# Large genetic diversity and strong positive selection in F-box and GPCR genes among the wild isolates of *Caenorhabditis elegans*

Fuqiang Ma[1] and Chaogu Zheng[1,2]

[1]School of Biological Sciences, The University of Hong Kong, Hong Kong SAR, China

[2]Correspondence: cgzheng@hku.hk (C.Z.)

**Short title:** Intraspecific adaptive evolution of F-box and GPCR genes

**Abstract**

The F-box and chemosensory GPCR (csGPCR) gene families are greatly expanded in nematodes, including the model organism *Caenorhabditis elegans*, compared to insects and vertebrates. However, the intraspecific evolution of these two gene families in nematodes remain unexamined. In this study, we analysed the genomic sequences of 330 recently sequenced wild isolates of *C. elegans* using a range of population genetics approaches. We found that F-box and csGPCR genes, especially the *Srw* family csGPCRs, showed much larger diversity than other gene families. Population structure analysis and phylogenetic analysis divided the wild strains into eight non-Hawaiian and three Hawaiian subpopulations. Some Hawaiian strains appeared to be more ancestral than all other strains. F-box and csGPCR genes maintained a large amount of the ancestral variants in the Hawaiian subpopulation and their divergence among the non-Hawaiian subpopulations contributed significantly to population structure. These genes are mostly located at the chromosomal arms and high recombination rate correlates with their large polymorphism. Gene flow might also contribute to their diversity. Moreover, we identified signatures of strong positive selection in the F-box and csGPCR genes in the non-Hawaiian population using both neutrality tests and Extended Haplotype Homozygosity analysis. A large number of high frequency derived alleles in these genes were found in non-Hawaiian population, leading to divergence from the ancestral genotype found in Hawaiian strains. In summary, we found that F-box and csGPCR genes harbour a large pool of natural variants, which may be subjected to positive

selection during the recent selective sweep in non-Hawaiian population. These selected variants may affect protein degradation and sensing of environmental cues, which may result in advantages during adaptation.

## Introduction

*C. elegans* genome contains over 500 F-box genes, compared to ~69 in human genome. This great expansion of the F-box gene family is the result of tandem gene duplication, which have also been observed in plants (Xu et al. 2009). F-box genes code for proteins sharing the F-box domain, a 42-48 amino acid-long motif that binds to Skp1 proteins during the assembly of the SCF (Skp1-Cullin-F-box) E3 ubiquitin ligase complexes, which ubiquitinate protein substrates and target them for degradation (Kipreos and Pagano 2000). F-box genes and the SCF complex-mediated protein degradation have diverse functions in *C. elegans*, including the regulation of lifespan (Ghazi et al. 2007), developmental timing (Fielenbach et al. 2007), sex determination (Jager et al. 2004), and neuronal differentiation (Bounoutas et al. 2009). The role of F-box proteins in the evolution of *Caenorhabditis* species has been noticed before in the study of sex determination and the rise of hermaphroditism. For example, through convergent evolution, *C. elegans* and *C. briggsae* independently evolved the hermaphroditic reproduction system by using two different F-box genes (*fog-2* and *she-1*, respectively) to suppress the translation of *tra-2* mRNA and promote spermatogenesis (Guo et al. 2009). The intraspecific variation of F-box genes and their contribution to adaptation within *C. elegans* have not been studied.

Chemoreception is a major way for the nematodes to sense environmental cues and is mediated by the chemosensory-type seven-transmembrane G-protein-coupled receptors (csGPCRs). The *C. elegans* genome contains more than 1,300 csGPCR genes (Thomas and Robertson 2008), an exceptionally large number given the small size of its nervous system (302 neurons in adult hermaphrodites). The csGPCR genes can be divided into four superfamilies and families (in parentheses): *Str* (*srd*, *srh*, *sri*, *srj*, and *str*), *Sra* (*sra*, *srab*, *srb*, and *sre*), *Srg* (*srg*, *srt*, *sru*, *srv*, *srx*, and *srxa*), and *Solo* (*srw*, *srz*, *srbc*, *srsx*, and *srr*) (Vidal et al. 2018). Evidence of extensive gene duplication and deletion and intron gain and loss were found in the *srh* genes among species in the Caenorhabditis genus, suggesting rapid interspecific evolution (Robertson 2000). Several of the csGPCRs were found to be essential for sensing some odors and pheromones (Sengupta et al. 1996; Kim et al. 2009; Park et al. 2012), but the function of most csGPCRs is unknown. The expansion of the csGPCR gene families and their roles in environmental sensing strongly suggest their involvement in evolution, but the evidence for intraspecific positive selection is missing.

Thanks to the sampling efforts in the past, a collection of ~330 wild isolates of *C. elegans* have been obtained and sequenced (Crombie et al. 2019; Stevens et al. 2019). Their genomic sequences were recently made available (Cook et al. 2017), providing an important resource for understanding the intraspecific evolution of *C. elegans*. Here we analyzed the nonsynonymous single nucleotide variants (SNVs) among the 330 wild isolates of *C. elegans* and compared the nucleotide diversity of genes belonging to

different gene families. We found that the F-box and the csGPCR genes showed much larger diversity than an average gene. Population structure analysis and phylogenetic analysis divided the wild strains into eight non-Hawaiian and three Hawaiian subpopulations. Some Hawaiian strains appeared to be more ancestral than all other strains. F-box and csGPCR genes maintained a large amount of potentially ancestral variant sites in the Hawaiian strains and their divergence among the eight non-Hawaiian groups contributed significantly to population structure. High recombination rate and gene flow might have contributed to the large diversity of these genes, given their location at mostly the chromosomal arms. Furthermore, both neutrality tests and Extended Haplotype Homozygosity analysis identified signs of strong positive selection in the F-box and csGPCR genes; their derived alleles in non-Hawaiian population may have altered gene functions, leading to selective advantages. In summary, our systematic analysis suggests that F-box and csGPCR genes harbour a large pool of natural variants, which were subjected to positive selection during the recent selective sweep and adaptive evolution of the wild *C. elegans* population.

## Materials and Methods

### Population genetic statistics

The VCF file containing 2,493,687 SNVs and 412,448 small indels was downloaded from the CeNDR website (https://www.elegansvariation.org/) (Cook et al. 2017). Variants were annotated by SnpEff (v4.3t) using the Ensemble WBcel235.94 genome assembly. The gene upstream and downstream regions are defined as 5 kb

flanking the CDS bases. Among all SNV sites, we found that 665,368 SNVs in 11199 genes had complete sequencing data in all 330 wild strains (660 alleles) using VCFtools (v0.1.13) (Danecek et al. 2011). This dataset is referred to as "the complete-case dataset". SNVs in the complete-case dataset were then subjected to calculation using DnaSP (Rozas et al. 2017) and PopGenome (Pfeifer et al. 2014). Both software produced similar results for length-normalized nucleotide diversity (*Pi*) and neutrality test statistic Tajima's *D* for the synonymous, nonsynonymous, intron and UTR sites (supplementary fig. S2, Supplementary Material online). Correlation analysis was done in R using Pearson correlation test.

Because the analysis of the complete-case dataset removed 73% of the variant sites, we tested whether similar results can be obtained if we include variant sites with incomplete data. For an average strain, 76,872 (2.6%) out of the total 2,906,135 variant sites did not have high-quality sequencing data, and for an average variant site, 17.5 (2.9%) out of the 660 alleles (330 strains) did not have valid genotype. So, the portion of missing data appears to be small, but stringent complete-case analysis discarded almost three quarters of the SNVs and possibly lost valuable information. To deal with this problem, we used the software VariScan (Hutter et al. 2006), which can set a threshold for the number of alleles containing valid data for a given site. We first translated the VCF formatted file to Hapmap style using Tassel (v5.0) (Bradbury et al. 2007) to facilitate the calculation of *Pi* (NEI 1987), Tajima's *D* (Tajima 1989), Fay and Wu's *H* (Fay and Wu 2000) by VariScan. We then tested the threshold (NumNuc) at

200 and 450, which qualified sites with more than 200 and 450 alleles, respectively. These two conditions included 253,600 and 235,283 nonsynonymous variants covering 18,797 and 18,643 genes, respectively, as compared to the complete-case dataset that contains only 85,260 sites covering only 9,948 genes. Preserving more SNVs made the CDS length-normalized *Pi* bigger, but Tajima's *D* and Fay and Wu's *H* appeared similar between the results obtained using the complete-case dataset and the conditioned full dataset (supplementary fig. S2, Supplementary Material online). Thus, the inclusion of variant sites with a few missing data points did not affect the results of neutrality test but added significant amount of genetic diversity data. Thus, for some analysis, we opted to use the full dataset that qualifies all sites with >200 valid alleles (referred to simply as "the full dataset").

To assess the significance of the *D* and *H* values, we performed coalescent simulations for each gene based on the number of segregating sites using DnaSP v5. Confidence interval was set as 95% and the number of replicates was 1000.

**Population structure analysis**

We first used PLINK (v1.9) (Purcell et al. 2007) to convert the VCF file containing 2,493,687 SNVs to a PED formatted file, which were then subjected to analysis using ADMIXTURE with the number of subpopulation (K value) ranging from 2 to 15. The cross-validation (CV) error for K=11 is the smallest (supplementary fig. S3, Supplementary Material online). The population structure was visualized using pophelper web apps (v1.0.10) (Francis 2017). The 11 ancestral groups are: Europe_1,

Europe_2, Europe_3, Europe_4, Europe_5, Europe_6, Hawaii_1, Hawaii_2, Hawaii_3, North America and Australasia. 266 strains out of the 330 wild strains showed an ancestral proportion above 0.5, indicating clear lineage history. Strain information was downloaded from CeNDR website (https://www.elegansvariation.org/strain/catalog).

**Grouping of Hawaiian and non-Hawaiian strains**

We grouped the 330 wild isolates into Hawaiian and non-Hawaiian populations based on genetic difference in population structure instead of geographic locations (supplementary table S4, Supplementary Material online). The Hawaiian population contains 45 strains carrying a dominant lineage (admixing proportion > 0.5) from "Hawaii_1" (10 strains), "Hawaii_2" (10 strains), and "Hawaii_3" (25 strains). The rest 285 strains were grouped as the non-Hawaiian population. 64 of the 285 strains did not have a dominant ancestral lineage and contained extensive admixing among mostly the eight non-Hawaiian ancestral subpopulations. So, they were included in the non-Hawaiian population. The 45 strains in the Hawaiian population were all extracted from Hawaiian Islands except five strains (ECA36, JU3226, QX1211, ECA593, and XZ2019), and five strains that were extracted from Hawaiian Islands were included in the non-Hawaiian population (ECA928, ECA923, ECA369, QX1791, and XZ1515) because they are genetically very different from Hawaiian strains (supplementary table S4, Supplementary Material online).

This grouping of Hawaiian and non-Hawaiian populations were used for the computation of polymorphism (*Pi*), Tajima's *D*, and Fay and Wu's *H* within each

population and were used for EHH analysis. For the calculation of $F_{ST}$ and the gene flow and migration analysis among the 11 subgroups, we removed the strains with ancestral proportion less than 0.5 and kept 221 strains for the eight non-Hawaiian subpopulations and 45 strains for the three Hawaiian subpopulations.

**Phylogenetic analysis**

To visualize the phylogenetic relationship of the Hawaiian and non-Hawaiian populations, we used the SNVs from all 45 Hawaiian strains and 24 non-Hawaiian strains (3 strains with the biggest ancestral proportion from each subgroup). These 24 strains represented the genetic diversity of the non-Hawaiian population, allowing easy visualization without making the tree too crowded. We used Tassel to convert VCF file to Phylip interleaved format and constructed the neighbour joining net with SplitsTree4 (v4.15.1) (Huson and Bryant 2006). For the trees with just csGPCR and F-box genes, we used VCFtools to extract nonsynonymous SNVs of these genes according to their genomic location.

Prior to tree construction, *Caenorhabditis briggsae* was chosen as the outgroup. The coding sequences of *C. elegans* genes and their *C. briggsae* orthologs were downloaded from Wormbase and then aligned using MegaX (Kumar et al. 2018) to identify variants. We then checked each nonsynonymous SNV that exists in the *C. elegans* wild isolates (VCF file from CeNDR) for their presence in *C. briggsae* genome. If the allele in *C. briggsae* matches the reference (N2) sequence, it is considered as a wild-type; if the *C. briggsae* allele matches the alternative sequence, it is considered to

carry that variant. If neither, we considered *C. briggsae* data missing for that SNV. In total, we found 38,695 *C. elegans* SNVs in *C. briggsae* and included these data in tree construction.

**Gene families analysis and gene enrichment analysis**

Based on previous publications, we compiled a list of genes in the chemosensory GPCR (csGPCR) gene family (Vidal et al. 2018), F-box gene family (Kipreos and Pagano 2000), transcription factor family (Reece-Hoyes et al. 2005), and protein kinase family (Manning 2005). For tissue-specific genes, we collected genes whose expression are enriched in muscle, intestine, germline (Pauli et al. 2006), and neurons (Von Stetina et al. 2007). To compare the normalized *Pi*, Tajima's *D*, and Fay and Wu's *H* values between different groups of genes, we performed non-parametric Wilcoxon's test to evaluate the statistical significance of the difference between groups.

For gene enrichment analysis, simple enrichment fold of csGPCR and F-box genes are calculated as observed gene frequency divided by expected gene frequency. We also subjected a list of specific genes to Gene Set Enrichment Analysis at wormbase.org (Angeles-Albores et al. 2016; Angeles-Albores et al. 2018). Q value threshold cutoff was set at 0.1 to generate results of Tissue Enrichment Analysis (TEA), Phenotype Enrichment Analysis (PEA), and Gene Enrichment Analysis (GEA).

**Fixation index ($F_{ST}$) calculation**

Hudson's $F_{ST}$ (Hudson et al. 1992) were estimated using PopGenome. SNVs from the 266 strains that have an ancestral proportion bigger than 0.5 (221 non-

Hawaiian and 45 Hawaiian strains) were subjected to the calculation. Prior to computation, we removed SNVs with valid genotype data in less than 100 strains to be consistent with VariScan analysis (NumNuc = 200).

**Gene flow analysis**

The migration events among subpopulations were analyzed by TreeMix (Pickrell and Pritchard 2012). We first used Stacks (Catchen et al. 2013) to convert VCF file into the input format required by treemix. In each run, 1000 SNP blocks were set for all genes, and 100 SNP blocks were set for the analysis of csGPCR or F-box genes. Hawaii_1 was used as the outgroup and three migration events were allowed. 1000 bootstrap replicates were performed for all five analyses. From the bootstrap results, we extracted the common migration events and calculated the probability of occurrence for each migration events among the 1000 replicates. The top three events were presented. We also calculated the average migration weight for each of the three events among the 1000 bootstrap replicates and the average weight were color-coded in fig. 5. To avoid possible interference by singletons and linkage disequilibrium, we repeated the analysis after removing the singletons and high linked SNPs (using plink --indep-pairwise 50 10 0.8) and got the very similar results.

**Estimation of recombination rate**

Recombination rate was estimated using an R package, FastEPRR (Gao et al. 2016). The window size was set to be 50,000 bp and the sliding step length was set as 25,000 bp. After obtaining the estimated recombination rate for each genomic window,

we assigned that recombination rate (Rho value) to the genes, whose CDS range overlap with the genomic window.

**Extended Haplotype homozygosity analysis**

We used Extended Haplotype Homozygosity (EHH) analysis to identify regions with selection footprints (Sabeti et al. 2002). VCF file was first phased by beagle (v5.1) (Browning and Browning 2007) and then subjected to haplotype analysis using the rehh (v3.0) R package (Gautier and Vitalis 2012) to calculate iHS and XPEHH value. Strains were grouped as non-Hawaiian and Hawaiian as described above when computing iHS and XPEHH and unpolarized data were used to avoid making assumption of ancestry.

**Protein domain structure and PROVEAN score**

Domain structure of F-box proteins are drawn based on the Wormbase (WS275) annotation. Prediction of transmembrane helices and extracellular region of csGPCR proteins are conducted using TMHMM Server v. 2.0 at http://www.cbs.dtu.dk/services/TMHMM/. The potential functional effect of the nonsynonymous mutations is predicted by the PROVEAN (Choi and Chan 2015) web server (PROVEAN v1.1.3) at http://provean.jcvi.org/index.php; a score lower than -2.5 is deemed as having significant effects on the function.

## Results

**Genetic polymorphism in wild isolates of *C. elegans***

Using the sequencing data released by the *Caenorhabditis elegans* natural diversity resource (CeNDR; https://www.elegansvariation.org/) (Cook et al. 2017), we analyzed the genetic polymorphisms of 330 distinct isotypes of wild strains. From the VCF files (20180527 release) available on CeNDR website, we identified in total 2,906,135 high-quality variants, including 2,493,687 single nucleotide variants (SNVs) and 412,448 small indels. Among those variants, 594,265 occurred in the protein-coding region (CDS) and 2,311,870 occurred in non-coding regions, including downstream, upstream, intron, and untranslated regions (UTRs). Among the CDS variants, 266,004 SNVs caused nonsynonymous mutations, 271,718 SNVs caused synonymous mutations, and 51,701 SNVs may affect mRNA splicing. In addition, 9,223 small indels caused frameshift mutations, and 8,245 indels caused in-frame deletion or insertion in the coding region.

Interestingly, almost half (1,124,958) of the SNVs were found in only one of the 330 isotypes, and they all occurred as homozygotes likely due to the hermaphroditism-driven homozygosity in *C. elegans*; we consider those SNVs as singleton (or private doubleton). The widespread presence of singletons is consistent with a recent genetic selective sweep that significantly reduced variations throughout the genome (Andersen et al. 2012).

