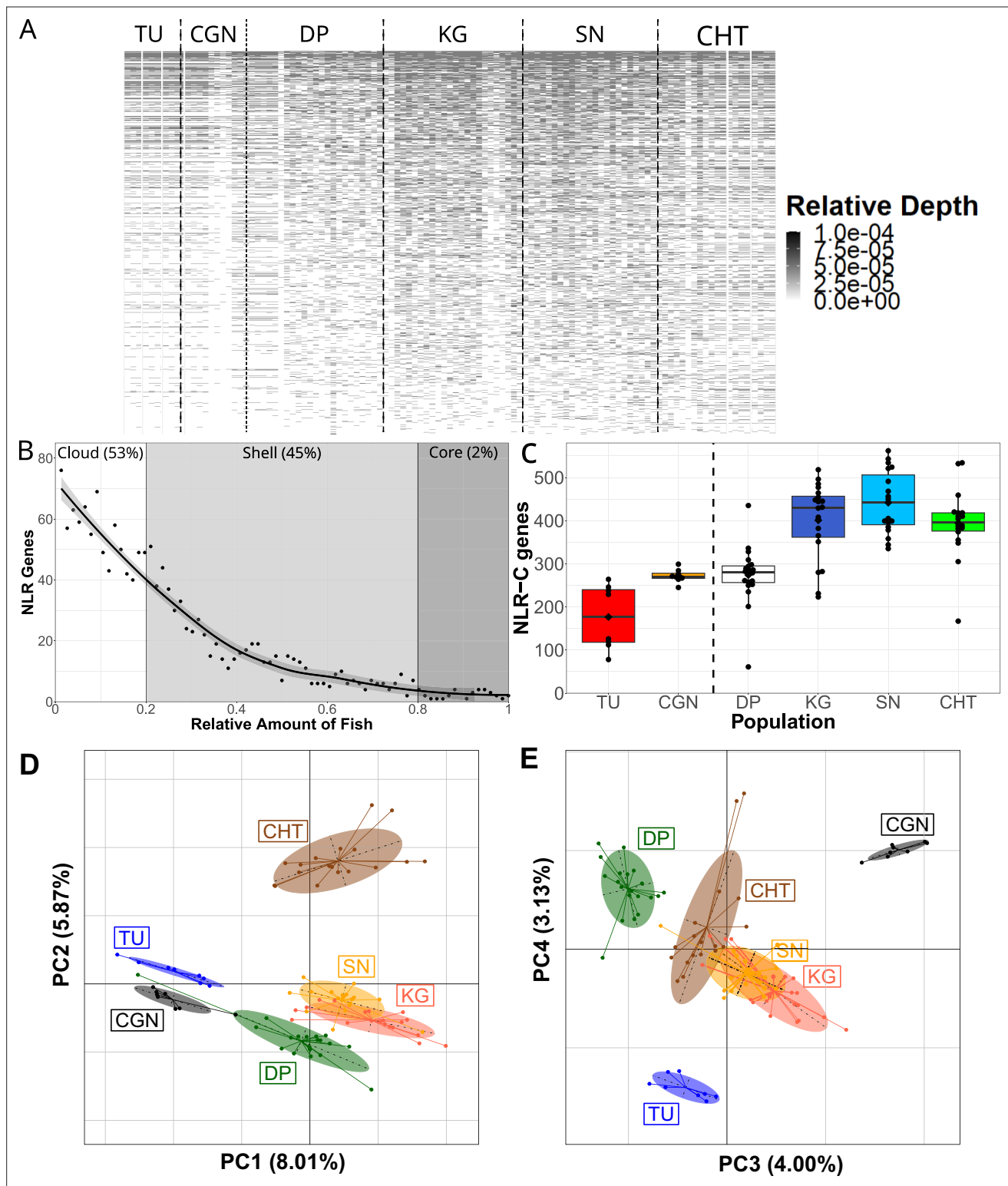


**Figure 3.** Copy number variation of NLR genes. **(A)** Sequence data from each individual zebrafish (vertical axis) was aligned to FISNA-NACHT exon sequences of the pan-NLRome (horizontal axis). Grayscale intensity shows, for each NLR, the proportion of NLR-aligning data in each given fish that matches this specific gene. Darker gray indicates a higher likelihood of this NLR being represented in multiple copies in the particular individual. Light gray indicates a single copy, white indicates absence. For clarity, only the 1235 FISNA-NACHT exons for which at least one fish had a minimum of 10 reads mapped to it are shown. **(B)** Numbers of pan-NLRome sequences (based on FISNACHT diagnosis) found in all three, two, or only one wild population. **(C)** Relative numbers of fish in which pan-NLRome sequences were found in wild populations. ‘Core’ pan-NLRome: genes which are found in at least 80% of the sample (from a total of 57 wild fish); ‘shell’: genes in at least 20%; ‘cloud’: rare genes found in less than 20% of the sample. **(D)** Observed and estimated sizes of population-specific pan-NLRomes. Data points (filled circles and squares) show the average number of totally discovered NLR genes (as identified via their FISNA-NACHT domain) when investigating  $x$  fish. The dashed line is obtained by non-linear fit of the data to the function given in **Equation 2**. For all populations, the hypothetical pan-NLRome size – when extrapolating  $x \rightarrow \infty$  – is finite (see **Table 1**).



**Figure 3—figure supplement 1.** Comparison of copy number variation in FISNA-NACHT and NLR-B30.2 exons. (**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 totally discovered NLR genes (as identified via their FISNA-NACHT domain) in  $x$  individuals. The dotted lines represent the result of non-linear curve fitting (detailed in 'Materials and methods').



