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

Fuqiang Ma, Chun Yin Lau, and Chaogu Zheng*

School of Biological Sciences, The University of Hong Kong, Hong Kong SAR, China

*Correspondence: 4S13, Kadoorie Biological Science Building, Pokfulam Road, The University of Hong Kong, Hong Kong, China, Email: cgzheng@hku.hk (C.Z.)

Key words: *C. elegans*, polymorphisms, positive selection, F-box, GPCR

Significance statement: The small nematode *Caenorhabditis elegans* has emerged as an important organism in studying the genetic mechanisms of evolution. F-box and chemosensory GPCR are two of the largest gene families in *C. elegans*, but their intraspecific evolution within *C. elegans* was not studied before. In this work, using the nonsynonymous SNV data of 330 *C. elegans* wild isolates, we found that F-box and chemosensory GPCR genes showed larger polymorphisms and stronger positive selection than other genes. The large diversity is likely the result of rapid gene family expansion, high recombination rate, and gene flow. Analysis of subpopulation suggests that positive selection of these genes occurred most strongly in the non-Hawaiian population, which underwent a selective sweep possibly linked to human activities.

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 analyzed 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 more 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 great 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. Accumulation 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 variants are mostly mapped to the substrate-recognition domains of F-box proteins and the extracellular regions of csGPCRs, possibly resulting in advantages during adaptation by affecting protein degradation and the sensing of environmental cues, respectively.

Introduction

C. elegans genome contains over 350 F-box genes, compared to ~69 in human genome (Kipreos and Pagano 2000; Thomas 2006). 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. F-box proteins also contain substrate-binding domains, including FOG-2 homology (FTH) domain, F-box associated (FBA) domain, Leucine-rich repeats (LRR), WD40 repeats etc, which recruit the substrate protein to the E3 ubiquitin ligase (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 divided the wild strains into eight non-Hawaiian and three Hawaiian subpopulations. 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. Given their location at mostly the chromosomal arms, high recombination rate might have contributed to the large diversity of these genes. 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

To obtain the genomic data of *C. elegans* wild isolates, We used the hard-filtered VCF file (20180527 release) provided by the *Caenorhabditis elegans* natural diversity resource (CeNDR; https://www.elegansvariation.org/) (Cook, et al. 2017). We chose the hard-filtered VCF over the soft-filtered VCF to avoid including low-quality reads and variants with low coverage depth in our analysis. The hard-filtered VCF file contained in total 2,906,135 high-quality variants, including 2,493,687 single nucleotide variants (SNVs) and 412,448 small indels, which were annotated by SnpEff (v4.3t) using the Ensemble WBcel235.94 genome assembly. About 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) and included them in most of our analysis. Among the 2,906,135 variants, 594,265 occurred in the protein-coding region (CDS) and 2,311,870 occurred in non-coding regions. 266,004 SNVs caused nonsynonymous mutations, 271,718 SNVs caused synonymous mutations, and 51,701 SNVs may affect mRNA splicing.

Among all SNV sites, we found that 665,368 SNVs in 11,199 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 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 (v3.6.1) using Pearson correlation test (R function cor.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 (2.6%) out of the 660 alleles (330 strains) did not have valid genotype; for the median site, 3.9 (0.6%) of the 660 alleles did not have 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 annotated the VCF to extract nonsynonymous SNVs and 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 contained only 85,260 sites covering only 9,948 genes. Preserving more SNVs made *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. For most 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 (R.R. 1990; Librado and Rozas 2009) 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. 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.

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 was 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. 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, which were named based on the geographic locations of most strains that carry the ancestral lineage (supplementary fig. S3 and table S2, Supplementary Material online). Out of the 330 strains, 266 have one dominant ancestral lineage (one ancestral proportion > 0.5); the other 64 strains showed considerable mixing between at least three ancestral populations. "Hawaii_1" and "Hawaii_2" are the same as the previous "Hawaiian Volcano" and "Hawaiian Divergent" subpopulations, and "Hawaii_3" is a combination of "Hawaiian Low" and "Hawaiian Invaded" subpopulation defined by Crombie, et al. (2019) (supplementary table S3, Supplementary Material online).

We then 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 remaining 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 extended haplotype homozygosity (EHH) analysis. For the calculation of F_{ST} and the gene flow and migration analysis among the 11 subgroups, we removed the strains without any ancestral proportion over 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 nonsynonymous 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. 1,000 bootstrap replicates were performed to make the tree by SplitsTree. Edges with 100% bootstrap support are labelled with "100".

Prior to tree construction, *Caenorhabditis briggsae, Caenorhabditis remanei*, and *Caenorhabditis brenneri* were chosen as the outgroups. The coding sequences of *C. elegans* genes

and their orthologs in *C. briggsae, C. remanei*, and *C. brenneri* were downloaded from WormBase (WS275) and then aligned using MegaX (Kumar, et al. 2018) to identify variants. We used a set of algorithms, including OrthoMCL, OMA, TreeFam, ParaSite-Compara, Inparanoid_8, WormBase-Compara, and Hillier-set, to identify the orthologs of *C. elegans* genes in the other three species. We then checked each nonsynonymous SNV that existed in the *C. elegans* wild isolates (VCF file from CeNDR) for their presence in *C. briggsae*, *C. remanei*, and *C. brenneri* genomes. If the allele in the three species matched the *C. elegans* reference (N2) sequence, it was considered as a wild-type; if the allele matches the alternative sequence, the species carried that variant. If neither, we considered it missing for that SNV. In the case of one species having multiple orthologs of the *C. elegans* genes, we checked the SNV against all orthologs and if any of them had the alternative sequence, we considered the species to carry the variant. In total, we found 78,833, 74,274, and 55,234 *C. elegans* SNVs in *C. briggsae*, *C. remanei*, and *C. brenneri* respectively 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; Thomas 2006), 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 *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. 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. 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 (EHH) analysis

We used 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 the Integrated Haplotype Score (iHS) and the Cross-population Extended Haplotype Homozygosity (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.

Assessing the influence of varying population size and bottleneck effects

Because different subpopulations have different numbers of strains, the varying population size may create bias when calculating neutrality statistics. We assessed this potential bias by comparing the SNV data extracted through different sampling schemes (scattering and pooling schemes), which were previously established (Stadler, et al. 2009; Li, et al. 2010). For scattered sampling, we randomly selected 5 strains from each of the 11 subpopulations based on population structure; for pooled sampling, we randomly selected 15 Hawaiian strains (3 subpopulations) and 40 non-Hawaiian strains (8 subpopulations). We repeated the sampling 100 times and then calculated the average *Pi,* Tajima's *D*, and Fay and Wu's *H*. The small differences in their values between the scattered and pooled sampling schemes suggest that the bias introduced by varying population size is not significant.

To assess the influence of demographic history and bottleneck effects on the neutrality tests, we simulated SNV data using the software MSMS (Ewing and Hermisson 2010) under constant population size model and bottleneck model. Parameters for the simulation were set according to previous studies (Andersen, et al. 2012). The command for simulating the two models are: msms -N 20000 -ms 440 1000 -t 100 -r 150 -SAA 500 -Sp 0.5 -SAa 200 (constant) and msms -N 20000 -ms 440 1000 -t 100 -r 150 -SAA 500 -Sp 0.5 -SAa 200 -eN 0.015 0.01 -eN 0.020 1.0 (bottleneck). The simulated data were then plotted as site frequency spectra (SFS), which were compared to the empirical site frequency spectrum data for nonsynonymous SNVs in *C.elegans*.

We also used the software SweeD to estimate the selective sweep position. SweeD appeared to be robust against the confounding effect of bottleneck on selective sweep prediction (Nielsen, et al. 2005; Pavlidis, et al. 2013). We identified the selected sites with significant score using the likelihood threshold of 0.01. Genes that harbour these selected sites were then identified.

Copy number variation (CNV) analysis

The raw sequencing data of the 330 wild isolates were downloaded from NCBI (PRJNA549503). Sequencing reads were aligned to the reference genome of *C.elegans* using BWA-mem (v0.7.17). Structural variants were called using Manta (Chen, et al. 2016). The output VCF was merged by bcftools (v1.9). Structural variants with <= 5bp position difference and <=20% size

difference were merged together. Deletions and duplications were considered as copy number variation. Large deletions or duplications with more than 1 Mbp and chromosome-level variation were discarded. Derived CNV allele frequency were calculated using XZ1516 as the outgroup.

Protein domain structure and PROVEAN score

We used PfamScan tools (ftp://ftp.ebi.ac.uk/pub/databases/Pfam/Tools/) to identify the F-box domain and the potential substrate-recognition domains (e.g. FTH, FBA, HTH-48, WD40 repeats, LRR, etc) of F-box proteins. We used the TMHMM server (http://www.cbs.dtu.dk/services/TMHMM/) to predict the transmembrane domain (TM) and the intracellular and extracellular regions of csGPCR proteins. SNVs falling into these different domains were then filtered accordingly. 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

Large polymorphisms in F-box and chemosensory GPCR genes among *C. elegans* wild isolates

From the sequencing data of 330 distinct isotypes of *C. elegans* wild strains (VCF files of 20180527 release on CeNDR), we identified in total 2,493,687 SNVs, including 271,718 SNVs synonymous and 266,004 nonsynonymous SNVs (fig. 1A). By 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 (0.06 ~ 5.7 kb) had no nonsynonymous mutation or small indels in CDS; within the 1,143 genes, 302 with average gene length of 1.4 kb (0.07 ~ 14 kb) did not have any SNVs or small indels in any of the CDS, intron, and UTR regions (53 of the 302 genes has no introns). 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) and the fact that they tend to be enriched in genomic regions with low recombination rate (Rho = 1.4 on average; see below for the genomic distribution of Rho). The lack of variation might also be partly due to purifying selection, as gene ontology and phenotype analysis found that many of these genes function in protein-protein interactions and

309 are essential for cell division and germline development (supplementary fig. S1, Supplementary
310 Material online).

311 Next, we focused on the 2,493,687 SNVs and calculated nucleotide polymorphism (*Pi*)
312 for each protein-coding gene and found that *Pi* is significantly bigger in introns and UTRs compared
313 to synonymous and nonsynonymous mutations, suggesting that coding regions had much less
314 variation than non-coding regions (supplementary fig. S2A, Supplementary Material online). For
315 nonsynonymous SNVs, we found that the F-box and csGPCR genes have much larger diversity than
316 average genes (fig. 1B-E; supplementary table S1, Supplementary Material online). For example,
317 among the 235 genes whose *Pi* is bigger than 0.01, 46 of them are F-box genes, indicating an over
318 ten-fold enrichment. For the 1685 genes with *Pi* bigger than 0.0025, F-box genes and csGPCRs
319 showed 4.91- and 2.27-fold enrichment, respectively (fig. 1A). Overall, compared to other gene
320 families like the transcription factor (TF) genes (891) and the protein kinase genes (402), F-box
321 genes (336) and the csGPCRs (1301) on average have significantly bigger *Pi* (fig. 1D and E;
322 significance by a non-parametric Wilcoxon's test). Large genetic diversity among the wild isolates
323 hints that the F-box and csGPCR genes might contribute to the adaptation of *C. elegans* in the
324 natural environment.

325 The csGPCRs can be further divided into *Str*, *Sra*, *Srg*, and *Solo* superfamilies, among
326 which the Solo superfamily genes have the biggest *Pi* (fig. 1F). Within the *Solo* superfamily, *Srw*-
327 type csGPCRs appeared to have the largest polymorphism on average, although the mean of *Pi* is
328 not significantly bigger than *Srz* and *Srbc* subfamilies. Interestingly, *Srw* genes appeared to be more
329 ancestral than the other csGPCR families and likely originated from the large Rhodopsin GPCR
330 family before the split of the nematode lineage (Krishnan, et al. 2014). Unlike other csGPCRs, *Srw*
331 genes have clear phylogenetic relationship with conserved FMRFamide/peptide receptors found in
332 insects and vertebrates (Robertson and Thomas 2006); variations in *Srw* genes might lead to
333 selective advantages in peptide sensing.