When analyzing the distribution of the variants across 20,222 protein-coding genes, we found that 1,143 genes with average CDS length of 0.6 kb had no nonsynonymous mutation or small indels in CDS and 302 genes with average gene

length of 1.4 kb did not have any SNVs or small indels in any of the CDS, intron, and UTR regions. Although the absence of coding variations may be explained by the small size of these genes (the genome average for CDS length is 1.2 kb and for gene length is 3.1 kb), gene ontology and phenotype analysis found that many of these genes are involved in protein-protein interactions and their functions are essential for cell divisions and germline development (supplementary fig. S1, Supplementary Material online), suggesting that the lack of variation may also be due to purifying selection.

Next, we focused on the 2,493,687 SNVs and calculated nucleotide polymorphism (*Pi*) for each protein-coding gene (see Methods for details) and found that the gene length-normalized *Pi* is significantly bigger in introns and UTRs compared to synonymous and nonsynonymous mutations, suggesting that coding regions had much less variation than regulatory regions (supplementary fig. S2A, Supplementary Material online). We also calculated CDS length-normalized *Pi* for the nonsynonymous SNVs and found that the F-box and csGPCR genes have much larger diversity than average genes (fig. 1A-D; supplementary table S1, Supplementary Material online). For example, among the 235 genes whose normalized *Pi* is bigger than 0.01, 46 of them are F-box genes, indicating an over ten-fold enrichment. For the 1685 genes with normalized *Pi* bigger than 0.0025, F-box genes and csGPCRs showed 4.91- and 2.27-fold enrichment, respectively (fig. 1A). Overall, compared to other gene families like the transcription factor (TF) genes (891) and the protein kinase genes (402), F-box genes (336) and the csGPCRs (1301) on average have significantly bigger *Pi* (fig.

1C and D; significance by a non-parametric Wilcoxon's test). Large genetic diversity among the wild isolates hints that the F-box and csGPCR genes might contribute to the adaptation of *C. elegans* in the natural environment.

The csGPCRs can be further divided into *Str*, *Sra*, *Srg*, and *Solo* superfamilies, among which the Solo superfamily genes have the biggest *Pi* (fig. 1E). Within the *Solo* superfamily, *Srw*-type csGPCRs appeared to have the largest diversity. Interestingly, unlike the other csGPCR families, *Srw* family genes have clear phylogenetic relationship with conserved FMRFamide/peptide receptors found in insects and vertebrates and were hypothesized to represent a novel origin of chemoreceptors that detect environmental peptides in the nematode lineage (Robertson and Thomas 2006). Thus, variations in *Srw* genes might lead to selective advantages in peptide sensing.

**Population structure of *C. elegans* wild isolates**

Because the 330 wild isolates we analyzed were isolated across the globe, we next asked whether the above genetic diversity is the result of divergence among different populations of *C. elegans*. Using Admixture software, we analyzed the population structure of the 330 strains and identified 11 ancestral populations (K = 11 gave the smallest cross validation error; supplementary fig. S3 and supplementary table S2, Supplementary Material online). Six European, one North American, one Australasian, and three Hawaiian ancestral populations emerged from our analysis; they were named based on the geographic locations of most strains that carry the ancestral lineage (fig. 2A). Out of the 330 strains, 266 have one dominant ancestral lineage (one

ancestral proportion > 0.5; fig. 2A); the other 64 strains showed considerable mixing between at least three ancestral populations (supplementary table S2, Supplementary Material online). Our results generally agreed with a recent study that used 276 strains and also found 11 distinct genetic groups (Crombie et al. 2019).

Moreover, some new insights emerged from our analysis. For example, although the "Europe_1" and "Europe_2" groups are heavily mixed with populations around the world, four other European groups appear to exist in relatively purer lineage (fig. 2B). "Europe_4" group represents strains mostly found in France except for a few (e.g. DL200 from Africa). Strains sharing predominantly "Europe_5" and "Europe_6" ancestral lineages were found mostly on Iberian Peninsula and Portuguese islands with a few exception (e.g. ED3040, LKC34, ED3052, and JU3169 from Africa). Similarly, strains in North American and a few other strains (e.g. CB4858 from Australasia and JU1172 from South America) shared the same "North_America" ancestral group. These results indicate that although Crombie et al. (2019) suggests that non-Hawaiian strains are heavily mixed and not well-differentiated, there may still be certain levels of geographic isolation-based genetic divergence among these strains.

Although previous report found that Hawaiian strains maintain high level of genetic diversity and are highly divergent from the non-Hawaiian strains (Crombie et al. 2019), whether the polymorphisms carried by Hawaiian strains are more ancestral than these of non-Hawaiian strains is unclear. To address this question, we constructed the neighbor-net phylogenetic tree using all nonsynymous SNVs found in *C. elegans*

wild isolates and included a closely related species *C. briggsae* as the outgroup (fig. 3A; see Methods). This analysis clearly indicated that Hawaiian strains, epseically "Hawaii_1" and "Hawaii_2" groups, are genetically closer to the sister species and contain more ancestral variations than the non-Hawaiian strains.

Among the three Hawaiian groups in our population structure analysis, "Hawaii_1" and "Hawaii_2" groups are the same as the previous "Hawaiian Volcano" and "Hawaiian Divergent" populations defined by Crombie et al. (2019), respectively, with the exception for XZ1516 and ECA701 (supplementary table S3, Supplementary Material online). These two strains are highly divergent from other Hawaiian strains (fig. 3) and may represent a separate ancestral subpopulation themselves. In fact, many of the variations in XZ1516 and ECA701 can be found in *C. briggsae*; among the 38,695 *C. elegans* wild SNVs carried by *C. briggsae*, 16,725 (43%) are found in either of the two strains. Thus, XZ1516 and ECA701 appeared to carry the most ancestral polymorphisms in *C. elegans*.

"Hawaii_3" group is a combination of the previously defined "Hawaiian Low" and "Hawaiian Invaded" populations, both of which have admixture with non-Hawaiian populations (supplementary table S3, Supplementary Material online). Nevertheless, these two groups could emerge as separate populations if we chose 12 ancestral populatons (K = 12; supplementary fig. S3A, Supplementary Material online), suggesting further structure in the "Hawaii_3" group. In phylogenetic analysis, strains carrying "Hawaii_3" clustered more closely with non-Hawaiian strains than the strains

carrying the other two Hawaiian lineages (fig. 3B).

With regards to geographic separation, "Hawaii_1" lineage is only found in strains extracted from Hawaiian islands (fig. 2C), whereas some strains isolated from North America (QX1211, XZ2019, and ECA593) and Australasia (ECA36 and JU3226) were found to carry predominantly the "Hawaii_2" or "Hawaii_3" lineage. Three Hawaiian strains (ECA369, QX1791, and XZ1515) have extensive mixing of one or two Hawaiian acenstral groups with non-Hawaiian groups and thus do not have a predominant lineage. These obervations suggest gene flow into and/or out of Hawaiian islands. ECA923 and ECA928 isolated on O'ahu island are two anormalies, because they do not carry lineage from any of the three Hawaiian groups. Based on the population structure and genetic grouping, we divided the 330 wild isolates into Hawaiian (45 strains) and non-Hawaiian (285) populations (supplementary table S4, Supplementary Material online; see Methods).

**Large divergence of F-box and csGPCR genes among *C. elegans* subpopulations**

Although at the genome level, the divergence among the strains carrying different Hawaiian lineages and between the Hawaiian and non-Hawaiian strains are obvious, we found that these genetic divergence are more profound in F-box and csGPCR genes as indicated in expanded neighbor-joining net and increased phylogenetic distance (fig. 3B and C; supplementary fig. S4, Supplementary Material online). Within csGPCRs, *Srw* genes appear to show even greater divergence among strains carrying different ancestral populations (fig. 3D). Thus, the divergence of DNA

sequences in F-box and csGPCR genes may contribute significantly to the genetic diversity among the *C. elegans* wild isolates. Moreover, the phylogenetic trees constructed using nonsynomymous SNVs of csGPCR or F-box genes had different shapes from the tree of all genes (fig. 3D). For example, looser clustering patterns and more admixture between Hawaiian and non-Hawaiian strains were observed for the F-box and *Srw* genes, suggesting that these genes may have a distinct evolutionary history than other genes among the wild strains.

Consistent with the idea that the Hawaiian *C. elegans* population contains ancestral genetic diversity before the recent selective sweep that reduced variations in non-Hawaiian strains (Andersen et al. 2012; Cook et al. 2017; Crombie et al. 2019), we found, using nonsynomymous SNVs, that the Hawaiian strains showed over two-fold larger nucleotide polymorphism than non-Hawaiian strains across all genes (fig. 4A). "Hawaii_3" strains have lower diversity than the other two Hawaiian groups, likely because "Hawaii_3" strains are genetically more similar to the non-Hawaiian strains. Interestingly, the diversity of F-box and csGPCR genes are bigger than the TF and protein kinase genes or an average gene in both non-Hawaiian and Hawaiian populations (fig. 4A). This large genetic diversity is correlated with the abundance of segregating sites, as F-box and csGPCR genes carried many more sites than an average gene (supplementary fig. S5, Supplementary Material online). For example, the *Srw* genes of the csGPCRs and the F-box genes both have almost four times more variant sites than the average of all genes. In extreme cases, *srw-57* has only 1071 nucleotide

in the CDS but carries 124 nonsynomymous variants; *fbxb-53* is 1020-bp long in the CDS and has 207 segregating sites.

We also found that a large number (6.3 per gene on average) of segregating sites only existed in Hawaiian strains and much fewer (3.1 per gene) sites are exclusively non-Hawaiian; a significant number (4.6 per gene) of sites are shared between some Hawaiian and non-Hawaiian strains (fig. 4B). As expected, F-box and csGPCR genes have a lot more exclusively Hawaiian sites than the TF or protein kinase genes. However, they don't carry many exclusively non-Hawaiian sites, and the large diversity of the F-box and csGPCR genes in non-Hawaiian strains mostly result from the large number of sites originated from the Hawaiian population (fig. 4B). This finding supports that the Hawaiian *C. elegans* (especially the "Hawaii_1" and "Hawaii_2" groups) maintains a relatively large pool of ancestral variation, and polymorphisms in the F-box and csGPCR genes contribute significantly to this ancestral diversity. Although selective sweep removed many ancestral alleles in non-Hawaiian population, the F-box and csGPCR genes still kept a significant number of variant sites, which might be related to adaptation.

Fixation index $F_{ST}$ is a measure for genetic difference between populations. We found that $F_{ST}$ for F-box and csGPCR genes were similar to other gene families when just considering Hawaiian and non-Hawaiian as two populations (fig. 4C). ~80% of all genes have $F_{ST} < 0.2$. We reasoned that this may be caused by large divergence among the subpopulations within each population. When calculating for the eight non-

Hawaiian subpopulations (supplementary table S5, Supplementary Material online), we found that only ~60% of the genes have $F_{ST} < 0.2$ and that F-box and csGPCR genes, especially *Srw* genes, have much higher mean $F_{ST}$ than TF and protein kinase genes or an average gene (fig. 4D and 4F). This finding suggests that the polymorphism of F-box and csGPCR genes contribute significantly to the population structure of the non-Hawaiian strains. Their divergence among subpopulations and fixation within subpopulation may be linked to local adaptation. For example, csGPCR *srw-66* ($F_{ST}$ = 0.76) contains 24 variants that were found in >75% of the strains in the "North_America" group and >55% of the "Europe_2" strains but not in any other non-Hawaiian groups. Similarly, F-box gene *fbxa-181* ($F_{ST}$ = 0.69) has 14 SNVs that are found in 73% of the "Europe_6" strains and not in any other groups.

Among the three Hawaiian subpopulations, the mean $F_{ST}$ values of F-box and csGPCR Srw genes appear to be significantly lower than other gene families or the genomic average (fig. 4E and 4F), which may be explained by their large diversities even within the same Hawaiian group (fig. 4A). Thus, the big variation of F-box and *Srw* genes don't seem to follow the population structure among the three Hawaiian subpopulations and they are not likely fixed within the Hawaiian groups.

**Gene flow shaped the diversity of the F-box and csGPCR genes**

Gene flow often affect genetic polymorphism (Ropars et al. 2018). The relatedness of "Hawaii_3" and the non-Hawaiian groups prompted us to analyze the gene flow for the F-box and csGPCR genes among the subpopulations. We used

TreeMix (Pickrell and Pritchard 2012) and nonsynonymous SNVs to build maximum-likelihood population trees among the 11 subpopulations discovered above; we allowed three migration events and performed 1000 boostraps to identify migration patterns. Using SNVs of all genes, we found gene flow from "North_America" and "Europe_2" groups to "Hawaiia_3", as well as from "Europe_6" to "Europe_4" (fig. 5A). This pattern is consistent with the admixing of "Hawaii_3" and non-Hawaiian lineages (fig. 2A). Surprisingly, the gene flow patterns of F-box and csGPCR genes appeared to be different from the overall pattern for the genome. We found that F-box genes have extensive gene flow between Hawaiian and non-Hawaiian populations in both directions, whereas csGPCR genes had only gene flow within the non-Hawaiian subpopulations (fig. 5B). Thus, variants in F-box genes may be exchanged frequently between Hawaiian and non-Hawaiian populations, which is supported by the great number of shared segregating sites in F-box genes between the two populations (fig. 4B). This gene flow may also contribute to the lack of fixation of F-box gene alleles within the Hawaiian subpopulations. On the other hand, csGPCR genes do not seem to contribute to this flow between Hawaiian and non-Hawaiian populations.

Previous work suggested that gene flow both into and out of Hawaii contributed to the global genetic diversity pattern of *C. elegans* (Crombie et al. 2019), our work appears to support gene flow into Hawaiian islands and identified the "North_America" and the "Europe_2" groups as the likely source of invasion. More importantly, our results indicate that the variants in F-box genes may be an important

part of the genetic information that flew into the Hawaiian populations and this gene flow in turn increased the global diversity of F-box genes.

Moreover, we also analyzed gene flow in the absence of the polymorphisms in F-box and csGPCR genes. Interestingly, the structure of the population tree changed after we removed the SNVs in F-box or csGPCR genes; instead of staying as a branch outside of the eight non-Hawaiian subpopulations, the "Hawaii_3" group moved into the non-Hawaiian groups and was placed next to "Europe_2" and "North_America" (fig. 5C). This finding supports that variations in the F-box and csGPCR genes played critical roles in distinguishing "Hawaii_3" strains from the non-Hawaiian populations and contributed significantly to intraspecific diversity.

**High recombination rate may contribute to the polymorphism of the F-box and csGPCR genes**

We next asked whether chromosomal locations of the csGPCR and F-box gene loci had effects on their diversity. Most of the csGPCR are located on chromosome II (13%), IV (9%), and V (70%), and most of F-box genes are located on the arms of chromosome II (33%), III (22%), and V (26%) (fig. 6A). We found that the chromosomal arms (the two distal quarters) have larger polymorphism than the center (the middle half) for all chromosomes, which is likely due to high frequency of recombination (Begun and Aquadro 1992; McGaugh et al. 2012) (fig. 6B). The fact that almost all F-box genes are located in the chromosomal arms may contribute to their high genetic diversity. csGPCR genes are also relatively more concentrated on the arms,

484 especially on chromosome II and V, although to a lesser extent than F-box genes. For
485 example, over 80% of the *Srw* genes are located on the arms of V. This chromosomal
486 clustering is likely the result of rapid gene duplication (Robertson and Thomas 2006).
487 In contrast, protein kinase and TF genes are more evenly spread out across
488 chromosomes (fig. 6A).

489 We estimated the recombination rate for Chromosome II, III, and V using the
490 FastEPRR software (Gao et al. 2016) and all nonsynonymous SNVs. As expected,
491 chromosomal arm regions have higher estimated recombination rate ( or *Rho*) than
492 the center (fig. 6C). Interestingly, the left arm of chromosome III where many F-box
493 genes are located have significantly higher recombination rate than the right arm, and
494 this high *Rho* value of F-box genes appear to correlate with the large polymorphism
495 (fig. 6D). Similarly, the arms of Chromosome V, which harbour many csGPCR genes,
496 have high recombination rate, which are correlated with large *Pi* (fig. 6D). Therefore,
497 because of the clustering of F-box and csGPCR genes on the chromosomal arms,
498 recombination likely contributed to their large diversity.

499 **Signs of strong positive selection on F-box and csGPCR genes**

500 Although previous studies hypothesized that positive selection of alleles that
501 confer fitness advantages under human influence reduced genetic variations in non-
502 Hawaiian population (Andersen et al. 2012), what genes are under selection is not clear.
503 Using the nonsynonymous SNVs, we performed neutrality test and calculated Tajima's
504 *D* and Fay and Wu's *H* values for every gene. The *D* value reflects the difference

between expected and observed diversity (Tajima 1989) and the *H* value measures the abundance of high-frequency derived allele (Fay and Wu 2000). Negative *D* and *H* values are both indicators of selective sweep and positive selection. To calculate the *H* value, we used XZ1516 or ECA701 as the outgroup, because these two strains likely carry the most ancestral genotypes according to the phylogenetic tree (fig. 3). *H* values calculated using the two strains as the outgroup were similar (supplementary fig. S6, Supplementary Material online), and in the following analysis we used XZ1516 as the outgroup.

In the neutrality tests, we found that Tajima' *D* were negative for most (>85%) genes in both CDS (synonymous and nonsynonymous variants) and the regulatory regions (intron and UTR variants); Fay and Wu's *H* were negative for the nonsynonymous variants of ~50% genes (fig. 7A; supplementary table S1, Supplementary Material online). This finding is consistent with the chromosome-wide sweep that occurred across the genome recently (Andersen et al. 2012). When assessing statistical significance of the *D* and *H* values using coalescence simulation (R.R. 1990; Librado and Rozas 2009), we found that vast majority (>95%) of the *D* value smaller than -2 and *H* value smaller than -20 have a *p* value lower than 0.05. Interestingly, F-box and csGPCR genes are overrepresented among the genes with significantly negative *D* and *H* values (fig. 7B). For example, among the 1038 genes with $H < -20$, 260 of them are csGPCRs (3.62-fold enriched) and 67 are F-box genes (3.61-fold enriched). Gene ontology analysis consistently showed strong enrichment (> 5 fold) in

genes involved in sensory perception of smell and chemical stimulus (supplementary fig. S7, Supplementary Material online).