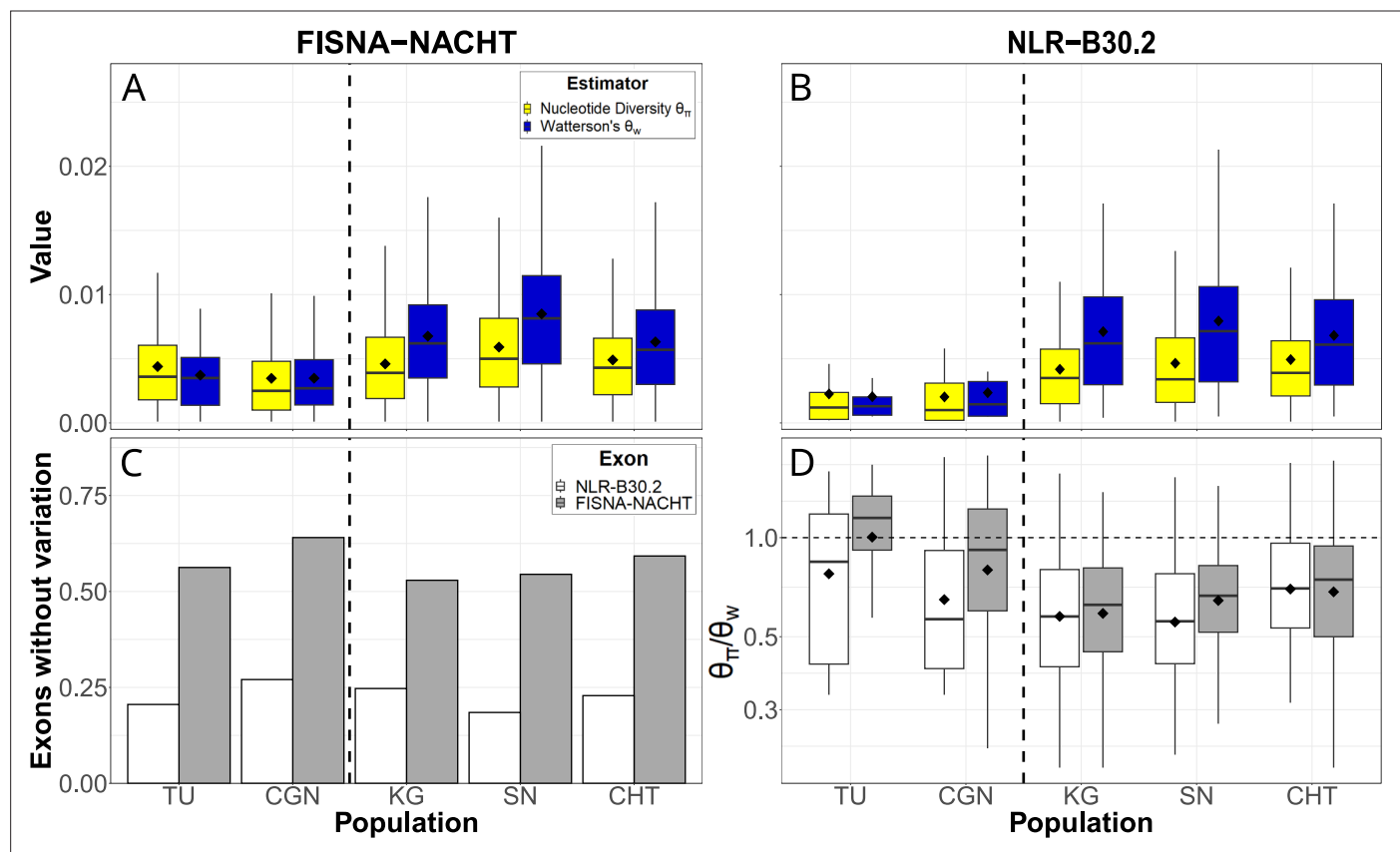


**Figure 3—figure supplement 2.** Copy number variation of NLR genes, including the DP population. **(A)** Sequence data from each individual zebrafish (x-axis) was aligned to FISNA-NACHT exon sequences of the pan-NLRome (y-axis). Grayscale intensity shows, for each NLR, the proportion of NLR-aligning data in each given fish that matches this specific gene. Darker colors can be interpreted as potentially having multiple copies. Lighter colors indicate a single copy, white color means that the sequence was not present. For clarity, only the 1235 FISNA-NACHT for which at least one fish had 10 reads mapped to it are shown. **(B)** Relative numbers of fish in which the pan-NLRome sequences were found in wild populations. Some belong to the core pan-NLRome (in at least 80% of fish), while others are classified as shell (in at least 20% of fish) or cloud (less than 20%). **(C)** Numbers of