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

335 We next conducted population structure analysis on the 330 wild isolates and found 3
336 Hawaiian and 8 non-Hawaiian subpopulations (supplementary fig. S3, Supplementary Material

online), which generally agrees with a recent study that used 276 strains and also found 11 distinct genetic groups (Crombie, et al. 2019). The three Hawaiian subpopulations contain strains mostly found in Hawaiian islands with a few exceptions. Among the non-Hawaiian subpopulations, some groups (e.g. "Europe_1" and "Europe_2") are heavily mixed with populations around the world, but others show correlation between geographical separation and genetic divergence. For example, "Europe_4" group represented strains mostly found in France, and "Europe_5" and "Europe_6" strains were mostly found on Iberian Peninsula and Portuguese islands. Similarly, strains extracted from North American shared the same "North_America" ancestral group.

Phylogenetic analysis using all nonsynymous SNVs and neighbor-joining methods (Huson and Bryant 2006) showed the evolutionary relationship among the *C. elegans* wild isolates. We rooted the tree using *C. briggsae*, *C. remanei* and *C. brenneri* as outgroups (fig. 2A; see Materials and Methods) to show that Hawaiian strains, especially "Hawaii_1" and "Hawaii_2" groups, are genetically closer to the sister species and contain more ancestral variations than the non-Hawaiian strains. Two strains in the "Hawaii_1" group, XZ1516 and ECA701, are highly divergent from other Hawaiian strains and appear to carry the most ancestral polymorphisms in among the *C. elegans* wild isolates.

"Hawaii_3" strains cluster more closely with non-Hawaiian strains (fig. 2A) and are more admixed with non-Hawaiian subpopulations (supplementary fig. S3, Supplementary Material online) compared to "Hawaii_"1 and "Hawaii_2" strains. Gene flow analysis identified migration events from "North Amercian" and "Europe_2" to "Hawaii_3", supporting the admixing of "Hawaii_3" with non-Hawaiian lineages (supplementary fig. S4A, Supplementary Material online).

Compared to the genomic average, divergence between the Hawaiian and non-Hawaiian strains are more profound in F-box and csGPCR genes, as shown in expanded neighbor-joining net and increased phylogenetic distance (fig. 2B and C). Within csGPCRs, *Srw* genes appear to show even greater divergence among the subpopulations (fig. 2D). Moreover, the phylogenetic trees constructed using nonsynomymous SNVs of csGPCR or F-box genes had different topologies from the tree of all genes (fig. 2D). 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.

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 Materials and Method) and calculated polymorphism for the two populations using nonsynomymous SNVs. Hawaiian population showed over two-fold larger *Pi* than non-Hawaiian population across all genes (fig. 3A), which is consistent with the hypothesis that recent selective sweep reduced variation in non-Hawaiian population, while Hawaiian strains kept part of the ancenstral diversity (Cook, et al. 2017; Crombie, et al. 2019). "Hawaii_3" has lower diversity than the other two Hawaiian subpopulations, likely because "Hawaii_3" strains are genetically more similar to the non-Hawaiian strains. Interestingly, the diversity of F-box and csGPCR genes is bigger than the TF, protein kinase genes, or an average gene in both non-Hawaiian and Hawaiian populations (fig. 3A). 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. 3B). As expected, F-box and csGPCR genes have a lot more exclusively Hawaiian sites than the TF or protein kinase genes. However, they do not 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. 3B). 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. 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. 3C). ~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), 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. 3D and 3F). 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. 3E and 3F), which may be explained by their large diversities even within the same Hawaiian group (fig. 3A). Thus, the big variation of F-box and *Srw* genes do not seem to follow the population structure among the three Hawaiian subpopulations and they are not likely fixed within the Hawaiian groups.

Gene flow also helped shape the diversity of the F-box and csGPCR genes. F-box genes have extensive gene flow between Hawaiian and non-Hawaiian populations in both directions (supplementary fig. S4B, Supplementary Material online), which is consistent with the great number of shared segregating sites in F-box genes between the two populations (fig. 3B). On the other hand, csGPCR genes had only gene flow within the non-Hawaiian subpopulations. Interestingly, when constructing the maximum-likehood population tree for gene flow analysis, we found that the tree structure changed after removing the variants 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" (supplementary fig. S4C, Supplementary Material online). 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. 4A). 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. 4B). Thus, the location of almost all F-box genes in the chromosomal arms may contribute to their high genetic diversity. csGPCR genes are also relatively concentrated on the arms, especially on chromosome II and V, although to a lesser extent than F-box genes. For example, over 80% of the *Srw* genes are located on the arms of V. This chromosomal clustering is likely the result of rapid gene duplication (Robertson and Thomas 2006). In contrast, protein kinase and TF genes are more evenly spread out across chromosomes (fig. 4A).

We estimated the recombination rate for Chromosome II, III, and V using the FastEPRR software (Gao, et al. 2016) and all nonsynonymous SNVs. As expected, chromosomal arm regions have higher estimated recombination rate (ρ or *Rho*) than the center (fig. 4C). Interestingly, the left arm of chromosome III where many F-box genes are located have significantly higher recombination rate than the right arm, and this high *Rho* value of F-box genes appear to correlate with the large polymorphism (fig. 4D). Similarly, the arms of Chromosome V, which harbour many csGPCR genes, have high recombination rate, which are correlated with large *Pi* (fig. 4D). Therefore, because of the clustering of F-box and csGPCR genes on the chromosomal arms, recombination likely contributed to their large diversity.

The analysis of copy number variants (CNVs) supports the idea that rapid expansion of F-box and csGPCR gene families led to large genetic diversity. By analysing the structural variants, we

found 8740 CNVs in 5586 genes. 185 (1.99 fold enrichment) F-box genes and 552 (1.54 fold enrichment) csGPCRs carried CNVs (supplementary fig. S6A, Supplementary Material online). Moreover, the average number of CNVs per gene is also higher for F-box and csGPCR genes compared to genomic average. Thus, large genetic polymorphisms for these genes were reflected in both the abundance of SNVs and CNVs.

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

Previous studies hypothesized that positive selection of alleles that confer fitness advantages under human influence reduced genetic variations in *C. elegans* (Andersen, et al. 2012), but what genes are under selection is not clear. Using the nonsynonymous SNVs, we performed neutrality tests and calculated Tajima's *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 (fig. 2). *H* values calculated using the two strains as the outgroup were similar (supplementary fig. S7, 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 the nonsynonymous SNVs for most (>85%) genes and Fay and Wu's *H* were negative for ~50% genes (supplementary table S1, Supplementary Material online). This finding is consistent with the chromosome-wide sweep that occurred across the genome (Andersen, et al. 2012). Interestingly, F-box and csGPCR genes are overrepresented among the genes with significantly negative *D* and *H* values (fig. 5A). 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. S8, 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. 5B and C), 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. 5D and E). 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. S9A, 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. S9B, Supplementary Material online). Thus, components of the ubiquitination-proteasome system (UPS) may co-evolve among the *C. elegans* wild isolates; genetic variations in UPS genes may alter the homeostasis of target proteins, leading to certain advantages during selection.

Another line of evidence for positive selection is the excess of nonsynonymous SNVs compared to synonymous SNVs, which is particularly obvious for F-box genes (supplementary fig. S10, Supplementary Material online). Polymorphism for synonymous SNVs is slightly bigger than nonsynonymous SNVs for genomic average, and the difference is more obvious in TF and protein kinase genes, which may be under purifying selection. F-box genes, however, have bigger *Pi* and more negative *D* and H values for nonsynonymous SNVs compared to synonymous SNVs, supporting that F-box genes are under positive selection. Interestingly, csGPCRs did not show this pattern and appeared to have a lot of synonymous SNVs, which have also very negative *H* values. The abundance of synonymous SNVs in csGPCRs may result from high recombination rate at the chromosomal arms; some synonymous SNVs might also be positively selected due to effects on codon usage and gene expression levels as previously seen in mammals (Resch, et al. 2007).

Positive selection of F-box and csGPCR genes in non-Hawaiian population

The above analysis detected signs of strong positive selection in F-box and csGPCR genes among all wild strains, we next found that Fay and Wu's *H* value is more negative in the non-

505 Hawaiian strains than the Hawaiian strains across all genes (fig. 6A). "Hawaii_3" group appeared
506 to have lower *H* values than "Hawaii_1" and "Hawaii_2" groups probably due to the admixing with
507 the non-Hawaiian strains. These observations are consistent with the selective sweep in non-
508 Hawaiian populations. Genes in the F-box and csGPCR (especially *Srw*) genes showed much more
509 negative *H* than the genomic average not only in non-Hawaiian strains but also in Hawaiian strains,
510 suggesting that they may also be under positive selection within the Hawaiian populations when
511 considering XZ1516 as the most ancestral strain.

512 Negative *H* values reflects the excess of high-frequency derived alleles. Indeed, the
513 number of high-frequency (>50%) SNVs distinct from the ancestral allele in XZ1516 is much higher
514 in csGPCR and F-box genes than in an average gene for both Hawaiian and non-Hawaiian
515 populations (fig. 6B). The large number of both segregating sites and high-frequency derived alleles
516 in csGPCR and F-box genes suggest that they evolve more rapidly than other parts of the genome
517 and may contribute to the adaptation to a changing environment in *C. elegans*. Analysis of copy
518 number variants (CNVs) is consistent with the above results. The allele frequency of derived CNVs
519 is much larger for F-box and csGPCR genes than the genomic average in both the entire population
520 and the non-Hawaiian population of wild isolates, with XZ1516 as the outgroup (supplementary
521 fig. S6B and C, Supplementary Material online).

522 Since the Hawaiian populations contained relatively more ancestral polymorphism than
523 the non-Hawaiian population, we next asked whether F-box and csGPCR genes accumulated high-
524 frequency derived alleles in non-Hawaiian population when considering Hawaiian strains as the
525 ancestral population. We found that *H* values calculated using a representative "Hawaii_1"
526 (ECA396) or "Hawaii_2" (ECA742) strain as the outgroup were significantly more negative in F-box
527 and csGPCR genes than genomic average, indicating positive selection within the non-Hawaiian
528 population, relative to the Hawaiian populations (fig. 6C and D). These *H* values were generally less
529 negative than the *H* values calculated using XZ1516 or ECA701 as the outgroup, confirming that
530 "Hawaii_1" and "Hawaii_2" strains are less distant to the non-Hawaiian strains than the two
531 outliers XZ1516 and ECA701.

532 **Different selection pressure on different domains of F-box and csGPCR proteins**

F-box proteins all have two distinct funcitonal domains, a N-terminal F-box domain that mediates the assembly of SCF complex and a C-terminal substrate recognition domain that binds the substrate proteins and target them for ubiquitination. Using Pfam scan, we identified the F-box domain and putative substrate-binding domain (e.g. FTH, FBA, etc) in all F-box proteins and extracted the nonsynonymous SNVs mapped to these domains. Interestingly, *Pi* is much bigger and *H* much more negative for the SNVs mapped to the substrate-binding domain compared to those mapped to the F-box domain (fig. 7A). The enrichment of variants and stronger positive selection in the substrate-recognition domains supports the hypothesis that variations in the F-box genes may result in selective advantages by altering the ubiquitination and degradation of certain cellular proteins.

As an example, F-box gene *fbxb-49* (*H* = -53.03) 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. 7C). Most of those sites are also high frequency derived sites when using a “Hawaii_1” (ECA396) or “Hawaii_2” (ECA742) strain as the outgroup. So, selective sweep may have fixed those sites in the non-Hawaiian population. Very few variants are located in the region encoding F-box domain, while many more sites occurred in domains that are responsible for recognizing the substrate protein. Four of the sites have PROVEAN (Protein Variation Effect Analyzer) score (Choi and Chan 2015) below -2.5, suggesting potentially significant functional impacts.