Compared with the TF and protein kinases genes or the genomic average, F-box and csGPCR genes have significantly lower *D* and *H* values (fig. 7C and D), suggesting that the csGPCR and F-box genes appear to be under stronger positive selection than other genes. Within the csGPCRs, Solo superfamily genes have the lowest *H* values and within the *Solo* superfamily, *Srw*-type csGPCRs have the lowest *H*, indicating that *Srw* genes may be under the strongest positive selection among all csGPCRs (fig. 7E and F). Putative functions of the *Srw* genes in sensing environmental peptides suggest they may be involved in adaptation.

Within the F-box genes, we did not observe significant difference in either *D* or *H* values or polymorphisms among the genes in *fbxa*, *fbxb*, and *fbxc* subfamilies (supplementary fig. S8A, Supplementary Material online). F-box proteins share an F-box domain, which complexes with Skp and Cullin to form the SCF complex that mediates protein ubiquitination and degradation. Five out of the 20 Skp-related genes in *C. elegans* (*skr-3*, *4*, *5*, *10*, and *15*) and three out of the 6 Cullin genes (*cul-1*, *3*, and *6*) have very negative *H*, suggesting strong selective sweep (supplementary fig. S8B, Supplementary Material online). Thus, components of the ubiquitination-proteasome system (UPS) may co-evolve among the *C. elegans* wild isolates. Because F-box proteins recognize numerous substrates and target them for degradation (Skaar et al. 2013), variations in F-box genes may alter the homeostasis of particular target proteins,

leading to certain advantages during selection.

In addition, we found that the 253 nuclear hormone receptor (NHR) genes had much bigger polymorphisms and more negative *D* and *H* values than other groups of TF genes (supplementary fig. S8 C and D, Supplementary Material online). In contrast, 52 TF genes (mostly homeodomain TFs) known to regulate neuronal differentiation (Hobert 2016) showed extremely small diversity and have relatively large or even positive *H*. This result suggests that genes essential for development may not be the target of positive selection in the recent sweep, whereas variations in NHR genes, which have undergone expansion and divergence in nematodes and mostly regulate molting, diapause, metabolism, and stress response (Antebi 2006; Jones et al. 2013), may play a bigger role than other TFs in the adaptation to a changing environment.

We also analyzed genes expressed in specific tissues, including muscle, intestine, germline, and neurons (Pauli et al. 2006; Von Stetina et al. 2007), and found no signs of stronger selection than an average gene (supplementary fig. S9; supplementary table S6, Supplementary Material online). Among the tissues, genes enriched in neurons had smaller *Pi* and bigger *D* and *H* values than genes in the other three tissues. Genes enriched in all muscle, intestine, and germline have smaller *Pi* and bigger *D* and *H* values than genes enriched in only one of the three tissues (supplementary fig. S9, Supplementary Material online). Thus, pan-neuronal genes and genes expressed in multiple tissues are likely not the target of positive selection perhaps due to their pleotropic effects on the development and behavior of the animal.

**Positive selection within *C. elegans* subpopulations**

The above analysis detected signs of strong positive selection in csGPCRs and F-box genes among the wild strains, we next calculated the Fay and Wu's *H* within subpopulations of *C. elegans* and found that the *H* value is more negative in the non-Hawaiian strains than the Hawaiian strains across all genes (fig. 8A). "Hawaii_3" group appeared to have lower *H* values than "Hawaii_1" and "Hawaii_2" groups probably due to the admixing with the non-Hawaiian strains. These observations are consistent with the selective sweep in non-Hawaiian populations. Genes in the csGPCR (especially *Srw* genes) and F-box gene families continued to show much more negative *H* than the genomic average even in Hawaiian strains, suggesting that these genes may also be under positive selection within the Hawaiian populations when considering XZ1516 as the most ancestral strain.

Negative *H* values reflects the excess of high-frequency derived alleles. Indeed, the number of high-frequency (>50%) SNVs distinct from the ancestral allele in XZ1516 is much higher in csGPCR and F-box genes than in an average gene for both Hawaiian and non-Hawaiian populations (fig. 8B). The large number of both segregating sites and high-frequency derived alleles in csGPCR and F-box genes suggest that they evolve more rapidly than other parts of the genome and may contribute to the adaptation to a changing environment in *C. elegans*.

Since the Hawaiian populations contained relatively more ancestral polymorphism than the non-Hawaiian population, we next asked whether the csGPCR

and F-box genes accumulated high-frequency derived alleles in non-Hawaiian population when considering Hawaiian strains as the ancestral population. We tested this idea by calculating Fay and Wu's *H* of non-Hawaiian strains using ECA396 (a representative "Hawaii_1" strain) or ECA742 (a representative "Hawaii_2" strain) as the outgroup and found that these *H* values were generally less negative than the *H* values calculated using XZ1516 or ECA701 as the outgroup (fig. 7C). This result confirmed that "Hawaii_1" and "Hawaii_2" strains are less distant to the non-Hawaiian strains than the two outliers XZ1516 and ECA701. F-box and csGPCR genes still showed significantly negative and smaller than average *H* values with ECA396 or ECA742 as the outgroup (fig. 8C and D), indicating signs of strong positive selection within the non-Hawaiian population, relative to the Hawaiian populations. Thus, the neutrality test statistics appeared to suggest that F-box and csGPCR genes were under strong selective pressure and experienced significant change in the recent sweep that affected *C. elegans* outside of Hawaii.

At the individual gene level, to understand how the nonsynonymous changes may affect functions of the F-box and csGPCR genes, we analyzed where these changes occur within the protein and predict whether the amino acid substitutations have significant functional effects by calculating the PROVEAN (Protein Variation Effect Analyzer) score (Choi and Chan 2015). For example, among the genes with very negative *H* values, F-box gene *fbxb-49* contains 48 high frequency derived sites in the non-Hawaiian population (considering XZ1516 as the outgroup) and the frequency of

those sites within the Hawaiian population are much lower (fig. 9A). Most of those sites are also high frequency derived sites when using a “Hawaii_1” or “Hawaii_2” strain as the outgroup. These results suggest that selective sweep may have fixed those sites in the non-Hawaiian population. Interestingly, very few variants are located in the region encoding F-box domain, which is responsible for the assembly of E3 ubiquitin ligase SCF complex. In contrast, most of the amino acid substitutions likely occurred in domains that are responsible for recognizing the substrate protein (fig. 9A). Four of the sites have PROVEAN scores below -2.5, suggesting significant functional impacts. These observations are consistent with the hypothesis that variations in the F-box genes may alter the turnover of some cellular proteins, leading to selective advantages.

Similarly, as an example of csGPCR genes, *srw-68* contains 44 high frequency derived sites in the non-Hawaiian population with XZ1516 as the outgroup, and 16 of them are also high frequency dervied sites compared to ECA396 and ECA742. All of those sites are near fixation among the non-Hawaiian strains. When mapping the variants onto the predicted protein structure organizaiton, we found that most of the amino acid substitutaitons occurred in the extracellular domain and very few in the transmembrane domain. This finding supports the idea that variations of csGPCR genes may result in changes in the ability to sense envrionmental signals, which may confer fitness advantages under human influence.

**Extended Haplotype Homozygosity analysis identified selection footprints in csGPCR and F-box genes**

In addition to the neutrality tests, we also applied the Extended Haplotype Homozygosity (EHH) method (Sabeti et al. 2002) to detect the selection footprints among the nonsynonymous SNVs across the genome. EHH identifies long-range haplotypes and can discover genomic regions with strong selective sweep. First, we computed the integrated Haplotype Score (iHS) for both non-Hawaiian and Hawaiian strains (supplementary table S7, Supplementary Material online). Interestingly, the regions that showed extended haplotype homozygosity (high |iHS| scores) were in the left arms of chromosome II and III, where F-box genes are located, and the two arms of chromosome V, where most csGPCRs are located (fig. 10A and B). Indeed, among the 335 genes carrying at least one SNV with |iHS| > 2 in non-Hawaiian strains, csGPCR and F-box genes are enriched for 4.5 and 1.5 fold, respectively (fig. 10D), indicating that these genes may be under selective pressure in the recent sweep among non-Hawaiian strains. Nevertheless, csGPCR and F-box genes may still be selected within the Hawaiian strains because of their enrichment in the regions with high |iHS| in the Hawaiian population (fig. 10D).

This genomic pattern of haplotype homozygosity is supported by that Fay and Wu's *H* values of genes on the left arms of II and III are much more negative than the center and right arms and that *H* values of genes on both arms of V are smaller than the center of V (fig. 10E). F-box and csGPCR may be driving this pattern, because they tend to have even more negative *H* than average genes in the arms. In addition, Chromosome V generally had much more negative *H* than other chromosomes,

suggesting signs of strong selective sweep, which is consistent with a previous observation of high haplotype homozygosity of V among non-Hawaiian strains (Andersen et al. 2012). Selection of the over 1000 csGPCR genes on V may explain this chromosomal pattern.

Genes that are under selection in the non-Hawaiian population but not in the Hawaiian population may be associated with the adaptation to human interference. So, we conducted the XP-EHH (Cross-Population EHH) test to identify SNVs with such selection pattern and found the left arm of chromosome II and both arms of V contain regions with significantly positive XP-EHH values (fig. 10C). F-box and csGPCR genes are highly enriched in those regions. 18 out of the 41 genes carrying SNVs with XP-EHH > 2 on the left arm of II are F-box genes. The enrichment of F-box and csGPCR genes is even more obvious if we only consider the outlier SNVs (the top 0.05%) or count all genes in extended regions that connect significant SNVs within a 50-kb range (Mohd-Assaad et al. 2018) (fig. 10D). In summary, both neutrality test and EHH analysis identified signs of strong positive selection on F-box and csGPCR genes in non-Hawaiian population.

As an example of highly selected genes, csGPCR *srw-56* contains 67 SNVs with significantly positive XP-EHH score (XP-EHH > 2; $p < 0.05$). 36 and 24 of those SNVs are high frequency derived sites in non-Hawaiian population with ECA396 ("Hawaii_1") and ECA742 ("Hawaii_2") as the outgroup, respectively. Thus, *srw-56* carries at least 24 variants that are prevalent in the non-Hawaiian strains but not the

Hawaiian population. All of the 24 amino acid substitution occurred in the extracellular domains of SRW-56, suggesting possible alteration in ligand binding (fig. 10F). Similarly, F-box gene *fbxa-85* carries 58 SNVs with high XP-EHH; 13 and 29 are high frequency derived sites among non-Hawaiian strains with ECA396 and ECA742 as the outgroup. Most of the missense mutation occurred in the FTH domain involved in recognizing target proteins and none in the F-box domain involved in forming the SCF complex (fig. 10G). Those mutations may affect the turnover rate of particular proteins in the non-Hawaiian population.

## Discussion

The nematode *C. elegans*, which is traditionally used as a model organism for molecular biology, has emerged as an important organism in studying the genetic mechanisms of evolution. The genomic sequences of the >50 species in the Caenorhabditis genus and the ~330 wild *C. elegans* isotypes provided an important resource for understanding the evolutionary history of *C. elegans* and nematodes in general, e.g. the rise of self-fertile hermaphroditism through convergent evolution in *C. elegans* and *C. briggsae* (Nayak et al. 2005) and the balancing selection maintaining genetic incompatibilities among *C. elegans* wild isolates (Seidel et al. 2008). In this study, we aimed at identifying genes or gene families that have large diversity among the *C. elegans* wild isolates and show signs of positive selection during the recent selective sweep. The F-box gene family and chemosensory GPCR genes emerged from

our analysis, suggesting that they may contribute to the adaptation of wild *C. elegans*.

**Intraspecific positive selection of F-box genes**

Compared to insects and vertebrates, *C. elegans* genome contains a large number of F-box genes. This increased number of F-box genes might have allowed selective recognition of target proteins for degradation in a precisely controlled manner and the increased precision in the regulation of protein turnover might have contributed to nematode evolution. In fact, an earlier study calculated the nonsynonymous (*dN*)/synonymous (*dS*) ratio among paralogous F-box genes in *C. elegans* reference genome (the N2 strain) and found evidence of purifying selection in the sequence encoding the F-box domain and positive selection in the substrate recognition domain (Thomas 2006). Our studies using the genomic sequences of 330 *C. elegans* wild isolates found extremely large intraspecific variations in the F-box genes, which is consistent with their rapid evolution, and strong positive selection of these genes in non-Hawaiian population, which supports their roles in adaptation. Interestingly, high frequency derived sites mostly occurred outside of the region that codes for the F-box domain, supporting that the function of substrate recognition but not Skp1 binding is the target of positive selection.

What kind of selective advantages can variants in F-box genes confer? Recent studies suggested a link between the SCF complex and antimicrobial immunity in *C. elegans*, because the transcription of many components of the SCF complex were upregulated upon Orsay virus and *Nematocida parisii* (a microsporidia fungi) infections

(Chen et al. 2017) and RNAi knockdown of the core SCF components promoted the infection (Bakowski et al. 2014). Among the upregulated genes are F-box genes that show strong signs of positive selection in our studies, e.g. *fbxc-19*, *fbxa-75*, *fbxa-135*, *fbxa-158*, *fbxa-165*, and *fbxa-182*, whose Fay and Wu's *H* are all below -20. Thus, an attractive hypothesis is that variations in F-box proteins allow or enhance the ability of SCF complex to ubiquitinate microbial and/or host proteins required for the replication of the pathogen, thus contributing to stronger immune defence. In addition to antiviral immunity, we also suspect certain alleles of F-box genes to confer other fitness advantages, given the importance of ubiquitination-proteasome system in many biological processes.

**Intraspecific adaptive evolution of csGPCRs**

The csGPCR family is the largest gene family in *C. elegans* and contains over 1,300 genes. Through the studies of specific phenotypes, a few csGPCRs were previously connected to adaptation. For example, the deletion of two csGPCR genes, *srg-36* and *srg-37*, which resulted in insensitivity to the dauer pheromone ascaroside and defects in entering dauer diapause, were acquired independently by two domesticated *C. elegans* strains grown in high density (McGrath et al. 2011). Similar loss-of-function deletions in the two csGPCRs were also found in natural isolates across the globe, suggesting that niche-associated variation in pheromone receptors may contribute to the boom-and-bust population dynamics (Lee et al. 2019). In addition, a frameshift-causing deletion in another csGPCR, *str-217*, in Hawaiian strain CB4856,

led to resistance to the insect repellents N,N-Diethyl-meta-toluamide (DEET) (Dennis et al. 2018) and similar deletions were found in nine other wild isotypes (our unpublished results), suggesting that *C. elegans* may have evolved to acquire resistance to harmful environmental chemicals by inactivating csGPCRs. The above examples showcased how the intraspecific evolution of individual csGPCR genes can have significant functional consequences and contribute to adaptation. Our study in a more systematic way indicate that csGPCRs are highly diverse and are under strong positive selection in the *C. elegans* wild population.

Among the four csGPCR superfamilies (*Str*, *Sra*, *Srg*, and *Solo*), our analysis using nonsynonymous SNVs found that *Str* genes had larger polymorphism and stronger positive selection than *Sra* and *Srg* genes, which is consistent with previous observation on the intraspecific variations of *Str* genes (Stewart et al. 2005). In fact, these variations created ~200 pseudogenes *Str* genes in *C. elegans* reference genome (the N2 strain) through often times only one apparent defect. Compared to *Str* genes, we found that *Solo* superfamily csGPCRs, especially *srw* genes have even larger diversity and stronger positive selection. Interestingly, *srw* genes are the only csGPCRs that have clear homology with insect and vertebrate GPCRs and likely code for FMRFamide/peptide receptors (Robertson and Thomas 2006). Large variations in *srw* genes might reflect the need to detect a wide range of environmental peptides. Moreover, most of the high frequency derived sites of *srw* genes in non-Hawaiian population are mapped to the regions that code for the extracellular domains, suggesting that altered

ligand recognition might be positively selected. A similar observation was made for *srz* genes in the *Solo* superfamily based on that *dN/dS* ratios among paralogous groups of *srz* genes in *C. elegans* and *C. briggsae* peak in the extracellular loops (Thomas et al. 2005). Thus, the large gene pool of csGPCRs may facilitate the adaptation to a changing environment by supplying alleles with specific ligand binding properties for positive selection.

**The correlation between large diversity and strong positive selection**

Compared to other gene families, F-box and csGPCR genes not only have large genetic diversity but also show strong signs of positive selection. This correlation is counterintuitive, because selection tends to reduce variation. We reason that gene families such as the TFs and protein kinases have low polymorphism because they play critical roles in the development of *C. elegans* and thus may be under purifying selection. In comparison, F-box and csGPCR genes maintain large polymorphism likely due to the lack of strong purifying selection, as well as high recombination rate and frequent gene flow. High recombination rate results from their clustering in the chromosomal arms, and gene flow between genetically divergent subpopulations helps maintain genetic diversity. Thus, even after the fixation of some derived variants that confer selective advantages in the non-Hawaiian population, F-box and csGPCR genes may still maintain a larger than average polymorphism because of their high recombination rate and frequent gene flow.

## Acknowledgments

We thank the *Caenorhabditis elegans* Natural Diversity Resource (CeNDR) at Northwestern University for sharing genomic data of the *C. elegans* wild isolates through their website. We thank Chun Yin Lau for helping with some bioinformatic analysis. This study is supported by start-up research grant and the seed fund for basic research (201910159087) from the University of Hong Kong to C.Z.

## References

Andersen EC, et al. 2012. Chromosome-scale selective sweeps shape Caenorhabditis elegans genomic diversity. Nature Genetics 44(3):285-290.