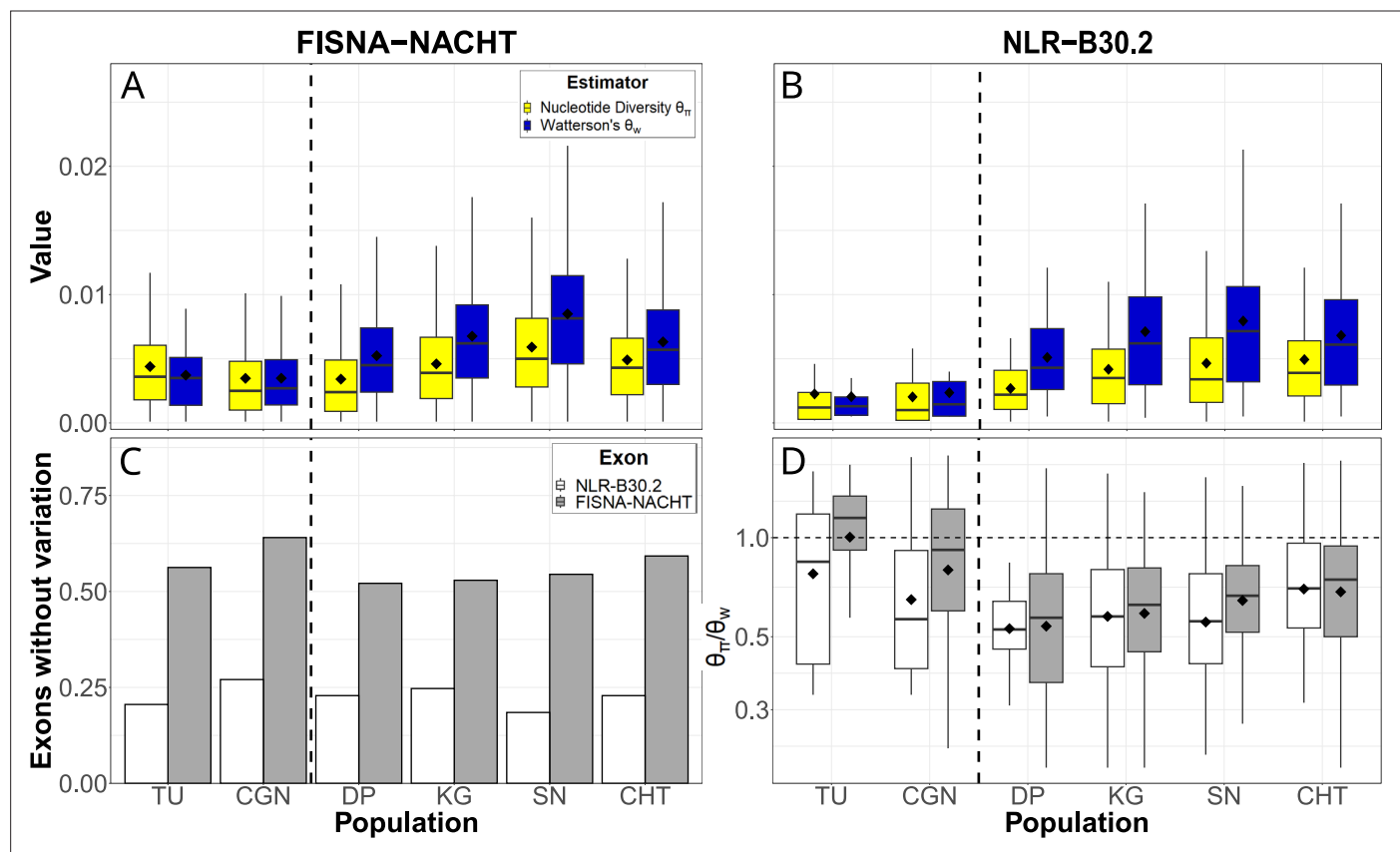
Figure 3—figure supplement 2 continued on next page

*Figure 3—figure supplement 2 continued*

unique NLR sequences (each with one or more copies per individual) found in fish of the sequenced strains. Black diamonds on the box plot denote means, horizontal lines denote medians. **(D, E)** Principal component analysis of scaled-per-individual NLR (FISNA-NACHT) copy numbers. The first two components appear to separate data based on differences between wild and laboratory zebrafish (PC1), and based on geographic distance (PC2).



**Figure 4.** Single-nucleotide variation in NLR exons. Pairwise nucleotide diversity ( $\theta_{\pi}$ ) and Watterson's estimator of the scaled mutation rate ( $\theta_w$ ) for FISNA-NACHT (A) and NLR-associated B30.2 (B) exons. (C) Proportion of exons without any single nucleotide polymorphisms. (D) Ratio of  $\theta_{\pi}/\theta_w$ . Only exons with at least one single-nucleotide polymorphism are shown. The dotted, horizontal line marks a ratio of 1, the expected value under neutrality and constant population size. The black diamonds on box plots denote means, horizontal lines denote medians.



**Figure 4—figure supplement 1.** Single-nucleotide polymorphisms of different NLR exons shown by population, including DP. **(A)** Nucleotide diversity ( $\theta_{\pi}$ ) and Watterson estimator ( $\theta_w$ ) for FISNA-NACHT exons. **(B)** Nucleotide diversity ( $\theta_{\pi}$ ) and Watterson estimator ( $\theta_w$ ) for NLR-associated B30.2 exons. **(C)** Proportion of exons which are completely monomorphic. **(D)** Ratio of  $\theta_{\pi}/\theta_w$ . Only exons with at least one variant are shown. The black, dotted line marks a ratio of 1. The black diamonds on box plots denote means, horizontal lines denote medians.