Similarly, we mapped nonsynonymous SNVs onto the domain structure of csGPCRs and found that SNVs affecting the extracellular domains or intracellular tails of csGPCRs have larger *Pi* and more negative *H* than SNVs mapped to the transmembrane (TM) domains (fig. 7B). Conservation of amino acid sequences in the TM domain is expected, as the membrane protein topology may be maintained by purifying selection. Variation in the extracellular domains, which are under stronger positive selection, could result in changes in the ability to sense envrionmental signals, which may confer fitness advantages. As an example, *srw-68* (*H* = -82.33) 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 (fig. 7D). All of those

sites are near fixation among the non-Hawaiian strains and mostly mapped to the extracellular regions. Four sites have PROVEAN scores below -2.5.

Extended Haplotype Homozygosity analysis identified selection footprints in F-box and csGPCR 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. 8A 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. 8D), indicating that these genes may be under selective pressure in the recent sweep among non-Hawaiian strains. Nevertheless, csGPCR and F-box genes may also be selected within the Hawaiian strains because of their enrichment in the regions with high |iHS| in the Hawaiian population (fig. 8D).

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. 8E). 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. 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. 8C). 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. 8D). 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 examples of highly selected genes, F-box gene *fbxa-85* carries 58 SNVs with significantly positive XP-EHH score (XP-EHH > 2; $p < 0.05$); 13 and 29 are high-frequency derived sites among non-Hawaiian strains using a "Hawaii_1" and "Hawaii_2" strain as the outgroup, respectively. Most of the sites occurred in the FTH domain involved in substrate binding and none in the F-box domain (fig. 8F). Similarly, sGPCR *srw-56* contains 67 SNVs with high XP-EHH; 36 and 24 of those SNVs are high frequency derived sites in non-Hawaiian population with a "Hawaii_1" and "Hawaii_2" strain as the outgroup, respectively. Most of them occurred in the extracellular domains of SRW-56 (fig. 8G).

Selection patterns in F-box and csGPCR genes are not likely affected by varying population size and demographic history

Varying population size and demographic history are known confounding factors for predicting selective sweep (Wakeley and Aliacar 2001; Przeworski 2002; Nielsen, et al. 2005). We next addressed whether these two factors confounded our neutrality test results. To assess whether the varying number of strains in the 11 subpopulation among the wild isolates had effects on the neutrality test statistics, we selected strains and SNVs using two different sampling schemes (scattering and pooling schemes) according to previous studies (Stadler, et al. 2009; Li, et al. 2010) (see Materials and Methods). Polymorphism, Tajima's *D*, and Fay and Wu's *H* calculated using the samples obtained with the two sampling methods are very similar (supplementary fig. S11, Supplementary Material online), indicating that the varying population sizes among the subpopulation do not significantly confound our results.

Previous population history analysis of *C. elegans* found that wild isolates in non-

Hawaiian population may have suffered a strong decline in population size about 10,000 generations ago (Thomas, et al. 2015). To assess the confounding effect of the potential bottleneck on selection detection, we simulated SNV data under neutrally constant population size model or bottleneck model and plotted site frequency spectra (SFS). Bottleneck leads to the enrichment of low and high frequency alleles in simulated data as expected (supplementary fig. S12, Supplementary Material online). However, SFS pattern of the empirical SNV data of non-Hawaiian populations are more similar to the constant population size model, suggesting that the potential bottleneck effect may not significantly change the site frequency in the *C. elegans* wild isolates we analyzed.

We next predicted selective sweep sites based on the site frequency spectrum using the software SweeD, which analyzes composite likelihood and is robust against recombination and demographic assumption (Nielsen, et al. 2005). We found that selected sites at the significance threshold of 1% are mostly located in the arms of chromosomes (supplementary fig. S13, Supplementary Material online), where F-box and csGPCRs are enriched, which is consistent with the results of neutrality tests and EHH analysis. Among the 564 significant sites located in 233 genes (mean α score, an indicator of selection coefficient, is 31), 31 sites are mapped to 10 F-box genes (2.5 fold enrichment) with an average α score at 54. csGPCR genes carry 28 significant sites with an average α score at 63. Thus, even considering demographic history, F-box and csGPCR genes still show strong selection during the sweep.

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 over 50 species in the Caenorhabditis genus and 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

to identify 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 large intraspecific variations in the F-box genes and signs of strong positive selection in non-Hawaiian population, which may imply their roles in adaptation. Interestingly, variants in the substrate-binding domain showed larger polymorphism and stronger selection than the variants in 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 expect certain

673 alleles of F-box genes to confer other fitness advantages, given the importance of ubiquitination-
674 proteasome system in many biological processes.

675 **Intraspecific adaptive evolution of csGPCRs**

676 The csGPCR family is the largest gene family in *C. elegans* and contains over 1,300 genes.
677 Through the studies of specific phenotypes, a few csGPCRs were previously connected to
678 adaptation. For example, the deletion of two csGPCR genes, *srg-36* and *srg-37*, which resulted in
679 insensitivity to the dauer pheromone ascaroside and defects in entering dauer diapause, were
680 acquired independently by two domesticated *C. elegans* strains grown in high density (McGrath,
681 et al. 2011). Similar loss-of-function deletions in *srg-36* and *srg-37* were also found in natural
682 isolates across the globe, suggesting that niche-associated variation in pheromone receptors may
683 contribute to the boom-and-bust population dynamics (Lee, et al. 2019). In addition, a frameshift-
684 causing deletion in another csGPCR, *str-217*, in Hawaiian strain CB4856, led to resistance to the
685 insect repellents N,N-Diethyl-meta-toluamide (DEET) (Dennis, et al. 2018) and similar deletions
686 were found in nine other wild isotypes (our unpublished results), suggesting that *C. elegans* may
687 have evolved to acquire resistance to harmful environmental chemicals by inactivating csGPCRs.
688 The above examples showcased how the intraspecific evolution of individual csGPCR genes can
689 have significant functional consequences and contribute to adaptation. Our study, in a more
690 systematic way, indicates that csGPCRs are highly diverse and are under strong positive selection
691 in the *C. elegans* wild population.

692 Among the four csGPCR superfamilies (*Str*, *Sra*, *Srg*, and *Solo*), our analysis using
693 nonsynonymous SNVs found that *Str* genes had larger polymorphism and stronger positive
694 selection than *Sra* and *Srg* genes (fig. 1F and 5D), which is consistent with previous observation on
695 the intraspecific variations of *Str* genes (Stewart, et al. 2005). In fact, these variations created ~200
696 pseudogenes *Str* genes in *C. elegans* reference genome (the N2 strain) through often times only
697 one apparent defect. Compared to *Str* genes, we found that *Solo* superfamily csGPCRs, especially
698 *Srw* genes have even larger diversity and stronger positive selection. Given that *Srw* genes are the
699 only csGPCRs that have clear homology with vertebrate GPCRs and likely code for
700 FMRFamide/peptide receptors (Robertson and Thomas 2006), large variations in these 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.

Rapid expansion of the F-box and csGPCR gene families and high rate of gene gain and loss also contributed to their large diversity and facilitated positive selection and adaptation. Indeed, our analysis of copy number variants found more frequent gene duplication and deletion in F-box and csGPCR genes than genomic average, supporting fast intraspecific evolution of these genes. Previous studies found that the *C. elegans* genome shows a higher duplication rate than *Drosophila* and *yeast* genomes (Lipinski, et al. 2011). This pattern is mostly likely driven by the duplication of F-box and csGPCR genes. Functional diversification of the duplicated genes could lead to novel functional characteristics. Although the function of most F-box and csGPCR genes in *C. elegans* are unknown, their expression pattern, to certain extent, reflects their potential functions. For example, among the 39 positively selected ($H < -20$) csGPCRs whose expression were

729 studied before (Vidal, et al. 2018), we found that these csGPCRs show distinct expression patterns
730 in a diverse range of tissues (supplementary fig. S14, Supplementary Material online). Although
731 expression is heavily enriched in sensory neurons, most csGPCRs are expressed in unique sets of
732 cells and identical expression patterns for two csGPCRs are rare. We suspect the diversification in
733 expression regulation is correlated with diversification in functions. Thus, our data supports a
734 model that duplications of F-box and csGPCR genes and accumulation of nonsynonymous SNVs
735 lead to functional diversities in protein degradation and chemosensation pathways, which allowed
736 positive selection to act upon during adaptation.
737
738 **Acknowledgments**
739 We thank the *Caenorhabditis elegans* Natural Diversity Resource (CeNDR) at Northwestern
740 University for sharing genomic data of the *C. elegans* wild isolates through their website. This study
741 is supported by funds from the Research Grant Council of Hong Kong [ECS 27104219], the Food
742 and Health Bureau of Hong Kong [HMRF 07183186], and seed funds from the University of Hong
743 Kong [201812159005 and 201910159087] to C.Z. Computational work were performed using
744 research computing facilities offered by Information Technology Services at the University of Hong
745 Kong.
746
747 **References**
748 Correction: A Randomised, Double-Blind, Controlled Efficacy Trial of the LiESP/QA-21 Vaccine in Naive
749 Dogs Exposed to Two Leishmania infantum Transmission Seasons. PLoS Negl Trop Dis 8:e3408.
750 Andersen EC, Gerke JP, Shapiro JA, Crissman JR, Ghosh R, Bloom JS, Felix MA, Kruglyak L. 2012.
751 Chromosome-scale selective sweeps shape Caenorhabditis elegans genomic diversity. Nat Genet
752 44:285-290.
753 Angeles-Albores D, Lee RYN, Chan J, Sternberg PW. 2018. Two new functions in the WormBase
754 Enrichment Suite. Micropublication: biology. Dataset.
755 Bakowski MA, Desjardins CA, Smelkinson MG, Dunbar TL, Lopez-Moyado IF, Rifkin SA, Cuomo CA,
756 Troemel ER. 2014. Ubiquitin-mediated response to microsporidia and virus infection in C. elegans. PLoS
757 Pathog 10:e1004200.
758 Begun DJ, Aquadro CF. 1992. Levels of naturally occurring DNA polymorphism correlate with
759 recombination rates in D. melanogaster. Nature 356:519-520.
760 Bounoutas A, Zheng Q, Nonet ML, Chalfie M. 2009. mec-15 encodes an F-box protein required for touch
761 receptor neuron mechanosensation, synapse formation and development. Genetics 183:607-617,

601SI-604SI.

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. 2007. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633-2635.

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 81:1084-1097.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. 2013. Stacks: an analysis tool set for population genomics. Molecular Ecology 22:3124-3140.

Chen K, Franz CJ, Jiang H, Jiang Y, Wang D. 2017. An evolutionarily conserved transcriptional response to viral infection in Caenorhabditis nematodes. BMC Genomics 18:303.

Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Kallberg M, Cox AJ, Kruglyak S, Saunders CT. 2016. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics 32:1220-1222.

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

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

Crombie TA, Zdraljevic S, Cook DE, Tanny RE, Brady SC, Wang Y, Evans KS, Hahnel S, Lee D, Rodriguez BC, et al. 2019. Deep sampling of Hawaiian Caenorhabditis elegans reveals high genetic diversity and admixture with global populations. Elife 8.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. Bioinformatics 27:2156-2158.

Dennis EJ, Dobosiewicz M, Jin X, Duvall LB, Hartman PS, Bargmann CI, Vosshall LB. 2018. A natural variant and engineered mutation in a GPCR promote DEET resistance in C. elegans. Nature 562:119-123.

Ewing G, Hermisson J. 2010. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics 26:2064-2065.

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

Fielenbach N, Guardavaccaro D, Neubert K, Chan T, Li D, Feng Q, Hutter H, Pagano M, Antebi A. 2007. DRE-1: an evolutionarily conserved F box protein that regulates C. elegans developmental age. Dev Cell 12:443-455.

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

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) 6:1563-1571.

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

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 104:5947-5952.

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

Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from DNA sequence data. Genetics 132:583-589.