Angeles-Albores D, Lee RYN, Chan J, Sternberg PW. 2018. Two new functions in the WormBase Enrichment Suite. Micropublication: biology. Dataset.

Angeles-Albores D, RY NL, Chan J, Sternberg PW. 2016. Tissue enrichment analysis for C. elegans genomics. Bmc Bioinformatics 17(1):366.

Antebi A. 2006. Nuclear hormone receptors in C. elegans. WormBook1-13.

Bakowski MA, et al. 2014. Ubiquitin-mediated response to microsporidia and virus infection in C. elegans. PLoS Pathog 10(6):e1004200.

Begun DJ, Aquadro CF. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature 356(6369):519-520.

Bounoutas A, Zheng Q, Nonet ML, Chalfie M. 2009. mec-15 encodes an F-box protein

required for touch receptor neuron mechanosensation, synapse formation and development. Genetics 183(2):607-617, 601SI-604SI.

Bradbury PJ, et al. 2007. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23(19):2633-2635.

Browning SR, Browning BL. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81(5):1084-1097.

Catchen J, et al. 2013. Stacks: an analysis tool set for population genomics. Molecular Ecology 22(11):3124-3140.

Chen CJ, Chen H, He YH, Xia R. 2018. TBtools, a Toolkit for Biologists integrating various biological data handling tools with a user-friendly interface. BioRxiv https://doi.org/10.1101/289660.

Chen K, et al. 2017. An evolutionarily conserved transcriptional response to viral infection in Caenorhabditis nematodes. BMC Genomics 18(1):303.

Choi Y, Chan AP. 2015. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics 31(16):2745-2747.

Cook DE, Zdraljevic S, Roberts JP, Andersen EC. 2017. CeNDR, the Caenorhabditis elegans natural diversity resource. Nucleic Acids Research 45(D1):D650-D657.

Crombie TA, et al. 2019. Deep sampling of Hawaiian Caenorhabditis elegans reveals high genetic diversity and admixture with global populations. Elife 8.

Danecek P, et al. 2011. The variant call format and VCFtools. Bioinformatics

27(15):2156-2158.

Dennis EJ, et al. 2018. A natural variant and engineered mutation in a GPCR promote DEET resistance in C. elegans. Nature 562(7725):119-123.

Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. Genetics 155(3):1405-1413.

Fielenbach N, et al. 2007. DRE-1: an evolutionarily conserved F box protein that regulates C. elegans developmental age. Dev Cell 12(3):443-455.

Francis RM. 2017. pophelper: an R package and web app to analyse and visualize population structure. Mol Ecol Resour 17(1):27-32.

Gao F, Ming C, Hu W, Li H. 2016. New Software for the Fast Estimation of Population Recombination Rates (FastEPRR) in the Genomic Era. G3 (Bethesda) 6(6):1563-1571.

Gautier M, Vitalis R. 2012. rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. Bioinformatics 28(8):1176-1177.

Ghazi A, Henis-Korenblit S, Kenyon C. 2007. Regulation of Caenorhabditis elegans lifespan by a proteasomal E3 ligase complex. Proc Natl Acad Sci U S A 104(14):5947-5952.

Guo Y, Lang S, Ellis RE. 2009. Independent recruitment of F box genes to regulate hermaphrodite development during nematode evolution. Curr Biol 19(21):1853-1860.

Hobert O. 2016. A map of terminal regulators of neuronal identity in Caenorhabditis elegans. Wiley Interdiscip Rev Dev Biol 5(4):474-498.

Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from

DNA sequence data. Genetics 132(2):583-589.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Molecular Biology and Evolution 23(2):254-267.

Hutter S, Vilella AJ, Rozas J. 2006. Genome-wide DNA polymorphism analyses using VariScan. Bmc Bioinformatics 7409.

Jager S, Schwartz HT, Horvitz HR, Conradt B. 2004. The Caenorhabditis elegans F-box protein SEL-10 promotes female development and may target FEM-1 and FEM-3 for degradation by the proteasome. Proc Natl Acad Sci U S A 101(34):12549-12554.

Jones LM, Rayson SJ, Flemming AJ, Urwin PE. 2013. Adaptive and specialised transcriptional responses to xenobiotic stress in Caenorhabditis elegans are regulated by nuclear hormone receptors. PLoS One 8(7):e69956.

Kim K, et al. 2009. Two chemoreceptors mediate developmental effects of dauer pheromone in C. elegans. Science 326(5955):994-998.

Kipreos ET, Pagano M. 2000. The F-box protein family. Genome Biol 1(5):REVIEWS3002.

Kumar S, et al. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Molecular Biology and Evolution 35(6):1547-1549.

Lee D, et al. 2019. Selection and gene flow shape niche-associated variation in pheromone response. Nat Ecol Evol 3(10):1455-1463.

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25(11):1451-1452.

Manning G. 2005. Genomic overview of protein kinases. WormBook1-19.

McGaugh SE, et al. 2012. Recombination modulates how selection affects linked sites in Drosophila. PLoS Biol 10(11):e1001422.

McGrath PT, et al. 2011. Parallel evolution of domesticated Caenorhabditis species targets pheromone receptor genes. Nature 477(7364):321-325.

Mohd-Assaad N, McDonald BA, Croll D. 2018. Genome-Wide Detection of Genes Under Positive Selection in Worldwide Populations of the Barley Scald Pathogen. Genome Biol Evol 10(5):1315-1332.

Nayak S, Goree J, Schedl T. 2005. fog-2 and the evolution of self-fertile hermaphroditism in Caenorhabditis. PLoS Biol 3(1):e6.

NEI M. 1987. Molecular Evolutionary Genetics. Columbia Univ. Press, New York.

Park D, et al. 2012. Interaction of structure-specific and promiscuous G-protein-coupled receptors mediates small-molecule signaling in Caenorhabditis elegans. Proc Natl Acad Sci U S A 109(25):9917-9922.

Pauli F, et al. 2006. Chromosomal clustering and GATA transcriptional regulation of intestine-expressed genes in C. elegans. Development 133(2):287-295.

Pfeifer B, Wittelsburger U, Ramos-Onsins SE, Lercher MJ. 2014. PopGenome: an efficient Swiss army knife for population genomic analyses in R. Molecular Biology and Evolution 31(7):1929-1936.

Pickrell JK, Pritchard JK. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet 8(11):e1002967.

Purcell S, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3):559-575.

R.R. H. 1990. Gene genealogies and the coalescent process. Oxford surveys in evolutionary biology.

Reece-Hoyes JS, et al. 2005. A compendium of Caenorhabditis elegans regulatory transcription factors: a resource for mapping transcription regulatory networks. Genome Biol 6(13):R110.

Robertson HM. 2000. The large srh family of chemoreceptor genes in Caenorhabditis nematodes reveals processes of genome evolution involving large duplications and deletions and intron gains and losses. Genome Research 10(2):192-203.

Robertson HM, Thomas JH. 2006. The putative chemoreceptor families of C. elegans. WormBook1-12.

Ropars J, et al. 2018. Gene flow contributes to diversification of the major fungal pathogen Candida albicans. Nat Commun 9(1):2253.

Rozas J, et al. 2017. DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. Molecular Biology and Evolution 34(12):3299-3302.

Sabeti PC, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419(6909):832-837.

Seidel HS, Rockman MV, Kruglyak L. 2008. Widespread genetic incompatibility in C. elegans maintained by balancing selection. Science 319(5863):589-594.

Sengupta P, Chou JH, Bargmann CI. 1996. odr-10 encodes a seven transmembrane

domain olfactory receptor required for responses to the odorant diacetyl. Cell 84(6):899-909.

Skaar JR, Pagan JK, Pagano M. 2013. Mechanisms and function of substrate recruitment by F-box proteins. Nat Rev Mol Cell Biol 14(6):369-381.

Stevens L, et al. 2019. Comparative genomics of 10 new Caenorhabditis species. Evol Lett 3(2):217-236.

Stewart MK, et al. 2005. High genetic diversity in the chemoreceptor superfamily of Caenorhabditis elegans. Genetics 169(4):1985-1996.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123(3):585-595.

Thomas JH. 2006. Adaptive evolution in two large families of ubiquitin-ligase adapters in nematodes and plants. Genome Res 16(8):1017-1030.

Thomas JH, et al. 2005. Adaptive evolution in the SRZ chemoreceptor families of Caenorhabditis elegans and Caenorhabditis briggsae. Proc Natl Acad Sci U S A 102(12):4476-4481.

Thomas JH, Robertson HM. 2008. The Caenorhabditis chemoreceptor gene families. BMC Biol 642.

Vidal B, et al. 2018. An atlas of Caenorhabditis elegans chemoreceptor expression. PLoS Biol 16(1):e2004218.

Von Stetina SE, et al. 2007. Cell-specific microarray profiling experiments reveal a comprehensive picture of gene expression in the C. elegans nervous system. Genome

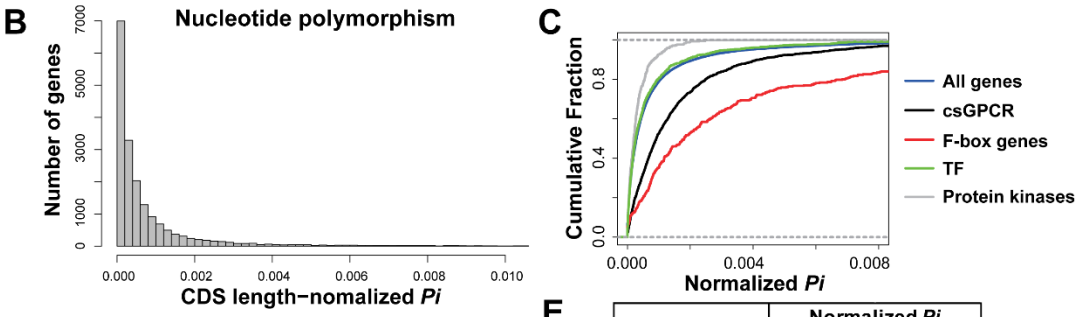
Biol 8(7):R135.

Xu G, Ma H, Nei M, Kong H. 2009. Evolution of F-box genes in plants: different modes of sequence divergence and their relationships with functional diversification. Proc Natl Acad Sci U S A 106(3):835-840.

**FIG. 1-10**

**A**

| | | Normalized *Pi* > 0.01 | | Normalized *Pi* > 0.005 | | Normalized *Pi* > 0.0025 | |
|---|---|---|---|---|---|---|---|
| | Total number | Number | Enrichment fold | Number | Enrichment fold | Number | Enrichment fold |
| csGPCRs | 1301 | 20 | 1.23 | 102 | 1.99 | 265 | 2.27 |
| F-box genes | 336 | 46 | 10.95 | 86 | 6.49 | 148 | 4.91 |
| All genes | 18797 | 235 | | 741 | | 1685 | |




**D**

| | Normalized *Pi* | |
|---|---|---|
| | Mean | Median |
| All genes | 0.00099 | 0.00033 |
| csGPCRs | 0.00179 | 0.00095 |
| F-box genes | 0.00424 | 0.00200 |
| TF | 0.00078 | 0.00027 |
| Protein kinase | 0.00036 | 0.00021 |

**E**

| csGPCR group | Normalized *Pi* | |
|---|---|---|
| | Mean | Median |
| Solo (293) | 0.00260 | 0.00144 |
| Str (567) | 0.00191 | 0.00116 |
| Sra (126) | 0.00087 | 0.00062 |
| Srg (315) | 0.00117 | 0.00056 |
| Srw (112) | 0.00338 | 0.00189 |
| Srz (65) | 0.00249 | 0.00181 |
| Srbc (69) | 0.00283 | 0.00154 |
| Srsx (37) | 0.00045 | 0.00039 |