804 Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Molecular
805 Biology and Evolution 23:254-267.
806 Hutter S, Vilella AJ, Rozas J. 2006. Genome-wide DNA polymorphism analyses using VariScan. Bmc
807 Bioinformatics 7:409.
808 Jager S, Schwartz HT, Horvitz HR, Conradt B. 2004. The Caenorhabditis elegans F-box protein SEL-10
809 promotes female development and may target FEM-1 and FEM-3 for degradation by the proteasome.
810 Proc Natl Acad Sci U S A 101:12549-12554.
811 Kim K, Sato K, Shibuya M, Zeiger DM, Butcher RA, Ragains JR, Clardy J, Touhara K, Sengupta P. 2009. Two
812 chemoreceptors mediate developmental effects of dauer pheromone in C. elegans. Science 326:994-
813 998.
814 Kipreos ET, Pagano M. 2000. The F-box protein family. Genome Biol 1:REVIEWS3002.
815 Krishnan A, Almen MS, Fredriksson R, Schioth HB. 2014. Insights into the origin of nematode
816 chemosensory GPCRs: putative orthologs of the Srw family are found across several phyla of
817 protostomes. PLoS One 9:e93048.
818 Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis
819 across Computing Platforms. Mol Biol Evol 35:1547-1549.
820 Lee D, Zdraljevic S, Cook DE, Frezal L, Hsu JC, Sterken MG, Riksen JAG, Wang J, Kammenga JE, Braendle
821 C, et al. 2019. Selection and gene flow shape niche-associated variation in pheromone response. Nat
822 Ecol Evol 3:1455-1463.
823 Li Y, Stocks M, Hemmila S, Kallman T, Zhu H, Zhou Y, Chen J, Liu J, Lascoux M. 2010. Demographic histories
824 of four spruce (Picea) species of the Qinghai-Tibetan Plateau and neighboring areas inferred from
825 multiple nuclear loci. Mol Biol Evol 27:1001-1014.
826 Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data.
827 Bioinformatics 25:1451-1452.
828 Lipinski KJ, Farslow JC, Fitzpatrick KA, Lynch M, Katju V, Bergthorsson U. 2011. High spontaneous rate of
829 gene duplication in Caenorhabditis elegans. Curr Biol 21:306-310.
830 Manning G. 2005. Genomic overview of protein kinases. WormBook:1-19.
831 McGaugh SE, Heil CS, Manzano-Winkler B, Loewe L, Goldstein S, Himmel TL, Noor MA. 2012.
832 Recombination modulates how selection affects linked sites in Drosophila. PLoS Biol 10:e1001422.
833 McGrath PT, Xu Y, Ailion M, Garrison JL, Butcher RA, Bargmann CI. 2011. Parallel evolution of
834 domesticated Caenorhabditis species targets pheromone receptor genes. Nature 477:321-325.
835 Mohd-Assaad N, McDonald BA, Croll D. 2018. Genome-Wide Detection of Genes Under Positive
836 Selection in Worldwide Populations of the Barley Scald Pathogen. Genome Biol Evol 10:1315-1332.
837 Nayak S, Goree J, Schedl T. 2005. fog-2 and the evolution of self-fertile hermaphroditism in
838 Caenorhabditis. PLoS Biol 3:e6.
839 NEI M. 1987. Molecular Evolutionary Genetics. Columbia Univ. Press, New York.
840 Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. 2005. Genomic scans for selective
841 sweeps using SNP data. Genome Research 15:1566-1575.
842 Park D, O'Doherty I, Somvanshi RK, Bethke A, Schroeder FC, Kumar U, Riddle DL. 2012. Interaction of
843 structure-specific and promiscuous G-protein-coupled receptors mediates small-molecule signaling in
844 Caenorhabditis elegans. Proc Natl Acad Sci U S A 109:9917-9922.
845 Pauli F, Liu Y, Kim YA, Chen PJ, Kim SK. 2006. Chromosomal clustering and GATA transcriptional regulation

846 of intestine-expressed genes in C. elegans. Development 133:287-295.
847 Pavlidis P, Zivkovic D, Stamatakis A, Alachiotis N. 2013. SweeD: likelihood-based detection of selective
848 sweeps in thousands of genomes. Molecular Biology and Evolution 30:2224-2234.
849 Pfeifer B, Wittelsburger U, Ramos-Onsins SE, Lercher MJ. 2014. PopGenome: an efficient Swiss army
850 knife for population genomic analyses in R. Molecular Biology and Evolution 31:1929-1936.
851 Pickrell JK, Pritchard JK. 2012. Inference of population splits and mixtures from genome-wide allele
852 frequency data. PLoS Genet 8:e1002967.
853 Przeworski M. 2002. The signature of positive selection at randomly chosen loci. Genetics 160:1179-
854 1189.
855 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly
856 MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses.
857 Am J Hum Genet 81:559-575.
858 R.R. H. 1990. Gene genealogies and the coalescent process. Oxford surveys in evolutionary biology.
859 Reece-Hoyes JS, Deplancke B, Shingles J, Grove CA, Hope IA, Walhout AJ. 2005. A compendium of
860 Caenorhabditis elegans regulatory transcription factors: a resource for mapping transcription regulatory
861 networks. Genome Biol 6:R110.
862 Resch AM, Carmel L, Marino-Ramirez L, Ogurtsov AY, Shabalina SA, Rogozin IB, Koonin EV. 2007.
863 Widespread positive selection in synonymous sites of mammalian genes. Mol Biol Evol 24:1821-1831.
864 Robertson HM. 2000. The large srh family of chemoreceptor genes in Caenorhabditis nematodes reveals
865 processes of genome evolution involving large duplications and deletions and intron gains and losses.
866 Genome Res 10:192-203.
867 Robertson HM, Thomas JH. 2006. The putative chemoreceptor families of C. elegans. WormBook:1-12.
868 Rozas J, Ferrer-Mata A, Sanchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, Sanchez-Gracia
869 A. 2017. DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. Molecular Biology and
870 Evolution 34:3299-3302.
871 Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ,
872 McDonald GJ, et al. 2002. Detecting recent positive selection in the human genome from haplotype
873 structure. Nature 419:832-837.
874 Seidel HS, Rockman MV, Kruglyak L. 2008. Widespread genetic incompatibility in C. elegans maintained
875 by balancing selection. Science 319:589-594.
876 Sengupta P, Chou JH, Bargmann CI. 1996. odr-10 encodes a seven transmembrane domain olfactory
877 receptor required for responses to the odorant diacetyl. Cell 84:899-909.
878 Stadler T, Haubold B, Merino C, Stephan W, Pfaffelhuber P. 2009. The impact of sampling schemes on
879 the site frequency spectrum in nonequilibrium subdivided populations. Genetics 182:205-216.
880 Stevens L, Felix MA, Beltran T, Braendle C, Caurcel C, Fausett S, Fitch D, Frezal L, Gosse C, Kaur T, et al.
881 2019. Comparative genomics of 10 new Caenorhabditis species. Evol Lett 3:217-236.
882 Stewart MK, Clark NL, Merrihew G, Galloway EM, Thomas JH. 2005. High genetic diversity in the
883 chemoreceptor superfamily of Caenorhabditis elegans. Genetics 169:1985-1996.
884 Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.
885 Genetics 123:585-595.
886 Thomas CG, Wang W, Jovelin R, Ghosh R, Lomasko T, Trinh Q, Kruglyak L, Stein LD, Cutter AD. 2015. Full-
887 genome evolutionary histories of selfing, splitting, and selection in Caenorhabditis. Genome Research

25:667-678.

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

Thomas JH, Kelley JL, Robertson HM, Ly K, Swanson WJ. 2005. Adaptive evolution in the SRZ chemoreceptor families of Caenorhabditis elegans and Caenorhabditis briggsae. Proc Natl Acad Sci U S A 102:4476-4481.

Thomas JH, Robertson HM. 2008. The Caenorhabditis chemoreceptor gene families. BMC Biol 6:42.

Vidal B, Aghayeva U, Sun H, Wang C, Glenwinkel L, Bayer EA, Hobert O. 2018. An atlas of Caenorhabditis elegans chemoreceptor expression. PLoS Biol 16:e2004218.

Von Stetina SE, Watson JD, Fox RM, Olszewski KL, Spencer WC, Roy PJ, Miller DM, 3rd. 2007. Cell-specific microarray profiling experiments reveal a comprehensive picture of gene expression in the C. elegans nervous system. Genome Biol 8:R135.

Wakeley J, Aliacar N. 2001. Gene genealogies in a metapopulation. Genetics 159:893-905.

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 106:835-840.

Figure legends

Figure 1. Large genetic polymorphism of csGPCR and F-box genes. (A) Flowchart of the analysis of SNVs among the wild isolates of *C. elegans*. (B) Genes with large *Pi* for nonsynonymous SNVs tend to be enriched in csGPCR and F-box gene families. *Pi* values of individual genes can be found in Table S1. (C) A histogram for the distribution of *Pi* across all genes. (D) The cumulative distribution of the *Pi* values for all genes, csGPCR, F-box, transcription factor (TF) and Protein kinases genes. (E) The mean and median of *Pi* for different gene families. (F) The mean and median of *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 Figures.

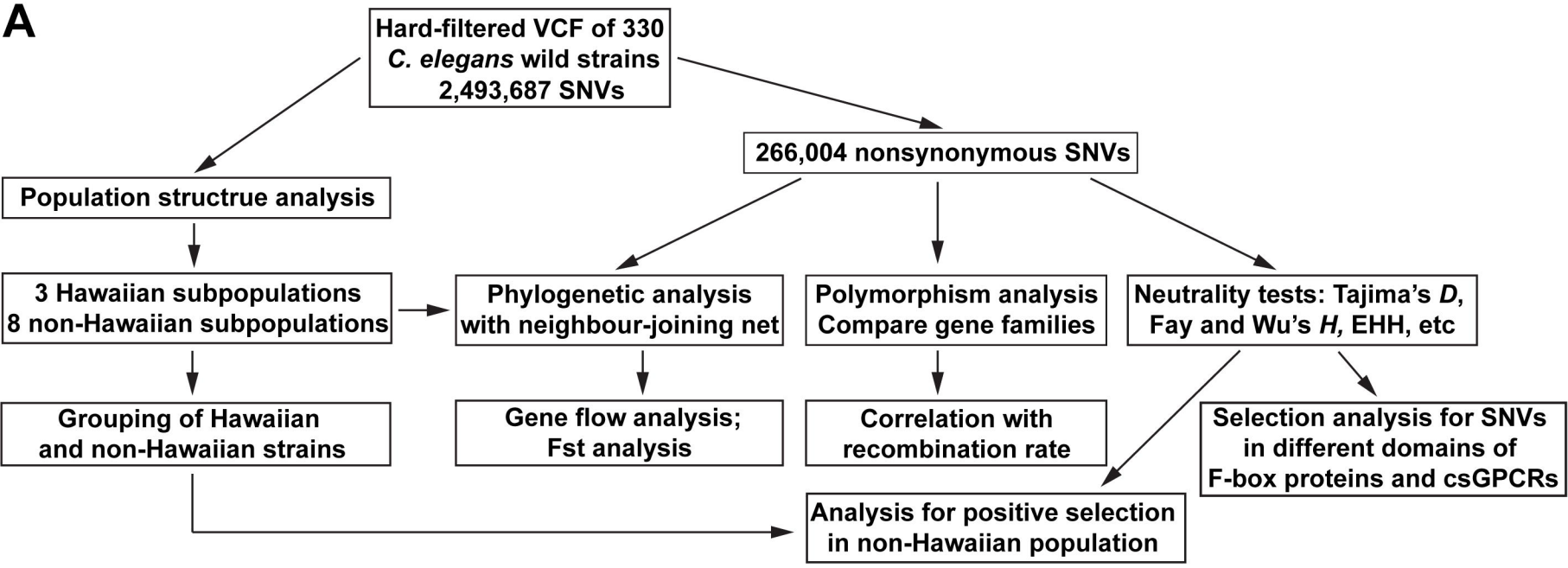
Figure 2. Phylogenetic relationship of the *C. elegans* wild isolates. Neighbour-joining nets plotted using the nonsynonymous SNVs of all genes (A), F-box genes (B), csGPCRs (C), or *Srw* genes (D). *C. brenneri*, *C. remanei*, and *C. briggsae* were used as outgroups for tree construction. Three representative non-Hawaiian strains (in black) with high ancestral population fraction were chosen from each of the eight non-Hawaiian groups. Edges are

921 labelled with "100", if 100% bootstrap support was attained in 1,000 bootstrap replicates. To
922 fit the trees into one figure, some branches connecting the three outgroups and the root are
923 manually shortened (dashed lines).
924
925 **Figure 3. csGPCR and F-box genes contribute to the large divergence of Hawaiian strains and**
926 **the differentiation among non-Hawaiian subpopulations.** (A) The mean of *Pi* for
927 nonsynonymous SNVs in all genes, csGPCRs, *Srw* genes, F-box genes, TF, and Protein kinase for
928 non-Hawaiian and Hawaiian populations, as well as the three Hawaiian subpopulations. (B)
929 The average number of segregating sites that belong to only Hawaiian or non-Hawaiian strains
930 and the sites that are shared by Hawaiian and non-Hawaiian strains for the six gene families.
931 The number is also normalized to the CDS length of individual genes. The number of non-
932 singleton segregating sites are in the parentheses. (C-E) The cumulative distribution of
933 Hudson's F_{ST} values for different gene families between the non-Hawaiian and Hawaiian
934 populations (C), among the eight non-Hawaiian subpopulations (D), and among the three
935 Hawaiian subpopulations (E). (F) The average F_{ST} value of different gene families among non-
936 Hawaiian and among Hawaiian subpopulations.
937
938 **Figure 4. High recombination rate may contribute to the large diversity of csGPCR and F-box**
939 **genes.** (A) Genomic location of F-box, csGPCR, Protein kinase, and TF genes plotted using
940 TBtools. (B) The mean of Pi values for the nonsynonymous SNVs in all genes, csGPCR, and F-
941 box genes in the arm or center of chromosome (Chr) II, III, IV and V. Chromosomes were
942 divided into three regions according to genomic coordinates: the left arm (one-fourth of the
943 chromosome from the start), the center (the central half), and the right arm (one-fourth of
944 the chromosome close to the end). The values in parentheses indicated gene numbers. (C)
945 Recombination rates (Rho) across Chr II, III and V in 50-kb windows. (D) The Pearson
946 correlation between recombination rate and Pi for F-box genes on Chr III and csGPCR genes
947 on Chr V.
948

949 **Figure 5. Positive selection on F-box and csGPCR gene.** (A) Enrichment of csGPCR and F-box
950 genes among the genes with Tajima's *D* < -2 and Fay and Wu's *H* < -20, respectively. Overlap
951 set include genes that fits both criteria. (B) The mean and median of Tajima's *D* and Fay and
952 Wu's *H* values of all genes, csGPCRs, F-box, TF, and Protein kinase. (C) The cumulative
953 distribution of different gene families. (D) The mean and median of Fay and Wu's *H* values of
954 genes in csGPCR superfamilies and *Solo* gene families. The number of genes are in parentheses.
955 (E) The cumulative distribution of genes in csGPCR subfamilies and *Solo* families. *H* values were
956 calculated using XZ1516 as the outgroup. The statistical significance was determined by
957 Wilcoxon rank-sum test.
958
959 **Figure 6. Accumulation of high-frequency derived alleles in F-box and csGPCR genes in non-**
960 **Hawaiian population.** (A) The average Fay and Wu's *H* values of all genes, csGPCRs, *Srw* genes,
961 F-box genes, TFs, and Protein kinase for the non-Hawaiian and Hawaiian populations, as well
962 as the three Hawaiian subpopulations. (B) The average number of high frequency (>50%)
963 derived sites in genes from different gene families with XZ1516 as the outgroup. The right
964 panel shows the Pearson correlation between Fay and Wu's *H* and number of high-frequency
965 derived sites. (C) The average Fay and Wu's *H* value in non-Hawaiian population calculated
966 using different strains as the outgroup. XZ1516 and ECA701 are distant from all other strains
967 and may be considered to contain the most ancestral alleles. ECA396 and ECA742 are
968 representative strains from "Hawaii_1" and "Hawaii_2" subpopulations. (D) The cumulative
969 distribution of the *H* values of all genes, F-box, or csGPCR genes calculated using ECA396 or
970 ECA742 as the outgroup.
971
972 **Figure 7. Different selective pressure on different domains of F-box and csGPCR proteins.** (A)
973 Cumulative distribution of *Pi* and Fay and Wu's *H* for nonsynonymous SNVs in the F-box domain
974 or putative substrate-binding domains of F-box proteins. (B) Distribution of *Pi* and *H* for SNVs
975 in the transmembrane domain or extracellular or intracellular domains of csGPCRs. (C) The
976 domain structure of an F-box protein encoded by *fbxb-49*. The F-box domain is in blue, and

977 the type 2 F-box associated (FBA_2) domain, likely involved in binding substrate, is in cyan. (D)
978 The domain structure of a csGPCR encoded by *srw-68*. The predicted transmembrane (TM)
979 domain is in green. Extracellular loops (Out.) and intracellular (In.) tails are indicated. In both
980 (A) and (B), the panels immediately below the domain structure indicate the position of high-
981 frequency (>0.5) derived sites in non-Hawaiian populations using XZ1516 as the outgroup. Y-
982 axis indicates the frequency of the derived alleles among the non-Hawaiian population (black
983 dots) or the Hawaiian population (red dots). Each dot indicates a nonsynonymous SNVs. SNVs
984 causing amino acid substitution with PROVEAN score below -2.5 were shown. The lower two
985 panels showed the high-frequency derived sites in the non-Hawaiian population calculated
986 using ECA396 ("Hawaii_1" strain; purple dots) or ECA742 ("Hawaii_2" strain; blue dots) as the
987 outgroup.
988
989 **Figure 8. F-box and csGPCR genes are enriched in the genomic regions with selective**
990 **footprint identified by extended haplotype homozygosity (EHH) analysis.** (A-C) Manhattan
991 plots of the extent of haplotype homozygosity measured by the integrated Haplotype Score
992 (iHS) within the non-Hawaiian population (A) and Hawaiian population (B). (C) Regions of
993 selection in non-Hawaiian population but not the Hawaiian population indicated by the
994 Manhattan plots of cross-population EHH (XPEHH). (D) The number of F-box and csGPCR genes
995 that contain SNVs with significant iHS or XPEHH and their folds of enrichment. For extended
996 regions, significant SNVs that are less than 50-kb apart were connected to generate regions
997 with selective footprints. (E) The mean Fay and Wu's *H* values for all genes, F-box, and csGPCR
998 genes in the arms and the center of chromosome (Chr) II, III, and V. (F) The domain structure
999 of a representative F-box protein coded by *fbxa-85*; the F-box domain is in blue and the FTH
1000 domain in cyan. (G) The domain structure of a representative csGPCR coded by *srw-56*; the
1001 predicted transmembrane (TM) domain is in green, and extracellular loops (Out.) and
1002 intracellular (In.) tails are also indicated. Among the sites whose XPEHH > 2 in the two genes,
1003 the ones that are also high-frequency (> 0.5) derived sites with ECA396 (purple dots) and
1004 ECA742 (blue dots) as the outgroup are shown.

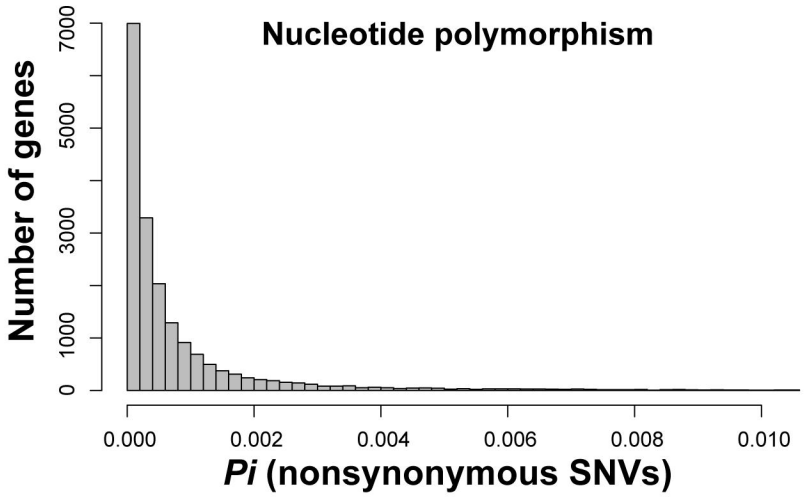
A



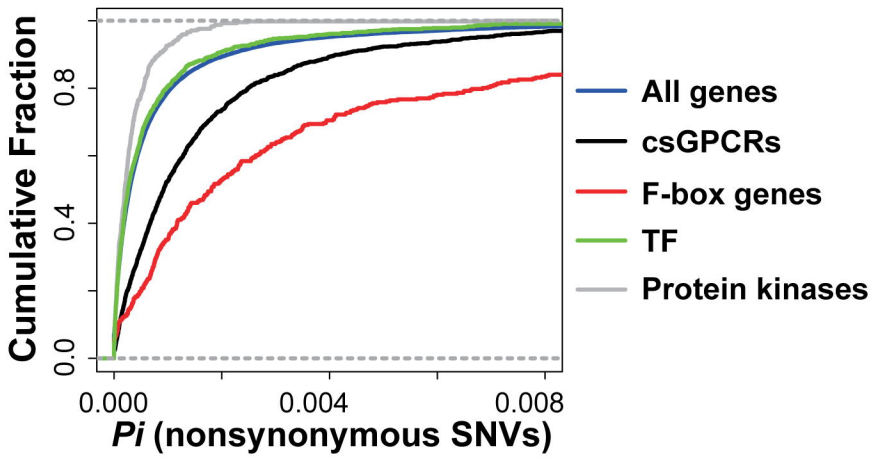
B

		Pi > 0.01		*Pi* > 0.005		*Pi* > 0.0025	
	Total #	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	