**FIG. 1. Large genetic polymorphism of csGPCR and F-box genes.**

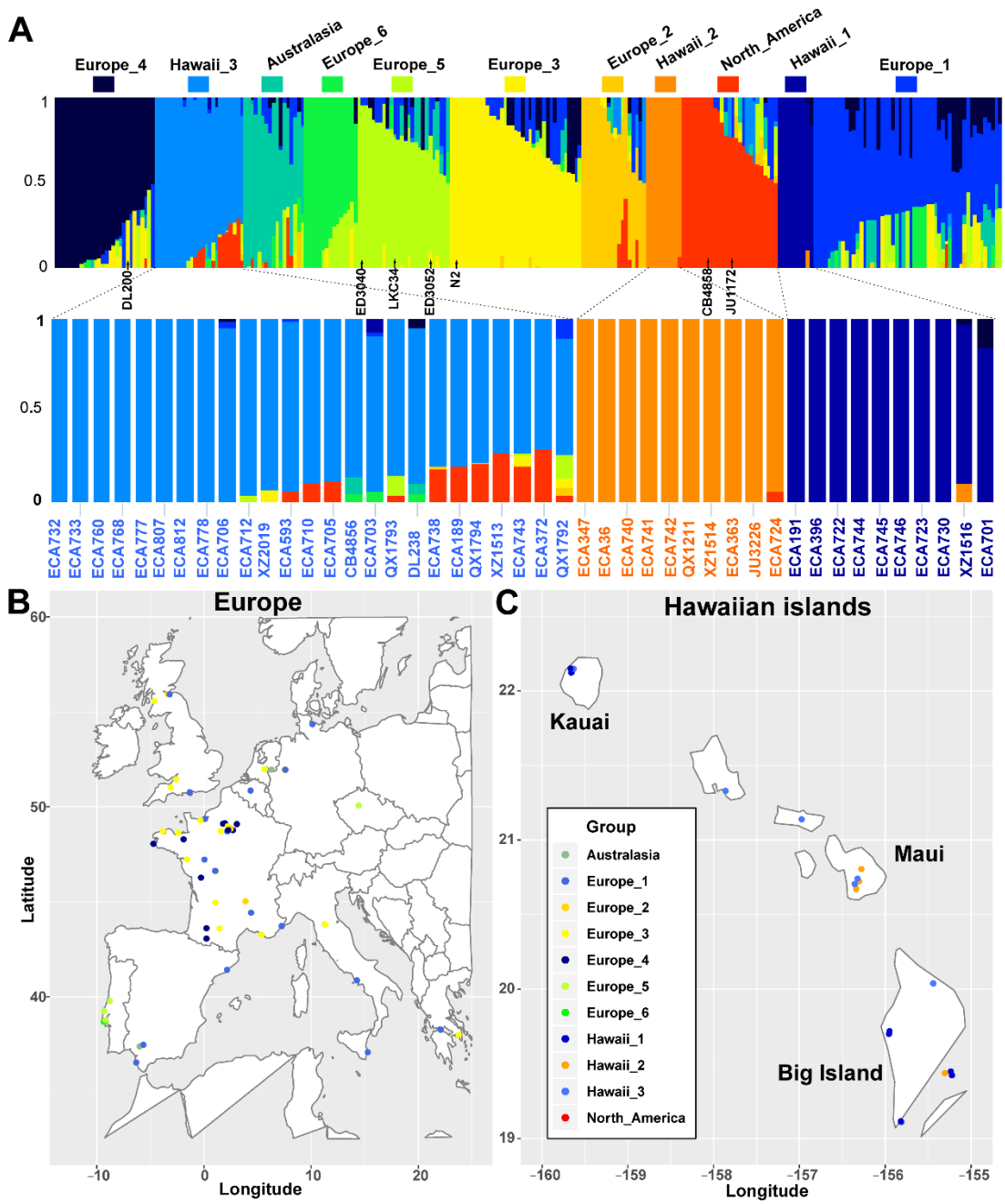



**FIG. 2. Population structure and geographical distribution of *C. elegans* wild isolates.**

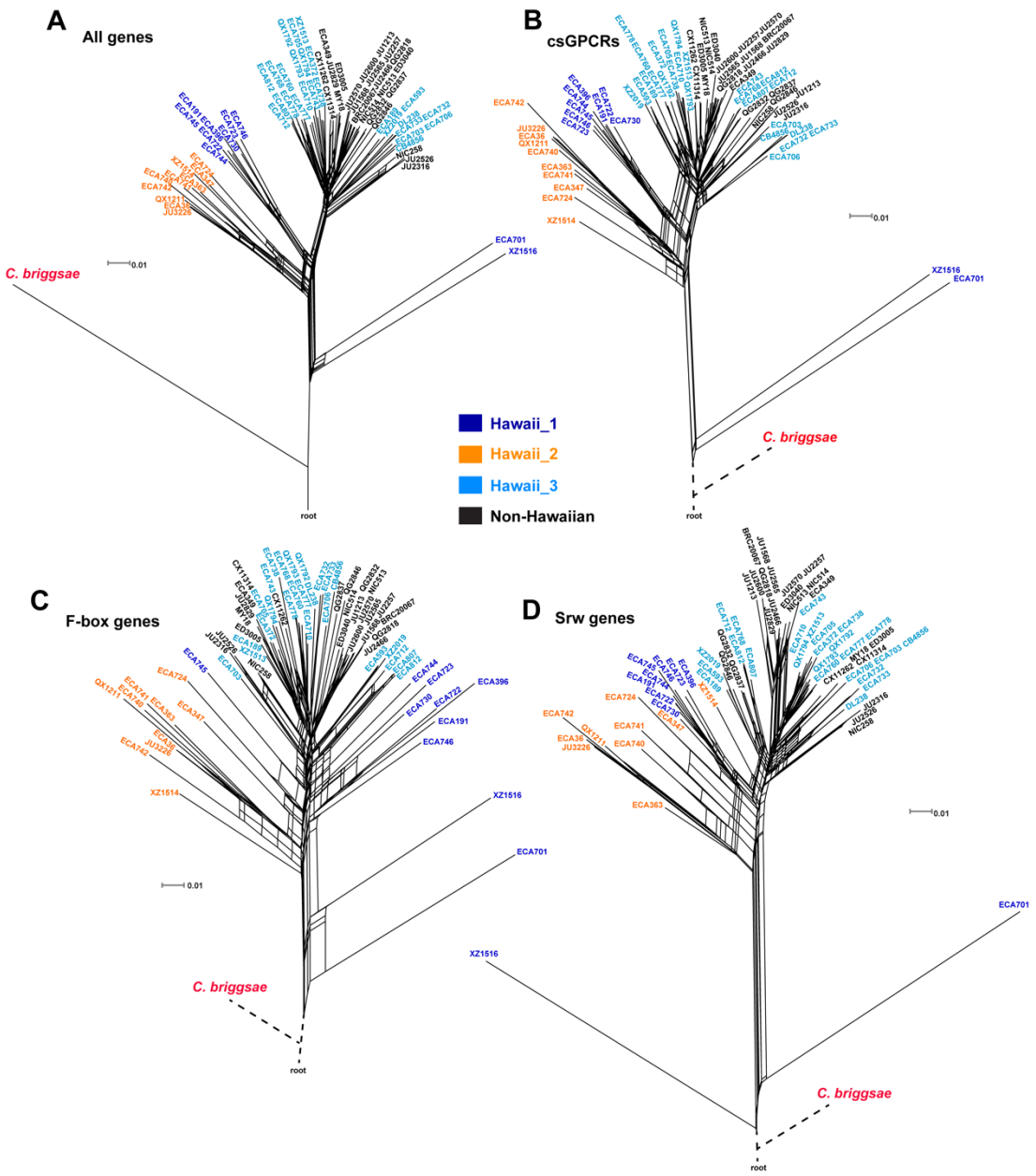


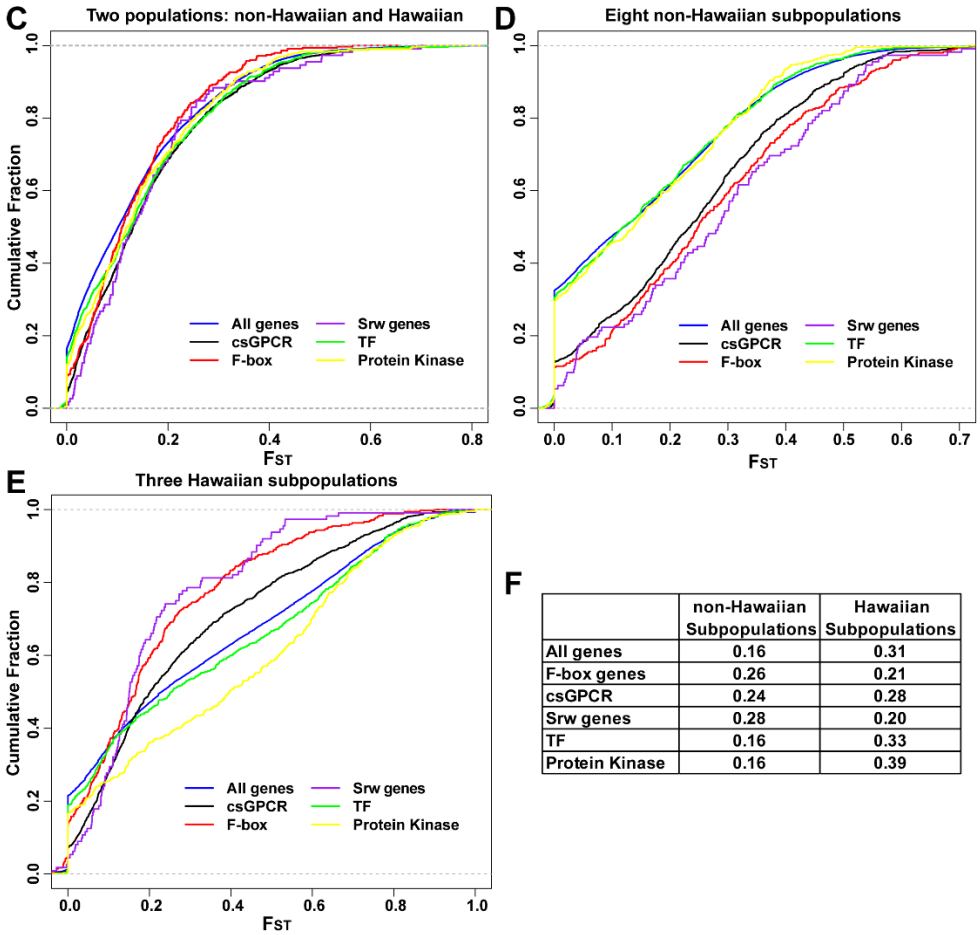

**FIG. 3. Phylogenetic relationship of the *C. elegans* wild isolates.**

**A**

| Normalized *Pi* | | | | | | |
|---|---|---|---|---|---|---|
| | All strains | Non-Hawaiian | Hawaiian | Hawaii_1 | Hawaii_2 | Hawaii_3 |
| All genes | 0.0010 | 0.0007 | 0.0017 | 0.0017 | 0.0012 | 0.0010 |
| csGPCRs | 0.0018 | 0.0013 | 0.0032 | 0.0035 | 0.0026 | 0.0019 |
| Srw GPCRs | 0.0034 | 0.0026 | 0.0059 | 0.0066 | 0.0052 | 0.0035 |
| F-box genes | 0.0042 | 0.0033 | 0.0069 | 0.0077 | 0.0064 | 0.0046 |
| TF | 0.0008 | 0.0005 | 0.0014 | 0.0014 | 0.0010 | 0.0009 |
| Protein kinase | 0.0004 | 0.0003 | 0.0007 | 0.0006 | 0.0004 | 0.0003 |

**B**

| Segregating sites (non-singletons) | | | | | | |
|---|---|---|---|---|---|---|
| | Exclusively Hawaiian site | | Exclusively nonHawaiian sites | | Shared sites | |
| | Number | Normalized | Number | Normalized | Number | Normalized |
| All genes | 6.3 (2.7) | 0.006 (0.0026) | 3.1 (1.3) | 0.0029 (0.0012) | 4.6 | 0.0045 |
| csGPCRs | 10.7 (4.4) | 0.011 (0.0044) | 4.3 (1.8) | 0.0043 (0.0019) | 9.1 | 0.0092 |
| Srw genes | 18.5 (7.1) | 0.017 (0.0067) | 6.3 (3.4) | 0.0060 (0.0033) | 19.5 | 0.0182 |
| F-box genes | 18.2 (7.4) | 0.018 (0.0071) | 7.2 (3.5) | 0.0073 (0.0035) | 20.2 | 0.0204 |
| TFs | 6.4 (2.7) | 0.005 (0.0022) | 3.2 (1.3) | 0.0025 (0.0010) | 4.5 | 0.0037 |
| Protein kinase | 5.6 (2.5) | 0.003 (0.0014) | 3.9 (1.7) | 0.0020 (0.0008) | 2.6 | 0.0014 |




**F**

| | non-Hawaiian Subpopulations | Hawaiian Subpopulations |
|---|---|---|
| All genes | 0.16 | 0.31 |
| F-box genes | 0.26 | 0.21 |
| csGPCR | 0.24 | 0.28 |
| Srw genes | 0.28 | 0.20 |
| TF | 0.16 | 0.33 |
| Protein Kinase | 0.16 | 0.39 |

**FIG. 4. csGPCR and F-box genes contribute to the large divergence of Hawaiian strains and the differentiation among non-Hawaiian subpopulations.**

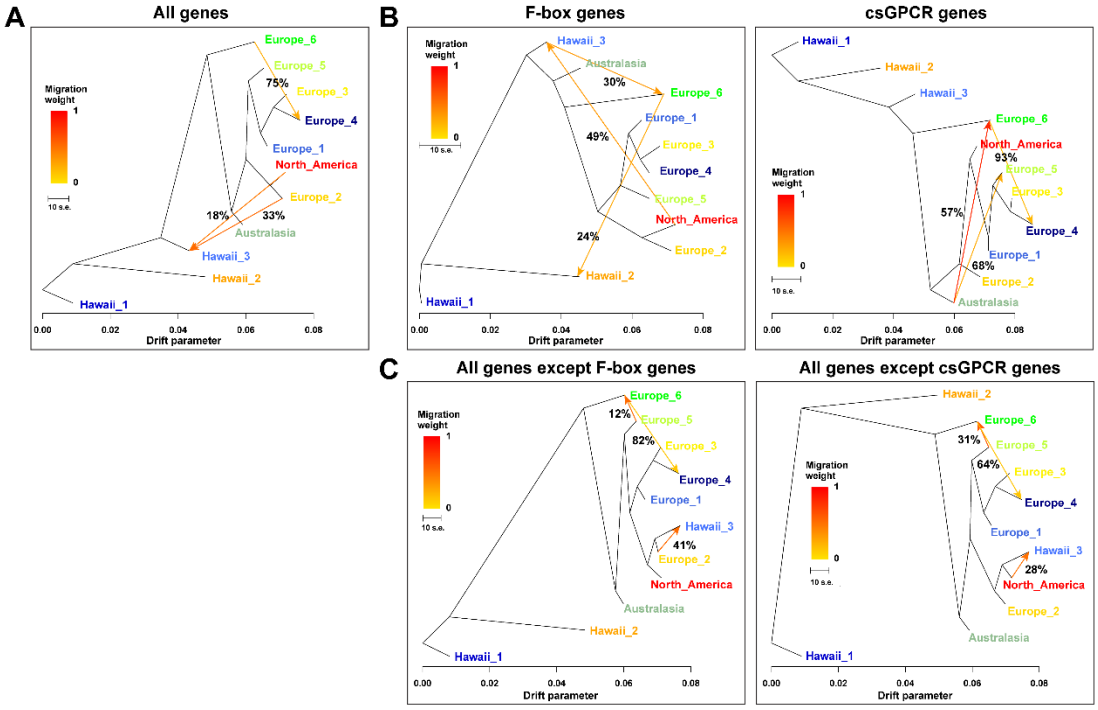
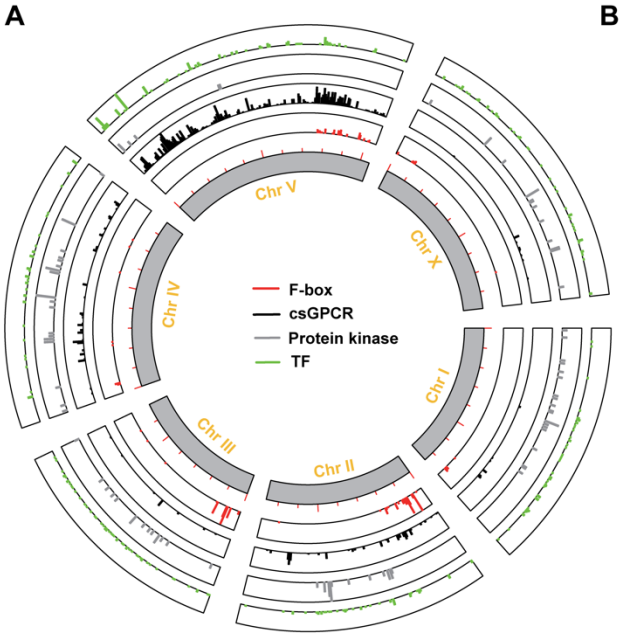
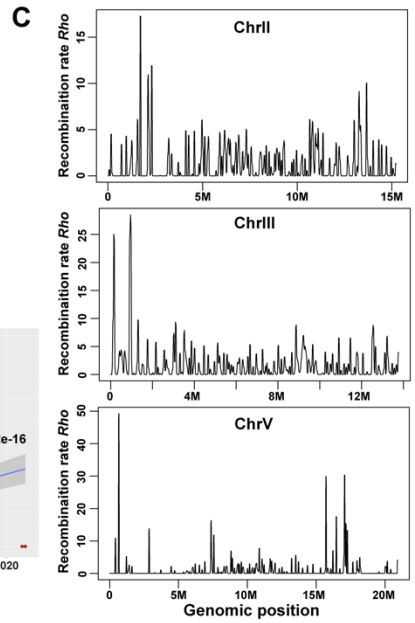



**FIG. 5. Gene flow of nonsynonymous variants of F-box and csGPCR genes among different subpopulations.**




B

| | Normalized *Pi* | | |
|---|---|---|---|
| | Left arm | Center | Right arm |
| | II: 0% ~ 25% | II: 25% ~ 75% | II: 75% ~ 100% |
| All genes | 0.0034 (998) | 0.0004 (1811) | 0.0010 (687) |
| csGPCR | 0.0025 (70) | 0.0006 (55) | 0.0016 (46) |
| F-box genes | 0.0058 (104) | 0.0004 (8) | 0.0028 (4) |
| | III: 0% ~ 25% | III: 25% ~ 75% | III: 75%~100% |
| All genes | 0.0024 (552) | 0.0003 (1572) | 0.0006 (547) |
| F-box genes | 0.0056 (65) | 0.0002 (6) | 0.0023 (9) |
| | IV: 0% ~ 25% | IV: 25% ~ 75% | IV: 75%~100% |
| All genes | 0.0009 (731) | 0.0003 (1957) | 0.0010 (617) |
| csGPCR | 0.0012 (28) | 0.0005 (85) | 0.0009 (42) |
| | V: 0% ~ 25% | V: 25% ~ 75% | V: 75%~100% |
| All genes | 0.0017 (1216) | 0.0005 (2662) | 0.0032 (1194) |
| csGPCR | 0.0021 (326) | 0.0009 (320) | 0.0034 (257) |
| Srw genes | 0.0025 (41) | 0.0025 (21) | 0.0055 (36) |
| F-box genes | 0.0018 (6) | 0.0011 (23) | 0.0055 (63) |



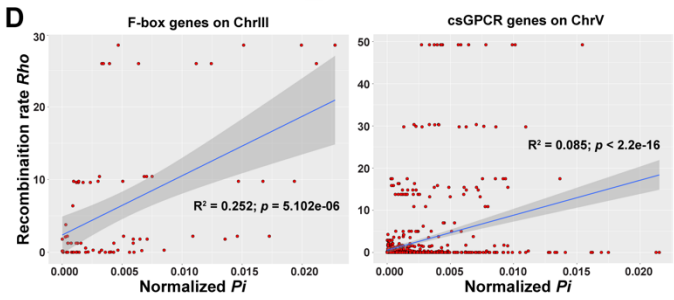




**FIG. 6. High recombination rate may contribute to the large diversity of csGPCR and F-box genes.**

**A**

| | Complete-case dataset | | Full dataset | |
|---|---|---|---|---|
| | DnaSP | PopGenome | VariScan_200 | VariScan_450 |
| Tajima's *D* | | | | |
| Syn | -0.93 | -1.09 | -0.92 | -0.93 |
| Nonsyn | -1.08 | -1.26 | -1.08 | -1.10 |
| Intron | -1.24 | -1.51 | -1.19 | -1.23 |
| UTR | -0.86 | -1.04 | -0.82 | -0.86 |
| Fay and Wu's *H* | | | | |
| Nonsyn | NA | NA | -3.87 | -3.72 |

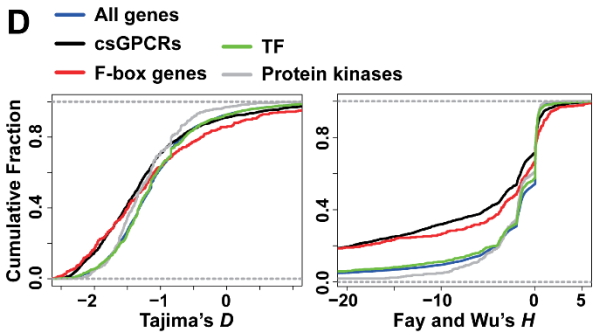
**B**

| | | Tajima's *D* < -2 | | Fay and Wu's *H* < -20 | | Overlap set | |
|---|---|---|---|---|---|---|---|
| | Total number | Number | Enrich. fold | Number | Enrich. fold | Number | Enrich. fold |
| csGPCRs | 1301 | 192 | 2.82 | 260 | 3.62 | 119 | 3.97 |
| F-box genes | 336 | 55 | 3.12 | 67 | 3.61 | 18 | 2.33 |
| All genes | 18797 | 985 | | 1038 | | 432 | |

**C**

| | Tajima's *D* | |
|---|---|---|
| | Mean | Median |
| All genes | -1.08 | -1.17 |
| csGPCRs | -1.20 | -1.35 |
| F-box genes | -1.08 | -1.28 |
| TF | -1.07 | -1.17 |
| Protein kinase | -1.22 | -1.31 |
| | Fay and Wu's *H* | |
| | Mean | Median |
| All genes | -4.17 | -1.38 |
| csGPCRs | -9.76 | -3.15 |
| F-box genes | -11.87 | -2.34 |
| TF | -4.28 | -1.44 |
| Protein kinase | -2.81 | -1.62 |

**D**




**E**

| csGPCR group | Fay and Wu's *H* | |
|---|---|---|
| | Mean | Median |
| Solo (293) | -14.10 | -4.69 |
| Str (567) | -10.17 | -3.85 |
| Sra (126) | -3.98 | -1.21 |
| Srg (315) | -7.31 | -2.35 |
| Solo - Srw (112) | -22.46 | -19.06 |
| Solo - Srz (65) | -6.84 | -1.19 |
| Solo - Srbc (69) | -14.14 | -9.79 |
| Solo - Srsx (37) | -3.04 | -1.09 |