C



D

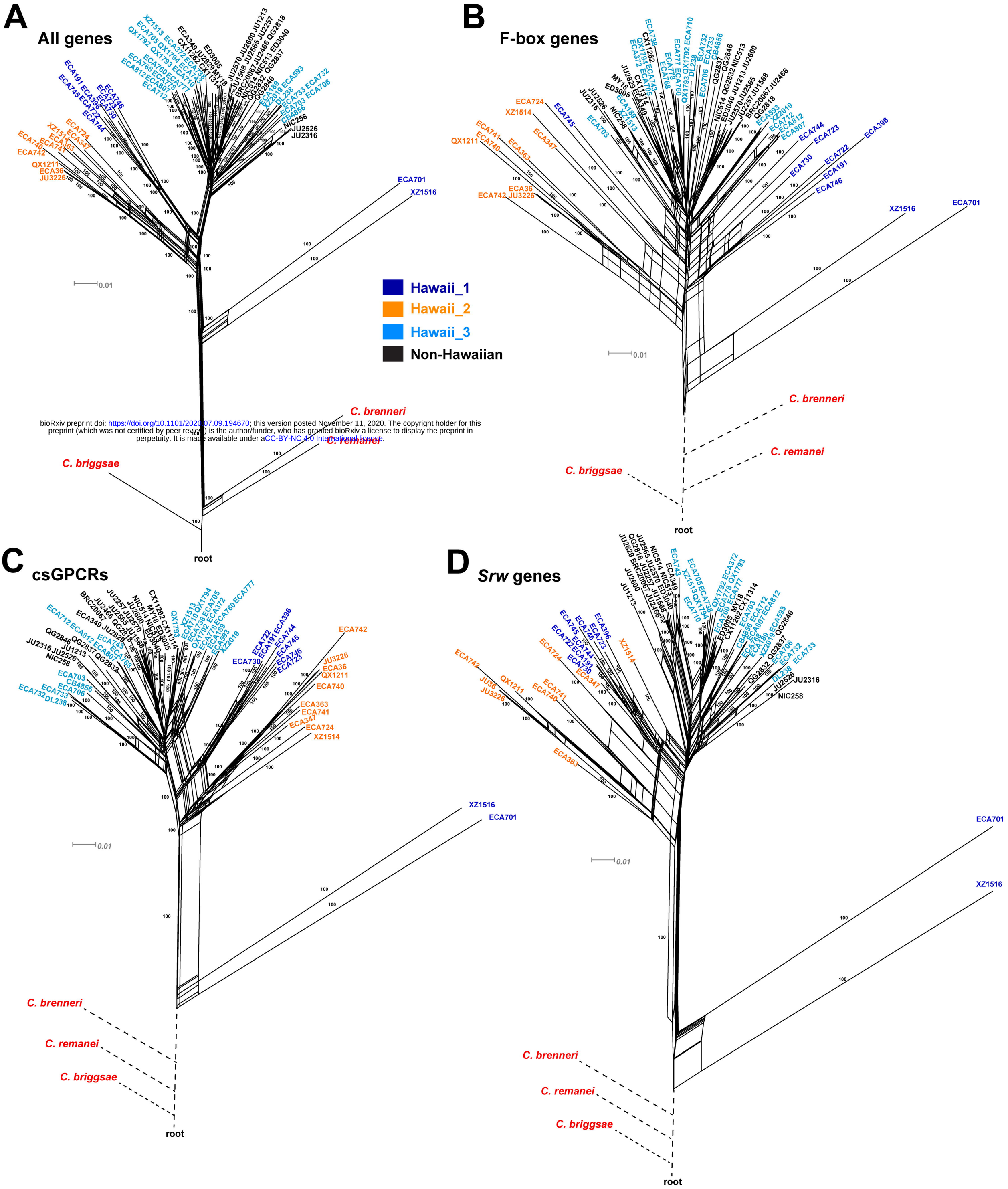


E

	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

F

csGPCR group	*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



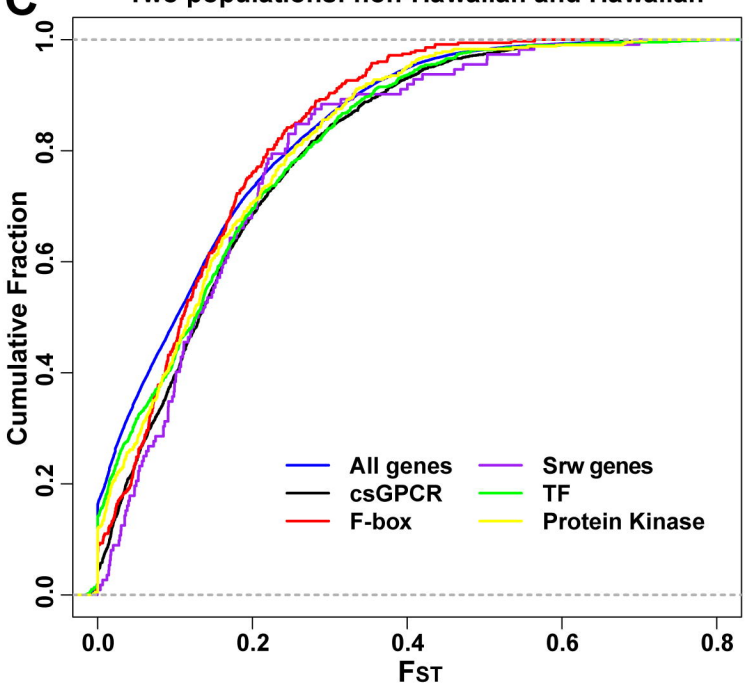
A

π	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

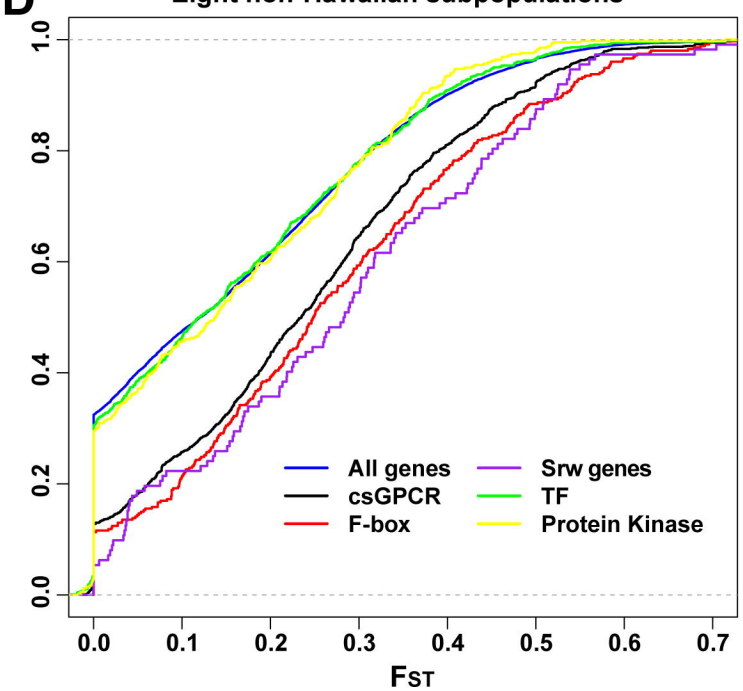
B

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

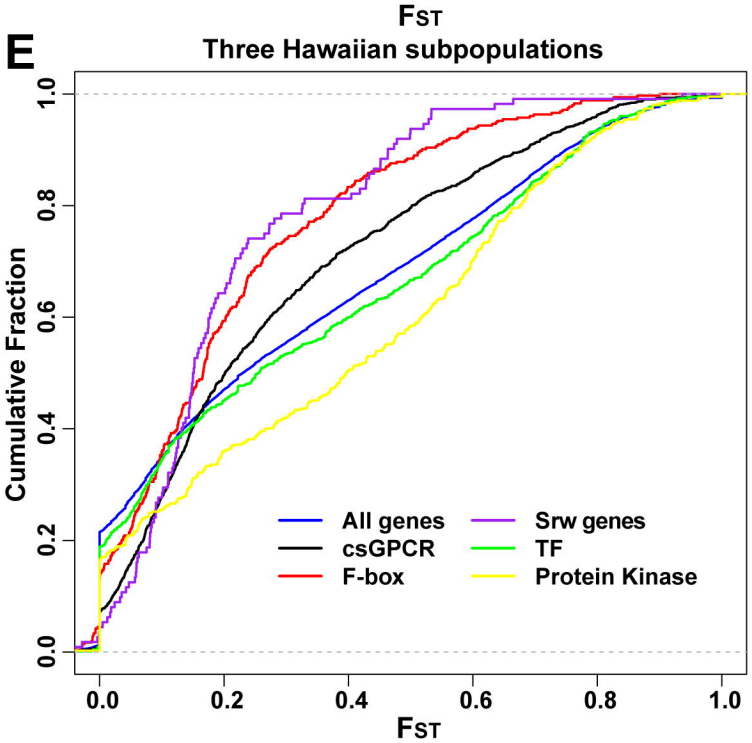
C



D



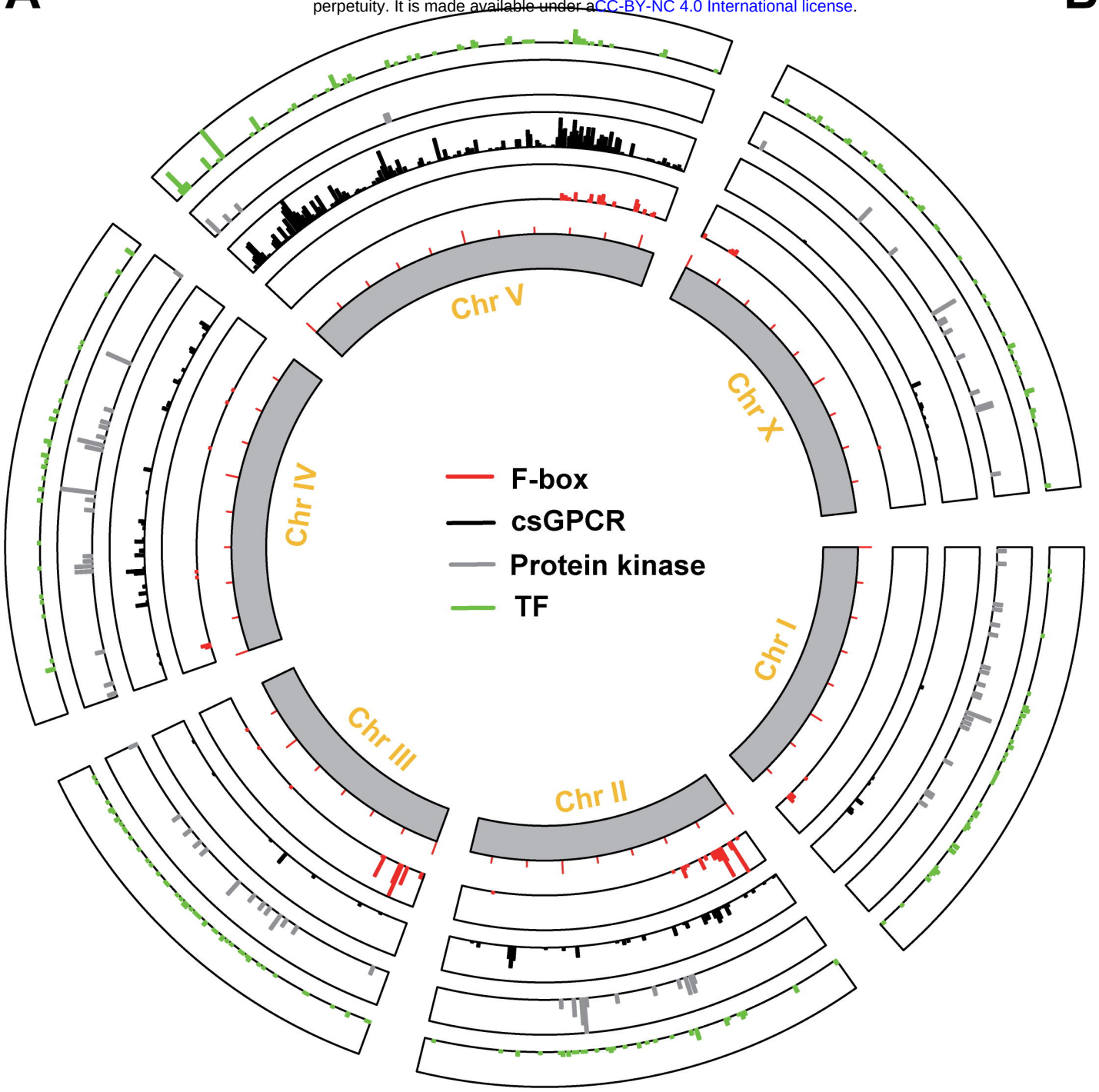