**F**

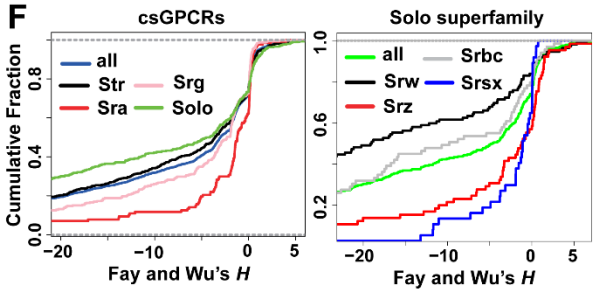


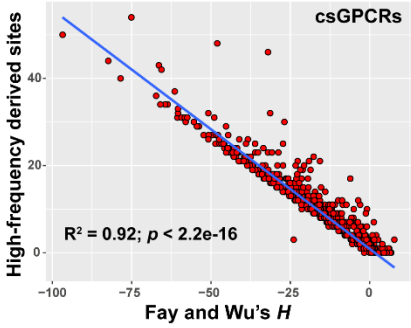

**FIG. 7. The signature of positive selection on F-box and csGPCR gene.**

**A**

| | Fay and Wu's *H* (XZ1516 as outgroup) | | | | | |
|---|---|---|---|---|---|---|
| | All strains | Non-Hawaiian | Hawaiian | Hawaii_1 | Hawaii_2 | Hawaii_3 |
| All genes | -4.17 | -4.24 | -2.48 | -2.24 | -2.64 | -3.45 |
| csGPCRs | -9.76 | -10.19 | -7.13 | -6.09 | -6.47 | -8.57 |
| Srw GPCRs | -22.46 | -23.68 | -17.69 | -14.40 | -13.54 | -20.91 |
| F-box genes | -11.87 | -12.31 | -6.53 | -4.22 | -6.10 | -9.22 |
| TF | -4.28 | -4.67 | -2.85 | -2.74 | -3.14 | -3.76 |
| Protein kinase | -2.81 | -3.03 | -1.43 | -1.56 | -1.83 | -2.45 |

**B**

| | Number of high frequency derived sites with XZ1516 as the outgroup | | |
|---|---|---|---|
| | Among all strains | Among Hawaiian | Among non-Hawaiian |
| All gene | 2.73 | 2.58 | 2.76 |
| csGPCRs | 6.24 | 6.18 | 6.28 |
| Srw GPCRs | 14.19 | 13.62 | 14.22 |
| F-box | 8.80 | 8.07 | 8.87 |
| TF | 2.97 | 3.00 | 3.01 |
| Protein Kinase | 1.90 | 1.63 | 1.93 |




**C**

| | Fay and Wu's *H* in non-Hawaiian Population | | | |
|---|---|---|---|---|
| Outgroup | XZ1516 (ancestral) | ECA701 (ancestral) | ECA396 (Hawaii_1) | ECA742 (Hawaii_2) |
| All genes | -4.24 | -4.03 | -2.42 | -2.90 |
| csGPCRs | -10.19 | -11.01 | -4.09 | -5.61 |
| F-box genes | -12.79 | -16.78 | -9.54 | -9.51 |

**D**

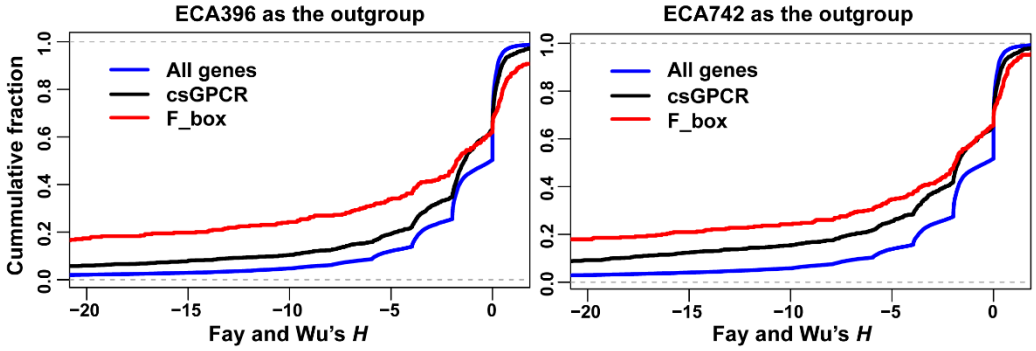



**FIG. 8. Accumulation of high-frequency derived alleles in F-box and csGPCR genes in non-Hawaiian population.**

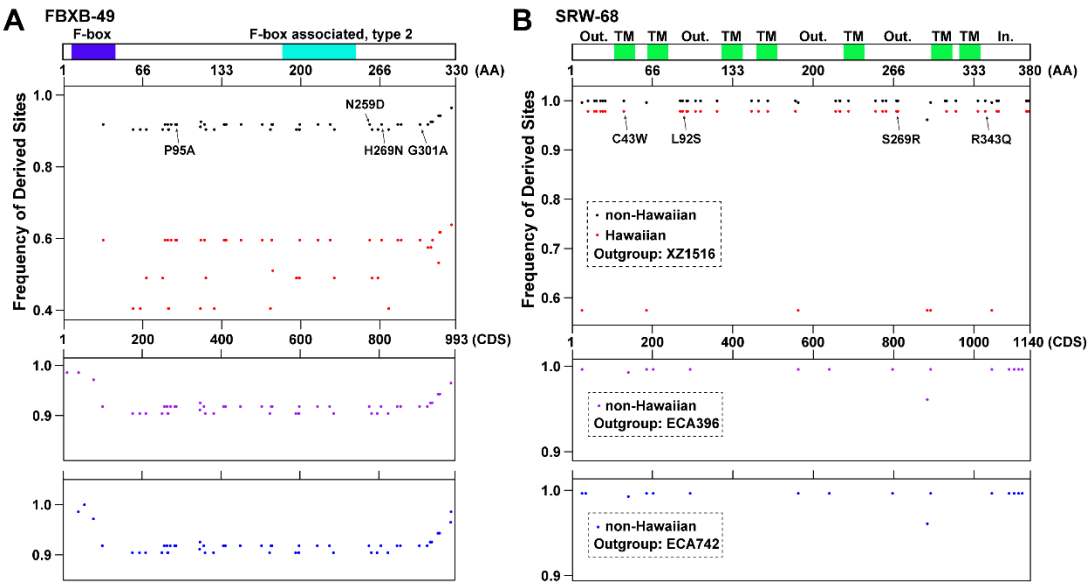
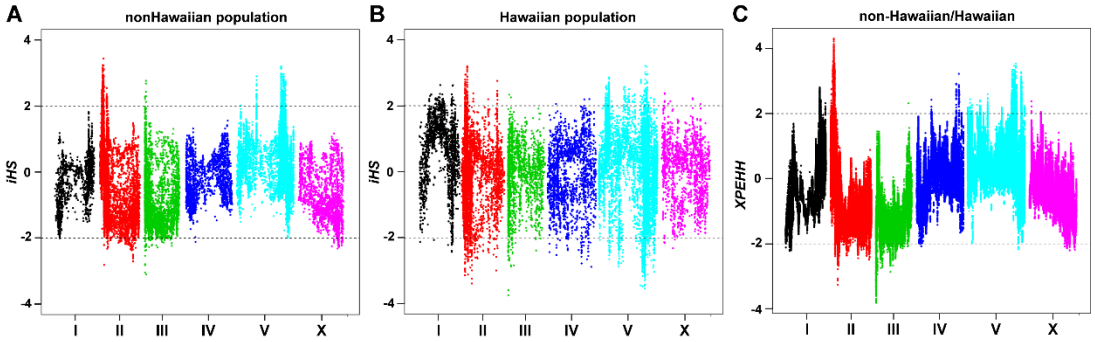



**FIG. 9. High-frequency derived sites were mapped to the substrate recognition domain of a reprehensive F-box protein and the extracellular loops of a representative csGPCR.**




D

| | non-Hawaiian | Hawaiian | nonHawaiian/Hawaiian |
|---|---|---|---|
| | \|iHS\| > 2 | \|iHS\| > 2 | XPEHH > 2 |
| # of genes | 335 | 698 | 193 |
| # of F-box genes | 27 (4.5 fold) | 40 (3.2 fold) | 31 (3.8 fold) |
| # of csGPCRs | 34 (1.5 fold) | 126 (2.6 fold) | 31 (2.3 fold) |
| | Extended region | Extended region | Extended region |
| # of genes | 1277 | 3110 | 596 |
| # of F-box genes | 83 (3.6 fold) | 112 (2.0 fold) | 82 (7.7 fold) |
| # of csGPCRs | 100 (1.1 fold) | 485 (2.3 fold) | 98 (2.4 fold) |
| | \|iHS\| top 0.05% | \|iHS\| top 0.05% | XEPHH top 0.05% |
| # of genes | 11 | 42 | 26 |
| # of F-box genes | 2 (10.1 fold) | 6 (8.0 fold) | 8 (17 fold) |
| # of csGPCRs | 4 (5.2 fold) | 18 (6.2 fold) | 3 (1.6 fold) |

E

| | Fay and Wu's *H* (XZ1516 as outgroup) | | |
|---|---|---|---|
| | Left arm | Center | Right arm |
| | II: 0% ~ 25% | II: 25% ~ 75% | II: 75% ~ 100% |
| All genes | -4.78 | -1.76 | -1.66 |
| F-box | -11.1 | -1.99 | -6.58 |
| | III: 0% ~ 25% | III: 25% ~ 75% | III: 75%~100% |
| All genes | -3.32 | -1.35 | -2.34 |
| F-box | -5.69 | -1.10 | -5.37 |
| | V: 0% ~ 25% | V: 25% ~ 75% | V: 75%~100% |
| All genes | -12.33 | -5.66 | -12.55 |
| csGPCRs | -15.9 | -11.47 | -11.32 |
| Srw genes | -24.96 | -38.58 | -14.2 |

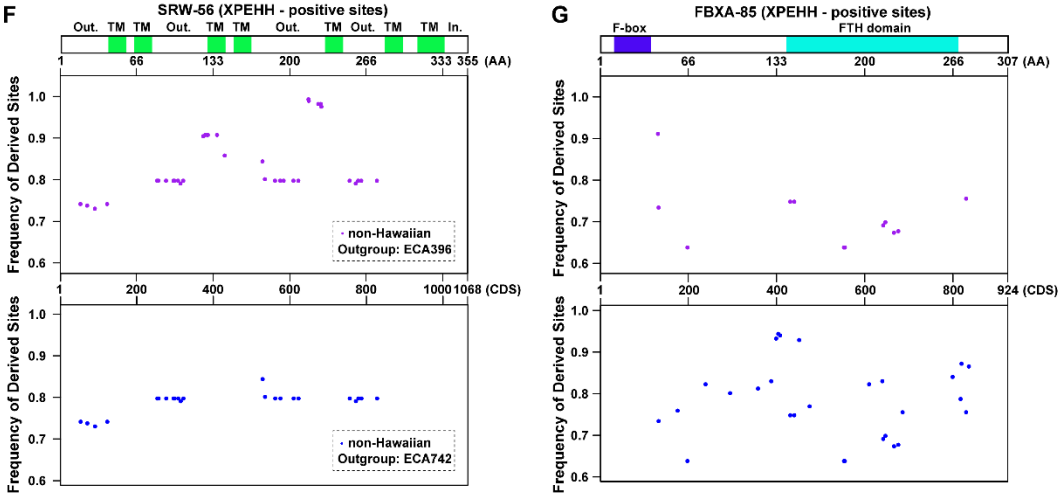



**FIG. 10. csGPCRs and F-box genes are enriched in the genomic regions with selective footprint identified by extended haplotype homozygosity (EHH) analysis.**

**FIG. legends:**

**FIG. 1.** Large genetic polymorphism of csGPCR and F-box genes. (A) Genes with large CDS length-normalized *Pi* tend to be enriched in csGPCR and F-box gene families. *Pi* values of individual genes can be found in supplementary table S1, Supplementary Material online. (B) A histogram for the distribution of normalized *Pi* across all genes. (C) The cumulative distribution of the normalized *Pi* values for all genes, csGPCR, F-box, transcription factor (TF) and Protein kinases genes. (D) The mean and median of normalized *Pi* for different gene families. (E) The mean and median of normalized *Pi* for different csGPCR superfamilies and families. The number of genes are in the parentheses. For statistical significance in a non-parametric Wilcoxon's rank-sum test, ns means not significant, a single asterisk means $p < 0.05$, and double asterisks mean $p < 0.01$. Similar annotations apply for the rest of the fig.s.

**FIG. 2.** Population structure and geographical distribution of *C. elegans* wild isolates. (A) The ancestral population proportions of wild isolates (color-coded) inferred by Admixture, when the number of population (K value) was set at 11. The bar indicates the ancestral fraction for each strain. Only strains with one ancestral proportion larger than 0.5 were shown (see supplementary table S2, Supplementary Material online for the data on all strains). Each group (or subpopulation) was named after the geographical location of most strains in that group. The names of the strains from the three Hawaiian subpopulations (Hawaii_1, Hawaii_2 and Hawaii_3) were shown in the enlarged panels.

Names of some non-Hawaiian strains were also shown. (B-C) Geographical locations of strains isolated in Europe (B) and Hawaii (C). Each dot represents a strain, but some dots are entirely overlapping given the scale of the map. The latitude and longitude data for each strain was obtained from CeNDR. Color code is consistent with (A).

**FIG. 3.** Phylogenetic relationship of the *C. elegans* wild isolates. Neighbour-joining nets plotted using the nonsynonymous SNVs of all genes (A), csGPCRs (B), F-box genes (C), or *Srw* genes (D). *C. briggsae* was used as the outgroup for tree construction. Three representative non-Hawaiian strains (in black) with high ancestral population fraction were chosen from each of the eight non-Hawaiian groups. Strains in the three Hawaiian groups are labelled according to the color code in fig. 2. To fit the trees into one figure, some branches connecting *C. briggsae* and the root are manually shortened (dashed lines). Trees in the original sizes can be found in supplementary fig. S4, Supplementary Material online.

**FIG. 4.** csGPCR and F-box genes contribute to the large divergence of Hawaiian strains and the differentiation among non-Hawaiian subpopulations. (A) The mean of CDS length-normalized *Pi* of all genes, csGPCRs, *Srw* genes, F-box genes, TF, and Protein kinase for non-Hawaiian and Hawaiian populations, as well as the three Hawaiian subpopulations (see the grouping in Materials and Methods). (B) The average number of segregating sites that belong to only Hawaiian or non-Hawaiian strains and the sites

that are shared by Hawaiian and non-Hawaiian strains for the six gene families. The number is also normalized to the CDS length of individual genes. The number of non-singleton segregating sites are in the parentheses. (C-E) The cumulative distribution of Hudson's $F_{ST}$ values for different gene families between the non-Hawaiian and Hawaiian populations (C), among the eight non-Hawaiian subpopulations (D), and among the three Hawaiian subpopulations (E). (F) The average $F_{ST}$ value of different gene families among non-Hawaiian and among Hawaiian subpopulations.

**FIG. 5.** Gene flow of nonsynonymous variants of F-box and csGPCR genes among different subpopulations. Gene flow among the three Hawaiian and eight non-Hawaiian subpopulations inferred by TreeMix using the nonsynonymous SNVs of all genes (A), the F-box genes, csGPCRs (B) and all genes excluding the F-box or csGPCR genes (C). Three migration events were allowed and 1000 bootstrap replicates were run with "Hawaii_1" set as the outgroup. The three most probable migration events from the 1000 bootstraps were shown and the percentage indicated the probability for the migration event. The color of the arrow indicates migration weight.

**FIG. 6.** High recombination rate may contribute to the large diversity of csGPCR and F-box genes. (A) Genomic location of F-box, csGPCR, Protein kinase, and TF genes plotted using TBtools (Chen et al. 2018). (B) The mean of normalized *Pi* values for all genes, csGPCR, and F-box genes in the arm or center of chromosome (Chr) II, III, IV

and V. Chromosomes were divided into three regions according to genomic coordinates: the left arm (one-fourth of the chromosome from the start), the center (the central half), and the right arm (one-fourth of the chromosome close to the end). The values in parentheses indicated gene numbers. (C) Recombination rates (*Rho*) across Chr II, III and V in 50-kb windows. (D) The Pearson correlation between recombination rate and *Pi* for F-box genes on Chr III and csGPCR genes on Chr V.

**FIG. 7.** The signature of positive selection on F-box and csGPCR gene. (A) The mean values of Tajima's *D* and Fay and Wu's *H* computed using different software (DnaSP, PopGenome and VariScan). Syn means synonymous SNVs; Nonsyn, nonsynonymous SNVs; UTR, SNVs in untranslated regions; NA, not calculated. Complete-case dataset only contains SNVs with no missing data; full dataset contains SNVs with valid genotype data for at least either 200 or 400 alleles out of the total 660 alleles (VariScan_200 and VariScan_450 respectively). (B) Enrichment of csGPCR and F-box genes among the genes with Tajima's $D < -2$ and Fay and Wu's $H < -20$, respectively. Overlap set include genes that fits both criteria. (C) The mean and median of Tajima's *D* and Fay and Wu's *H* values of all genes, csGPCRs, F-box, TF, and Protein kinase. (D) The cumulative distribution of different gene families. (E) The mean and median of Fay and Wu's *H* values of genes in csGPCR superfamilies and *Solo* gene families. The number of genes are in parentheses. (F) The cumulative distribution of genes in csGPCR subfamilies and *Solo* families. The statistical significance was determined by

Wilcoxon rank-sum test.

**FIG. 8.** Accumulation of high-frequency derived alleles in F-box and csGPCR genes in non-Hawaiian population. (A) The average Fay and Wu's *H* values of all genes, csGPCRs, *Srw* genes, F-box genes, TFs, and Protein kinase for the non-Hawaiian and Hawaiian populations, as well as the three Hawaiian subpopulations. (B) The average number of high frequency (>50%) derived sites in genes from different gene families with XZ1516 as the outgroup. The right panel shows the Pearson correlation between Fay and Wu's *H* and number of high-frequency derived sites. (C) The average Fay and Wu's *H* value in non-Hawaiian population calculated using different strains as the outgroup. XZ1516 and ECA701 are distant from all other strains and may be considered to contain the most ancestral alleles. ECA396 and ECA742 are representative strains from "Hawaii_1" and "Hawaii_2" subpopulations. (D) The cumulative distribution of the *H* values of all genes, csGPCR, or F-box genes calculated using ECA396 or ECA742 as the outgroup.