E



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

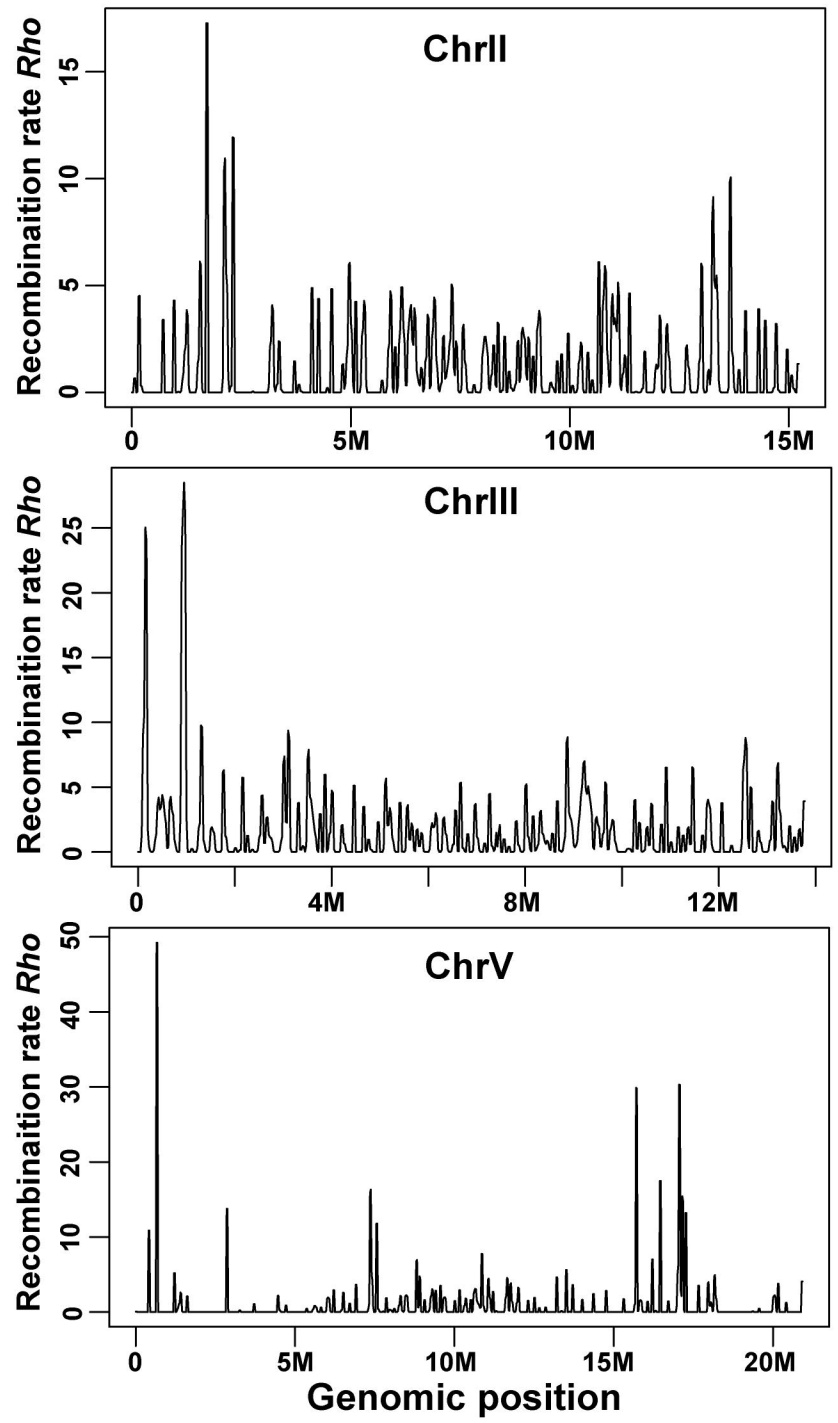
A



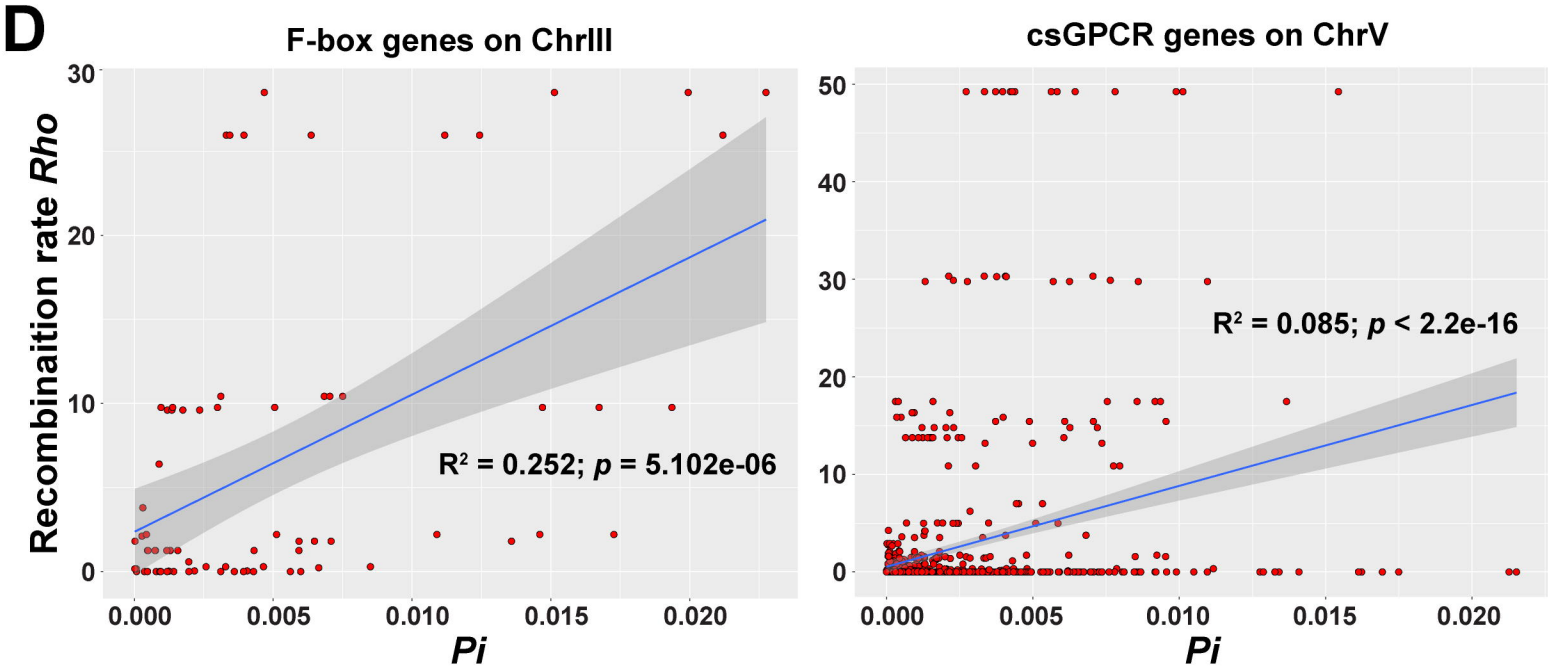
B

	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)

C



D



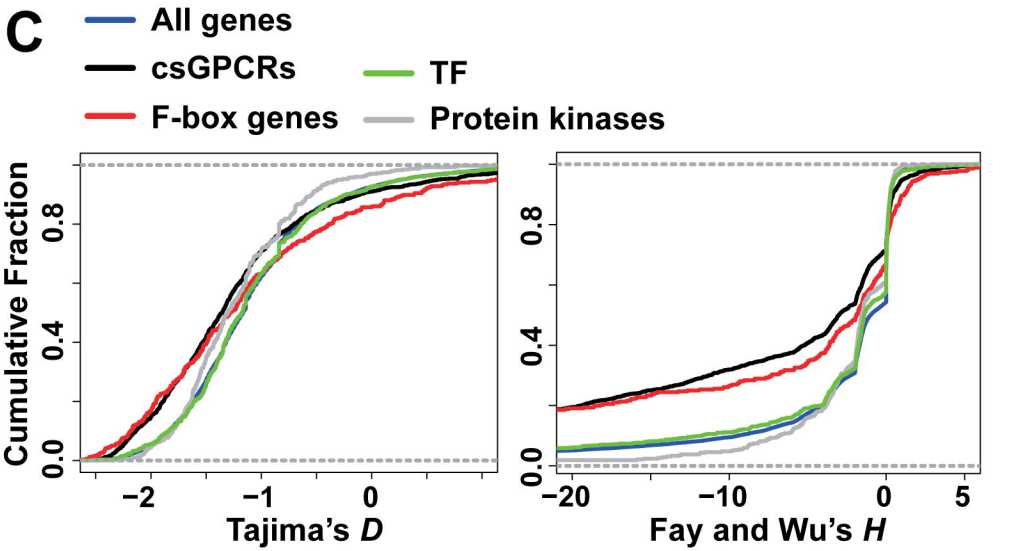
A

		Tajima's *D* < -2		Fay and Wu's *H* < -20		Overlap set	
	Total number	Number	Enrichment fold	Number	Enrichment fold	Number	Enrichment 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	