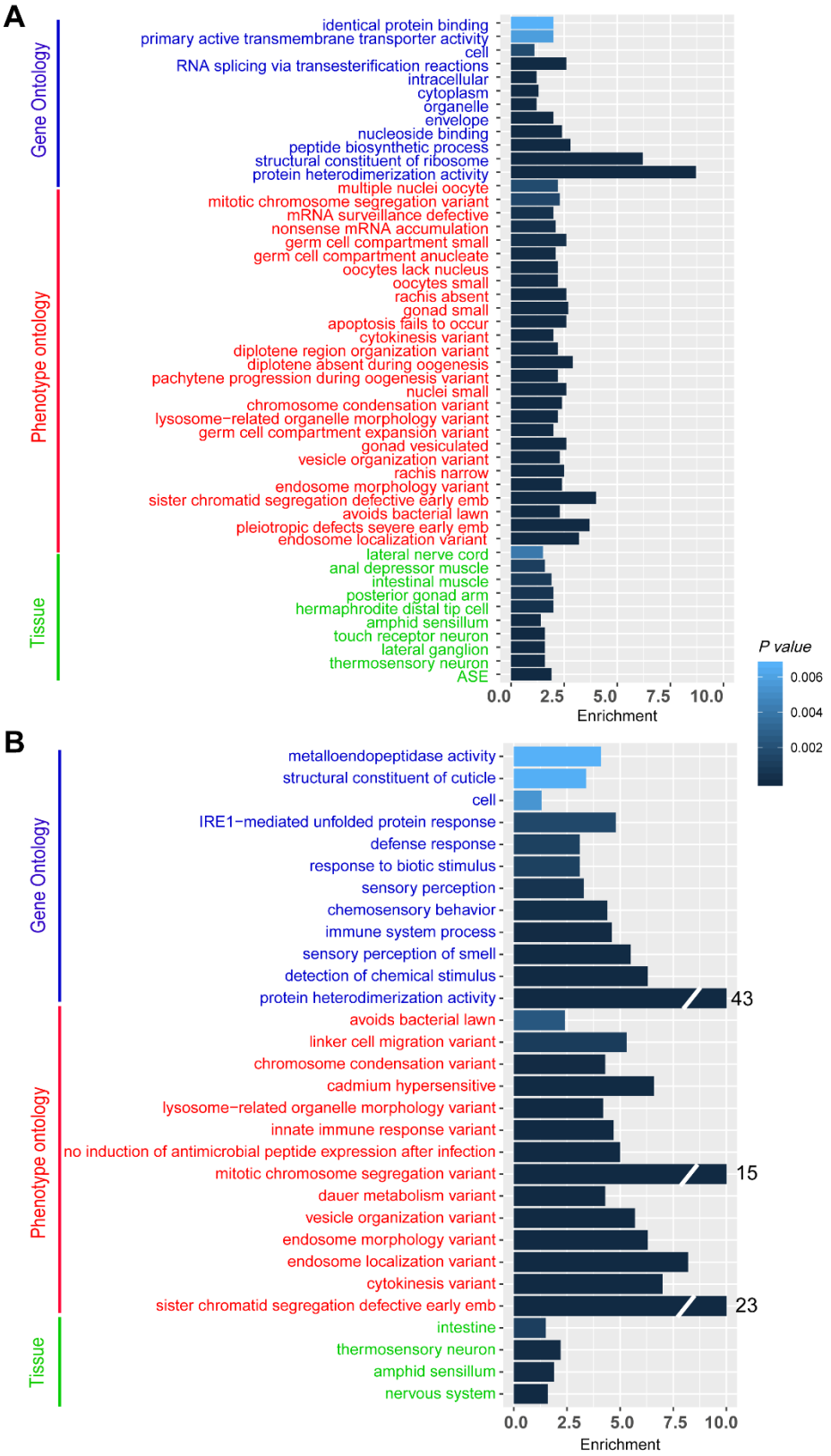
**FIG. 9.** High-frequency derived sites were mapped to the substrate recognition domain of a reprehensive F-box protein and the extracellular loops of a representative csGPCR. (A) A The domain structure of a F-box protein encoded by *fbxb-49*. The F-box domain is in blue, and the type 2 F-box associated domain, likely involved in binding substrate, is in cyan. (B) The domain structure of a csGPCR encoded by *srw-68*. The predicted

transmembrane (TM) domain is in green. Extracellular loops (Out.) and Intracellular (In.) tails are indicated. In both (A) and (B), the panel immediately below the domain structure indicate the position of high-frequency (>0.5) derived sites in non-Hawaiian populations using XZ1516 as the outgroup. Y-axis indicate the frequency of the derived alleles among the non-Hawaiian population (black dots) or the Hawaiian population (red dots). Each dot indicates a nonsynonymous SNVs. SNVs causing amino acid substitution with PROVEAN score below -2.5 were shown. The lower two panels showed the high-frequency derived sites in the non-Hawaiian population calculated using ECA396 (purple dots) and ECA742 (blue dots) as the outgroup.

**FIG. 10.** csGPCRs and F-box genes are enriched in the genomic regions with selective footprint identified by extended haplotype homozygosity (EHH) analysis. (A-C) Manhattan plots of the extent of haplotype homozygosity measured by the integrated Haplotype Score (iHS) within the non-Hawaiian population (A) and Hawaiian population (B). (C) Regions of selection in non-Hawaiian population but not the Hawaiian population indicated by the Manhattan plots of cross-population EHH (XPEHH). (D) The number of csGPCRs and F-box genes that contain SNVs with significant iHS or XPEHH and their folds of enrichment. For extended regions, significant SNVs that are less than 50-kb apart were connected to generate regions with selective footprints. (E) The mean Fay and Wu's *H* values for all genes, csGPCRs, and F-box genes in the arms and the center of chromosome (Chr) II, III, and V. (F) The

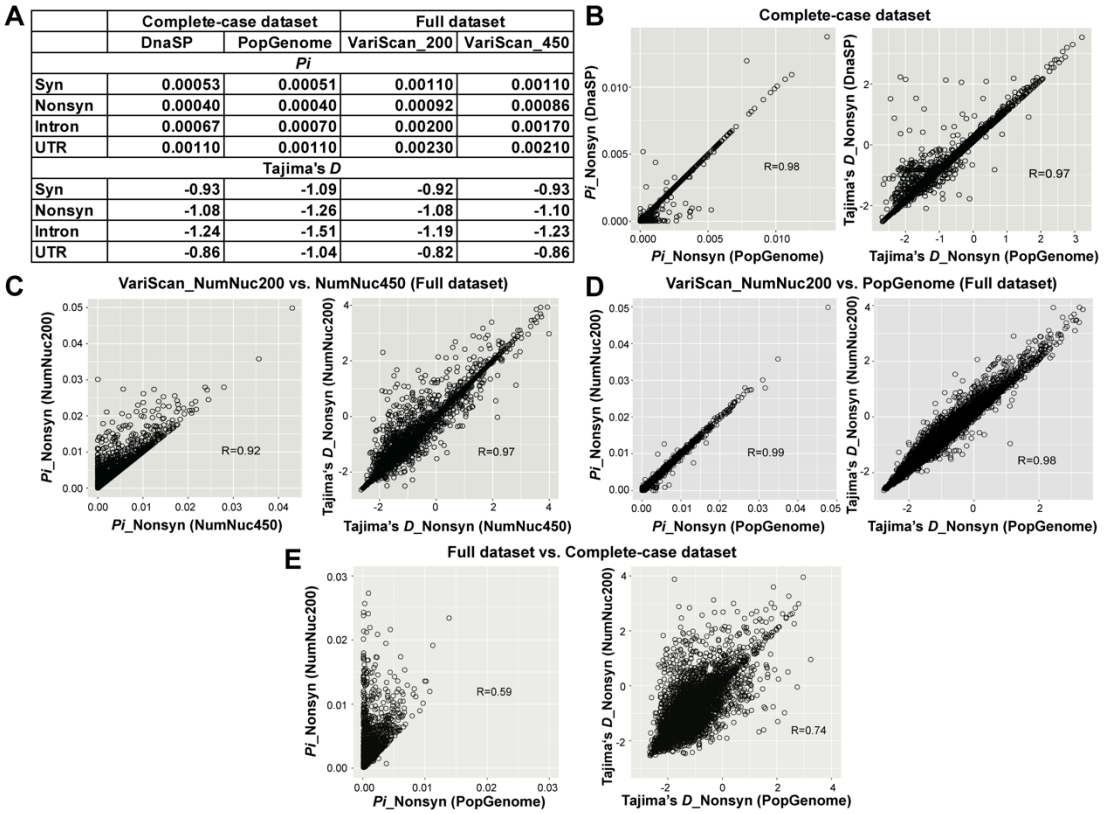
domain structure of a representative csGPCR coded by *srw-56*; the predicted transmembrane (TM) domain is in green, and extracellular loops (Out.) and intracellular (In.) tails are also indicated. (G) The domain structure of a representative F-box protein coded by *fbxa-85*; the F-box domain is in blue and the FTH domain in cyan. Among the sites whose XPEHH > 2 in the two genes, the ones that are also high-frequency (> 0.5) derived sites with ECA396 (purple dots) and ECA742 (blue dots) as the outgroup are shown.

**Supplemental Figures**




**Supplementary fig. S1. Gene ontology enrichment analysis of genes with no nonsynonymous SNVs.** (A) Gene enrichment analysis for the 1143 genes with no nonsynonymous SNVs and small indels in the CDS region. (B) Gene enrichment analysis for the 302 genes that have no nonsynonymous SNVs and small indels in the whole gene. Results of gene ontology

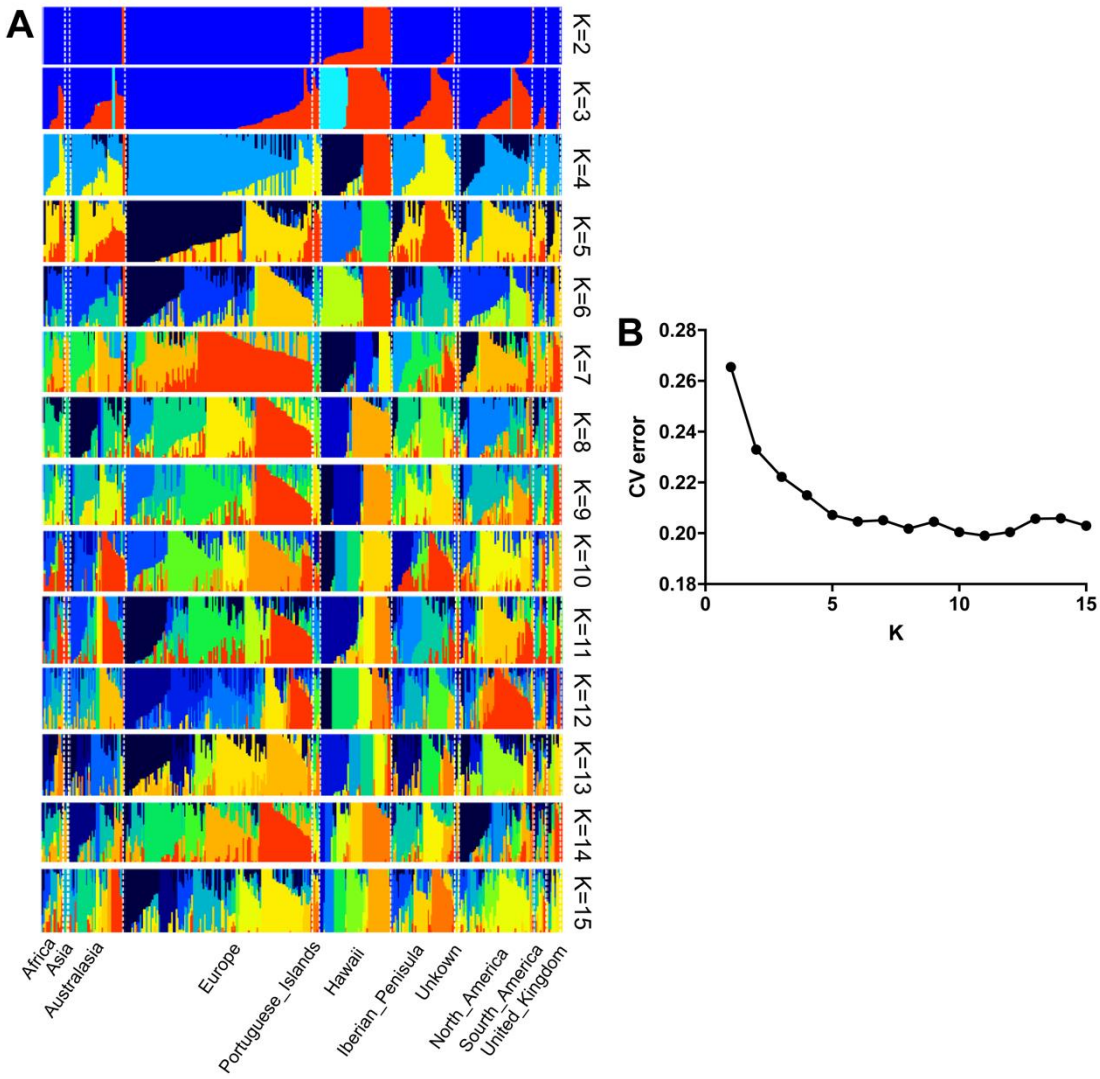
enrichment (blue), phenotype enrichment (red), and tissue expression enrichment (green) were obtained using the enrichment analysis tools on Wormbase (WS275).



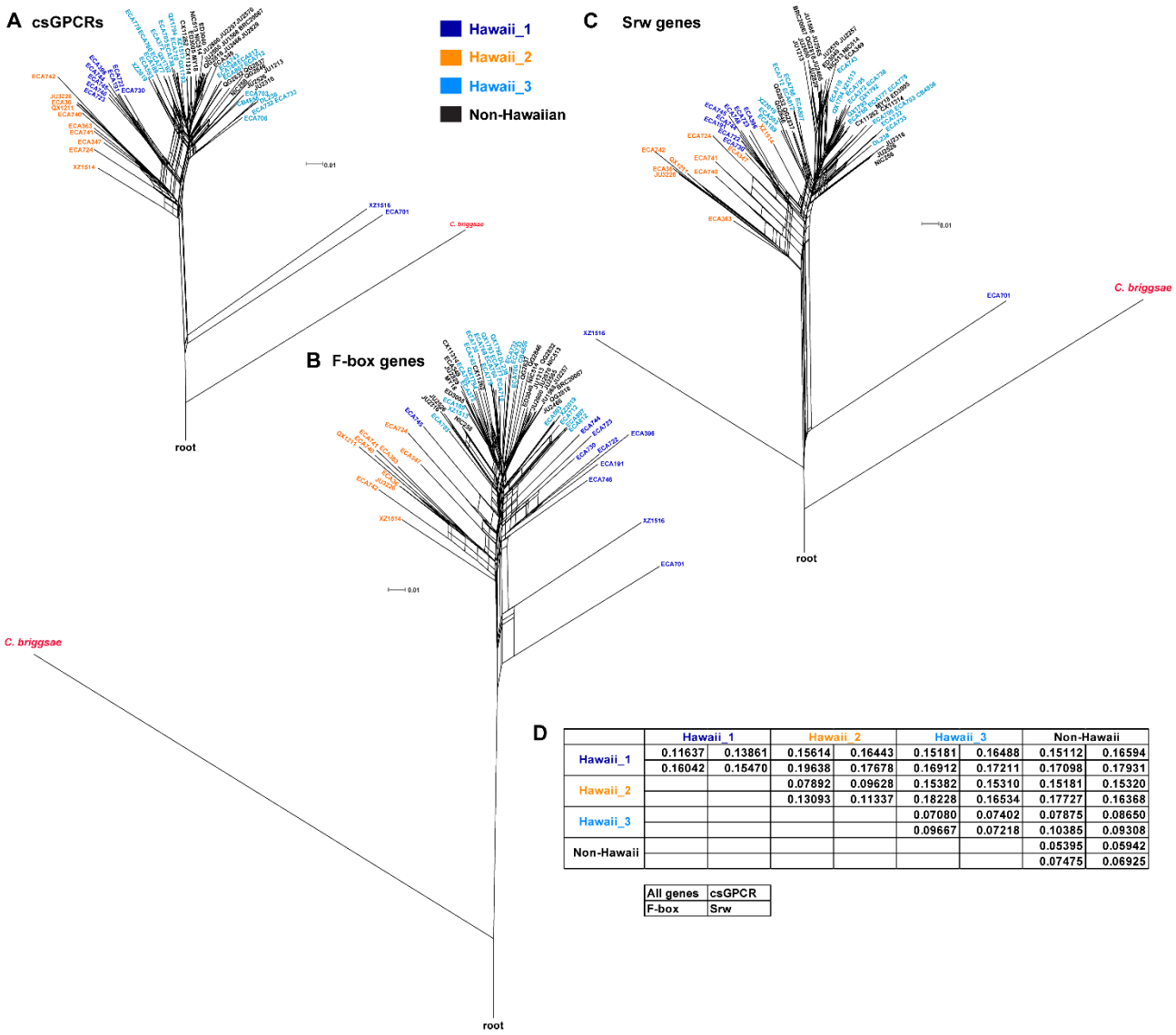
| | Complete-case dataset | | Full dataset | |
|---|---|---|---|---|
| | DnaSP | PopGenome | VariScan_200 | VariScan_450 |
| | *Pi* | | | |
| Syn | 0.00053 | 0.00051 | 0.00110 | 0.00110 |
| Nonsyn | 0.00040 | 0.00040 | 0.00092 | 0.00086 |
| Intron | 0.00067 | 0.00070 | 0.00200 | 0.00170 |
| UTR | 0.00110 | 0.00110 | 0.00230 | 0.00210 |
| | Tajima's *D* | | | |
| Syn | -0.93 | -1.09 | -0.92 | -0.93 |
| Nonsyn | -1.08 | -1.26 | -1.08 | -1.10 |
| Intron | -1.24 | -1.51 | -1.19 | -1.23 |
| UTR | -0.86 | -1.04 | -0.82 | -0.86 |



**Supplementary fig. S2. Computation of nucleotide polymorphism (*Pi*) and Tajima's *D* for synonymous, nonsynonymous, intron, and UTR SNVs in genes using different software.** (A) The mean of CDS length-normalized *Pi* and Tajima's *D* calculated for synonymous, nonsynonymous, intron and UTR SNVs in all genes. (B) The correlation of normalized *Pi* and Tajima's *D* calculated by DnaSP and PopGenome using the complete-case dataset that only contains SNVs with no missing genotype data among the 330 strains. Full dataset contains SNVs with valid genotype data for at least either 200 or 400 alleles out of the total 660 alleles (VariScan_200 and VariScan_450, respectively). (C) The correlation of normalized *Pi* and Tajima's *D* for nonsynonymous SNVs calculated by VariScan with different threshold for valid number of alleles (NumNuc = 200 and 450). (D) The correlation of normalized *Pi* and Tajima's *D* computed by VariScan (NumNuc = 200) and PopGenome including all sites. (E) The correlation of normalized *Pi* and Tajima's *D* computed using the full dateset (VariScan; NumNuc = 200) and the complete-case dateset (PopGenome; excluding sites with any missing genotype data).




**Supplementary fig. S3. Population structure predicted by Admixture using a range of K values.** (A) Population structure of the 330 wild isolates constructed using Admixture with the number of ancestral populations (K value) ranging from 2 to 15. The strains are arranged based on their geographic locations and are separated by white dash lines. (B) The cross validation (CV) error for a range of K values from 2 to 15.

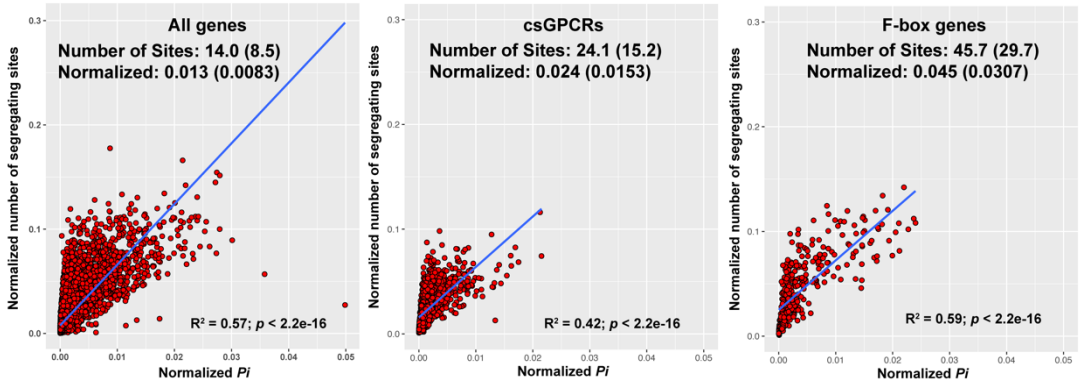



D

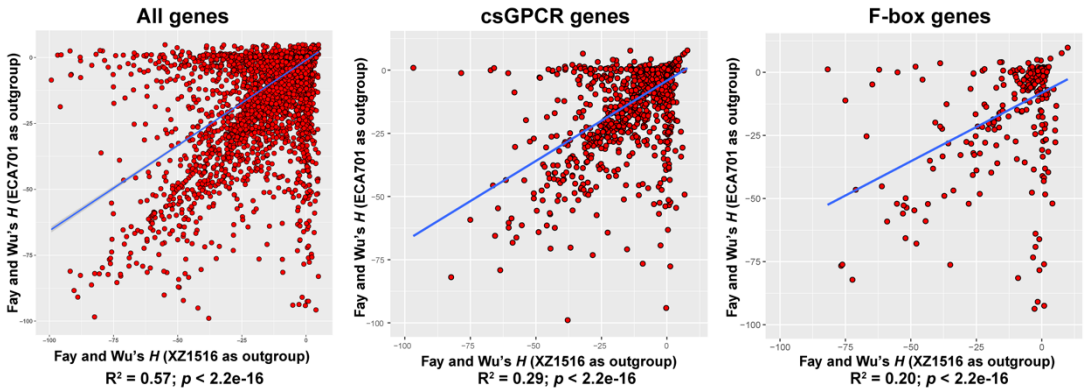
| | Hawaii_1 | | Hawaii_2 | | Hawaii_3 | | Non-Hawaii | |
|---|---|---|---|---|---|---|---|---|
| Hawaii_1 | 0.11637 | 0.13861 | 0.15614 | 0.16443 | 0.15181 | 0.16488 | 0.15112 | 0.16594 |
| | 0.16042 | 0.15470 | 0.19638 | 0.17678 | 0.16912 | 0.17211 | 0.17098 | 0.17931 |
| Hawaii_2 | | | 0.07892 | 0.09628 | 0.15382 | 0.15310 | 0.15181 | 0.15320 |
| | | | 0.13093 | 0.11337 | 0.18228 | 0.16534 | 0.17727 | 0.16368 |
| Hawaii_3 | | | | | 0.07080 | 0.07402 | 0.07875 | 0.08650 |
| | | | | | 0.09667 | 0.07218 | 0.10385 | 0.09308 |
| Non-Hawaii | | | | | | | 0.05395 | 0.05942 |
| | | | | | | | 0.07475 | 0.06925 |

| | |
|---|---|
| All genes | csGPCR |
| F-box | Srw |

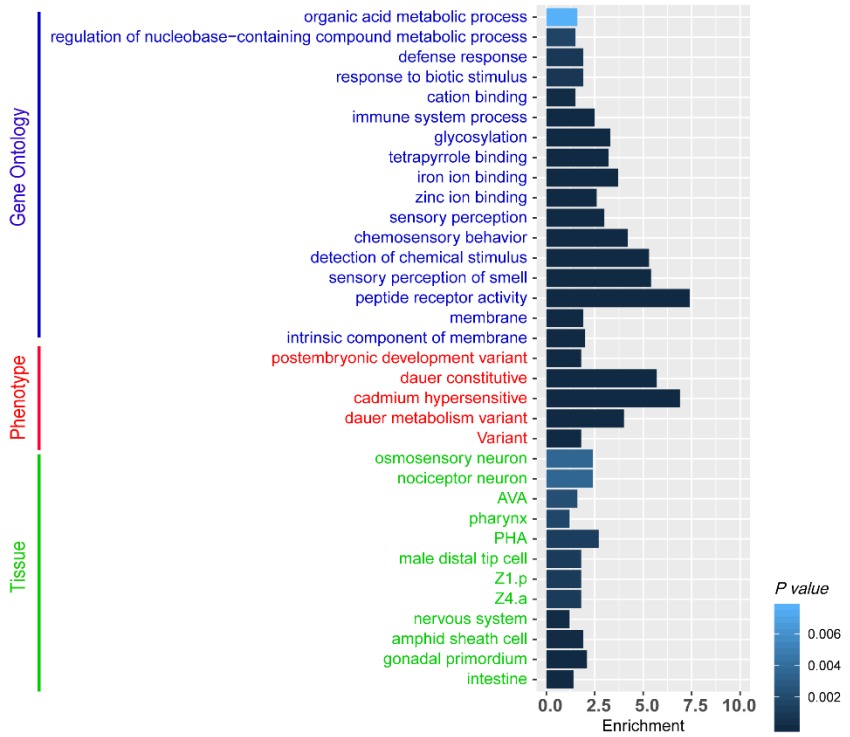
**Supplementary fig. S4. Phylogenetic relationship of *C. elegans* F-box and csGPCR genes.** Neighbour-joining nets plotted using the nonsynonymous SNVs of csGPCRs (A), F-box genes (B), or *Srw* genes (C). *C. briggsae* was used as the outgroup, and the trees are shown in their original sizes. (D) The means of pairwise distances among Hawaiian and non-Hawaiian subpopulations in the neighbour-joining nets. The four values in each pairwise comparison represent the mean distance calculated using SNVs from all genes, csGPCRs, F-box genes, or *Srw* genes.



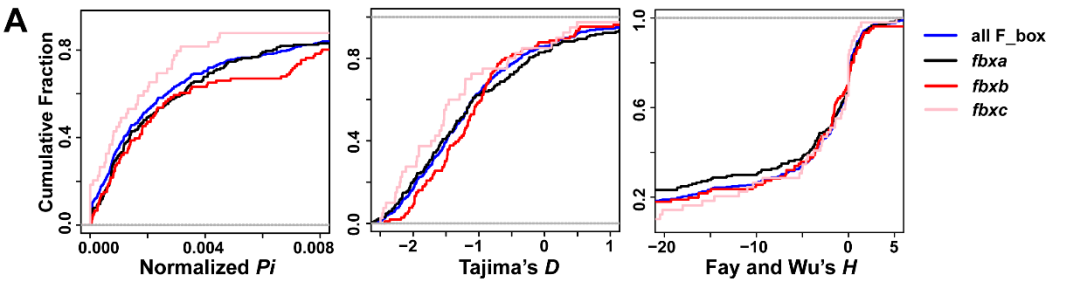

**Supplementary fig. S5. Correlation between polymorphism and the number of segregating sites.** Significantly positive Pearson correlation between CDS length-normalized number of segregating sites and normalized *Pi*, for all genes, csGPCRs, and F-box genes. The number in parentheses indicate the corresponding values excluding singletons.




**Supplementary fig. S6. Correlation between Fay and Wu's *H* values with XZ1516 and ECA701 as outgroups.** Significantly positive Pearson correlation between the *H* values calculated using XZ1516 and ECA701 as outgroups for all genes, csGPCRs, and F-box genes.




**Supplementary fig. S7. Gene enrichment analysis for genes whose Fay and Wu's *H* lower than -20.** Results of gene ontology enrichment (blue), phenotype enrichment (red), and tissue expression enrichment (green) were obtained using the enrichment analysis tools on Wormmbase (WS275).

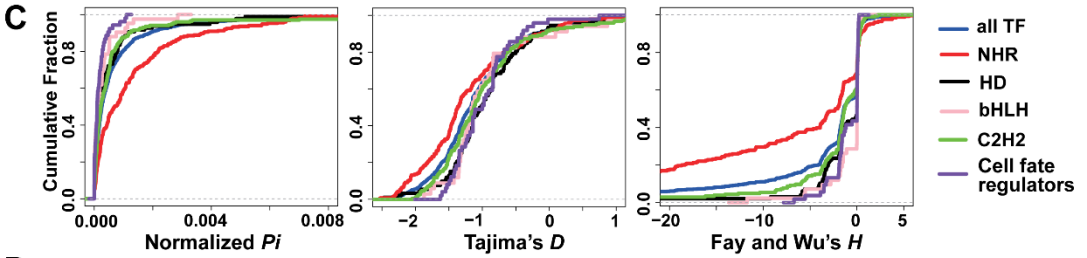



B

| Gene | Tajima's *D* | *Pi* | Fay and Wu's *H* |
|---|---|---|---|
| *skr-1* | -0.84 | 0.00002 | 0.01 |
| *skr-2* | -0.51 | 0.00055 | -3.18 |
| *skr-3* | -1.13 | 0.00100 | -4.40 |
| *skr-4* | -1.36 | 0.00600 | -7.84 |
| *skr-5* | 1.14 | 0.02300 | -17.06 |
| *skr-6* | 0.20 | 0.00410 | 0.99 |
| *skr-7* | -2.03 | 0.00041 | -1.75 |
| *skr-8* | -1.80 | 0.00020 | -1.87 |
| *skr-9* | -1.14 | 0.00100 | 0.25 |
| *skr-10* | -2.00 | 0.00014 | -9.86 |

| Gene | Tajima's *D* | *Pi* | Fay and Wu's *H* |
|---|---|---|---|
| *skr-12* | -1.82 | 0.00030 | 0.15 |
| *skr-13* | -1.02 | 0.00058 | 0.09 |
| *skr-14* | -1.40 | 0.00110 | -1.50 |
| *skr-15* | -2.05 | 0.00130 | -37.58 |
| *skr-16* | -1.04 | 0.00008 | 0.04 |
| *skr-17* | 0.41 | 0.00140 | -1.44 |
| *skr-18* | -0.55 | 0.00110 | -2.72 |
| *skr-19* | -0.79 | 0.00004 | 0.02 |
| *skr-20* | -1.11 | 0.00005 | -1.96 |
| *skr-21* | -0.40 | 0.00084 | -0.61 |

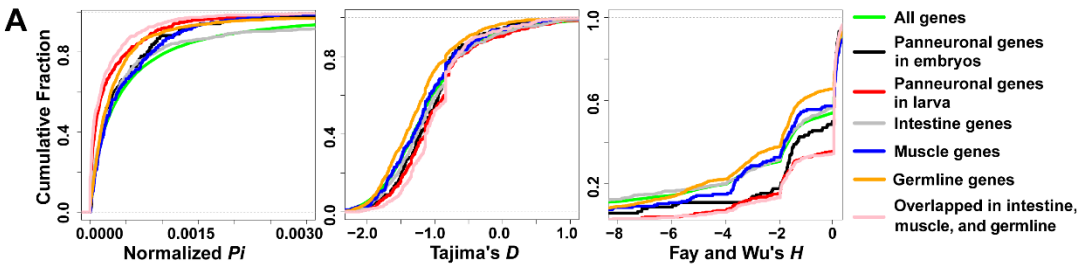
| Gene | Tajima's *D* | *Pi* | Fay and Wu's *H* |
|---|---|---|---|
| *cul-1* | -1.16 | 0.00029 | -3.49 |
| *cul-2* | -1.33 | 0.00018 | 0.41 |
| *cul-3* | -1.46 | 0.00002 | -3.89 |
| *cul-4* | -1.34 | 0.00007 | -1.83 |
| *cul-5* | -1.69 | 0.00003 | 0.07 |
| *cul-6* | -0.83 | 0.00046 | -6.78 |




D

| | Normalized *Pi* | | Tajima's *D* | | Fay and Wu's *H* | |
|---|---|---|---|---|---|---|
| | Mean | Median | Mean | Median | Mean | Median |
| All TFs | 0.00078 | 0.00027 | -1.07 | -1.17 | -4.28 | -1.44 |
| NHR | 0.00140 | 0.00078 | -1.21 | -1.37 | -9.72 | -2.36 |
| HD | 0.00069 | 0.00024 | -0.93 | -1.00 | -2.48 | -0.04 |
| bHLH | 0.00039 | 0.00019 | -0.93 | -1.16 | -1.15 | 0.02 |
| C2H2 | 0.00079 | 0.00024 | -0.98 | -1.13 | -2.82 | -1.42 |
| Fate regulators | 0.00019 | 0.00011 | -0.98 | -1.06 | -0.97 | 0.01 |

**Supplementary fig. S8. Population genetics analysis for SCF complex factors and genes in TF subgroups.** (A) The cumulative distribution of normalized *Pi*, Tajima's *D*, and Fay and Wu's *H* values for *fbxa*, *fbxb*, and *fbxc* type F-box genes. (B) The normalized *Pi*, Tajima's *D*, and Fay and Wu's *H* values of Skp1-homologus genes and Cullin genes. *skr-11* is a pseudogene and not shown. (C) The cumulative distribution of normalized *Pi*, Tajima's *D*, and Fay and Wu's *H* value of genes in TF subgroups. See Table S1 for the classification of TF genes into subgroups. (D) The mean and median value of normalized *Pi*, Tajima's *D*, and Fay and Wu's *H* values for all TFs and the five TF subgroups. Double asterisks indicate $p < 0.05$ in Wilcoxon's rank-sum test.




B

| | Number of genes | | Normalized *Pi* | | Tajima's *D* | | Fay and Wu's *H* | |
|---|---|---|---|---|---|---|---|---|
| Genes in tissues | Total | No SNV | Mean | Median | Mean | Median | Mean | Median |
| All genes | 20222 | 1425 | 0.00099 | 0.00033 | -1.08 | -1.17 | -4.17 | -1.38 |
| Intestine | 583 | 26 | 0.00103 | 0.00026 | -1.03 | -1.14 | -3.36 | -1.13 |
| Muscle | 195 | 21 | 0.00056 | 0.00027 | -1.11 | -1.18 | -2.87 | -1.40 |
| Germline | 998 | 29 | 0.00058 | 0.00022 | -1.23 | -1.30 | -3.97 | -1.69 |
| Overlapped in intestine muscle, and germline | 445 | 74 | 0.00034 | 0.00013 | -0.95 | -1.01 | -1.56 | 0.01 |
| Embryonic neurons | 136 | 17 | 0.00056 | 0.00021 | -1.03 | -1.07 | -3.10 | -0.02 |
| Larval neurons | 367 | 47 | 0.00038 | 0.00015 | -0.95 | -1.06 | -1.48 | 0.01 |

**Supplementary fig. S9. Population genetics analysis for tissue-specific genes.** (A) The cumulative distribution of normalized *Pi*, Tajima's *D*, and Fay and Wu's *H* values for pan-neuronal genes in the embryo and larva, and genes expressed in intestine, muscle, and germline and genes expressed in all three tissues. (B) The number of the tissue-specific genes with no nonsynonymous SNVs, as well as the mean and median of normalized *Pi*, Tajima's *D*, and Fay and Wu's *H* values for these tissue-specific genes. Double asterisks indicate $p < 0.05$ in Wilcoxon's rank-sum test. Intestine-, muscle-, and germline-specific genes were compared individually with the genes expressed in all three tissues and statistical significance were found in all three comparisons.