B

	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

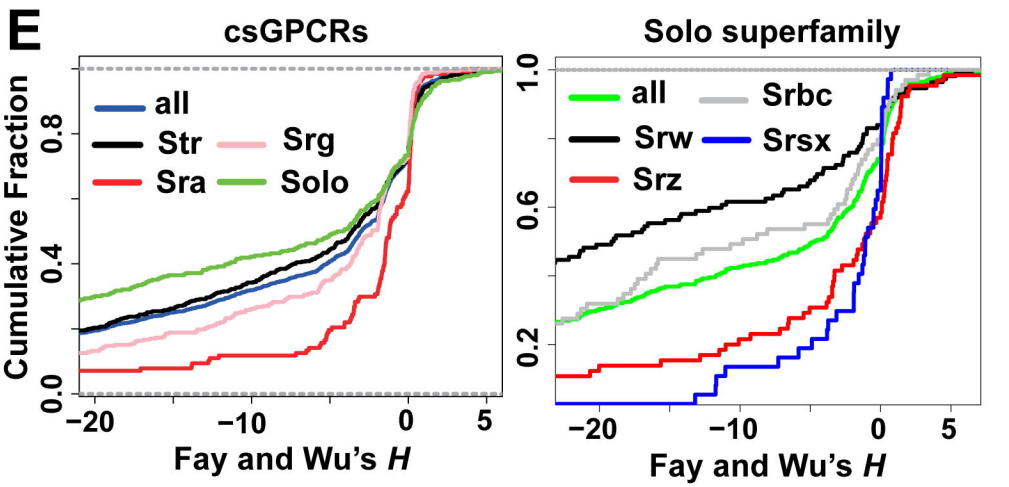
C



D

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

E

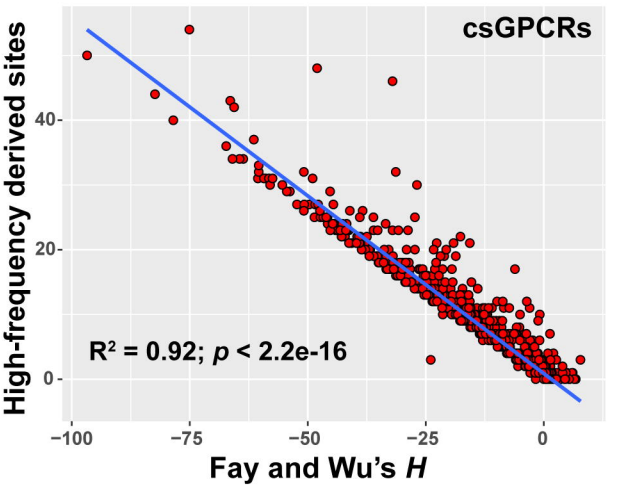


A

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

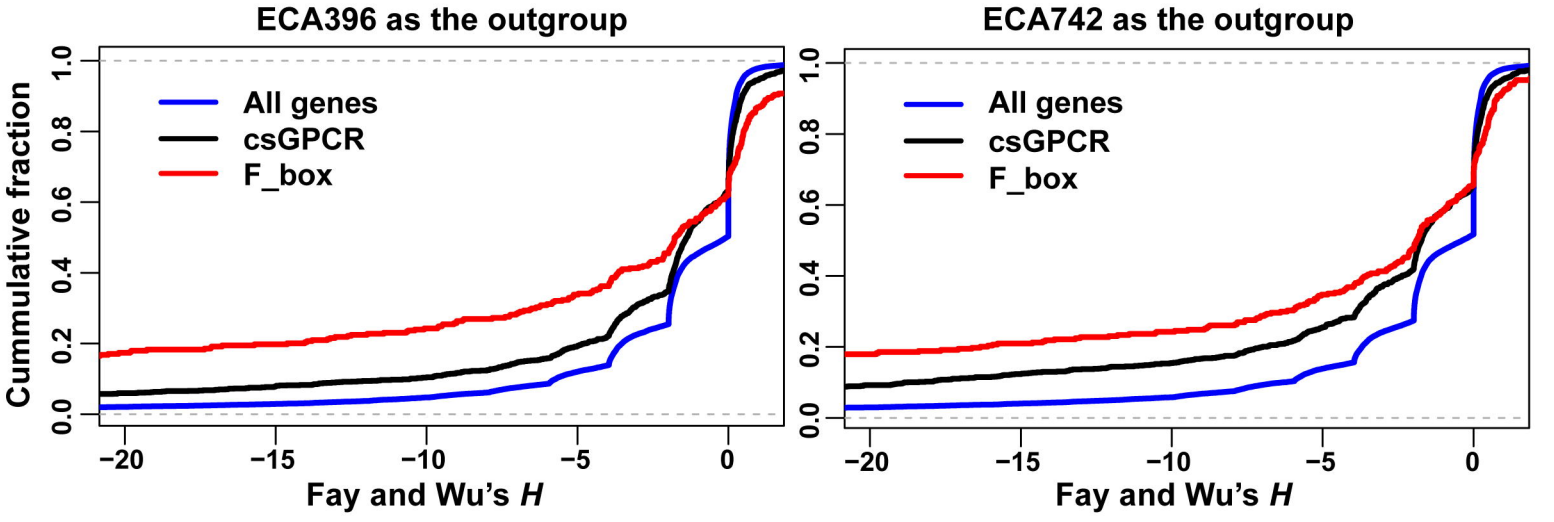
	Number of high frequency derived sites with XZ1516 as the outgroup		
	Among all strains	Among Hawaiian	Among non-Hawaiian
All gene	2.73	2.58	2.76
csGPCRs	6.24	6.18	6.28
Srw GPCRs	14.19	13.62	14.22
F-box	8.80	8.07	8.87
TF	2.97	3.00	3.01
Protein Kinase	1.90	1.63	1.93

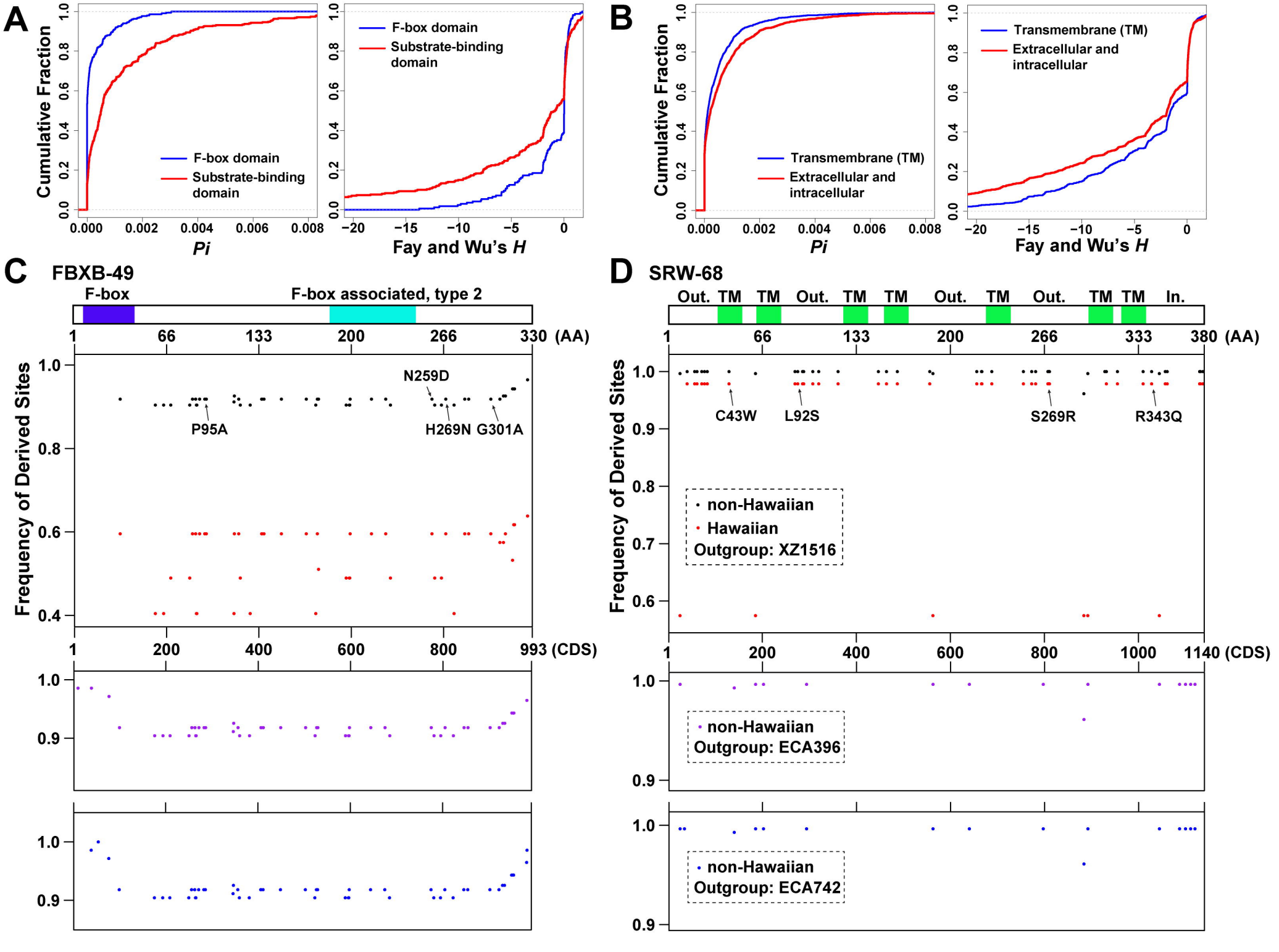


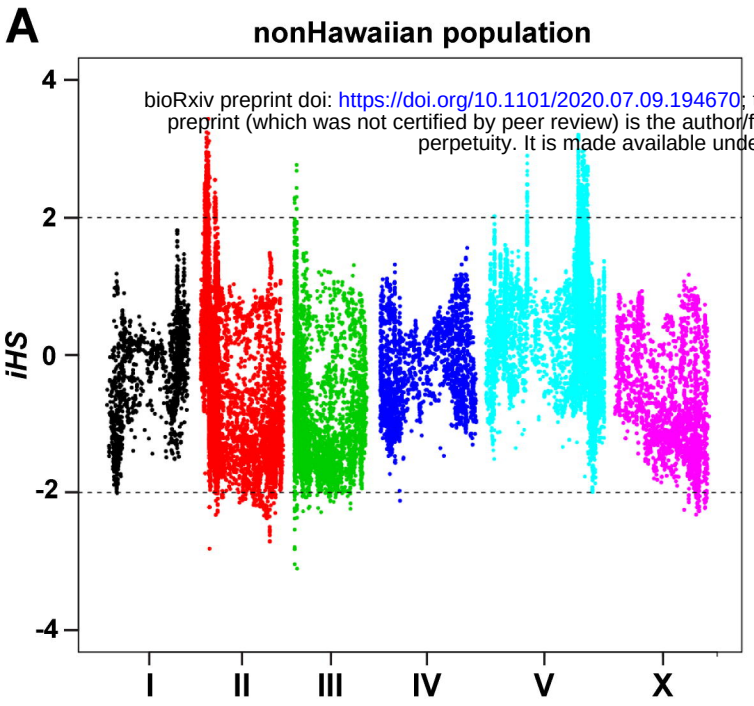
C

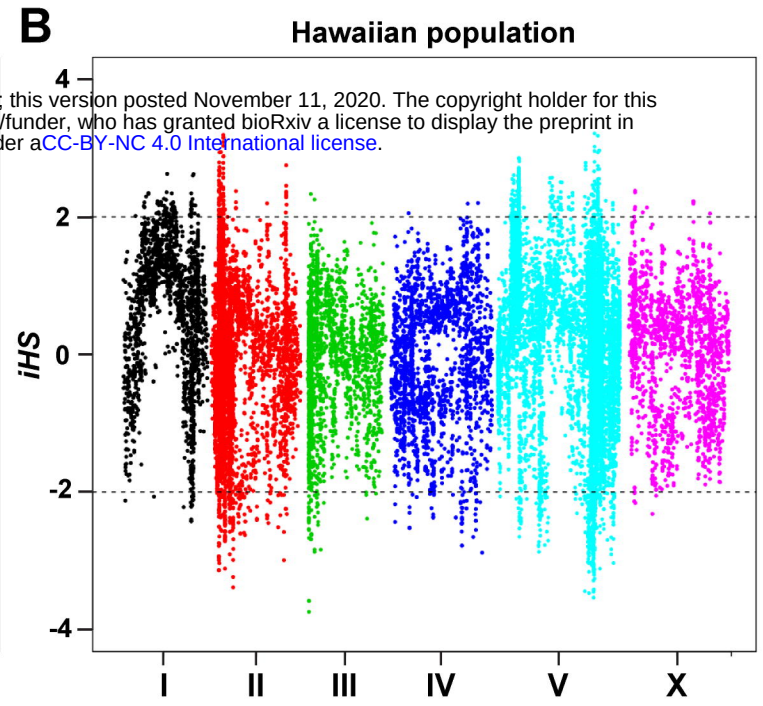
	Fay and Wu's *H* in non-Hawaiian Population			
Outgroup	XZ1516 (ancestral)	ECA701 (ancestral)	ECA396 (Hawaii_1)	ECA742 (Hawaii_2)
All genes	-4.24	-4.03	-2.42	-2.90
csGPCRs	-10.19	-11.01	-4.09	-5.61
F-box genes	-12.79	-16.78	-9.54	-9.51

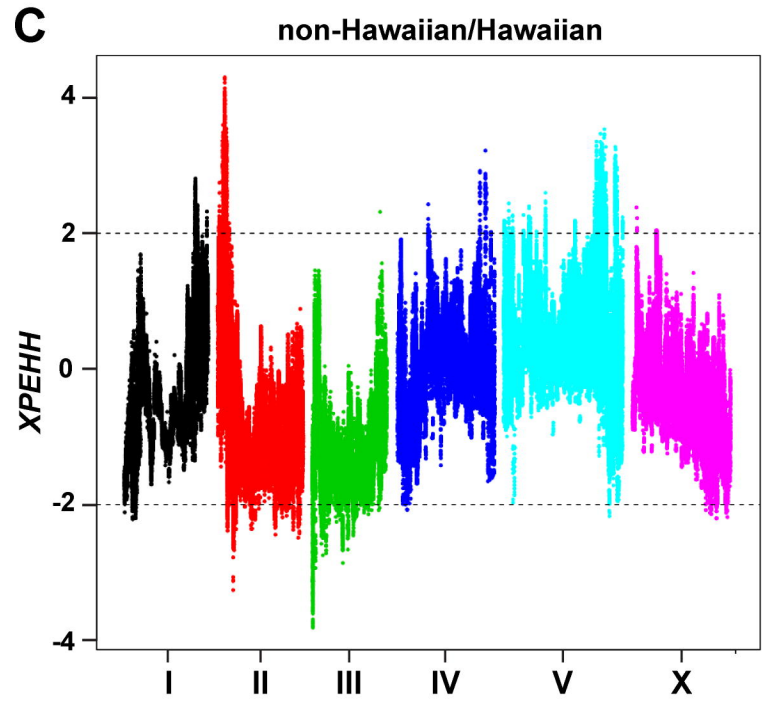
D




A
Cumulative Fraction
F-box domain
Substrate-binding domain
Pi
Fay and Wu's H
B
Transmembrane (TM)
Extracellular and intracellular
C
FBXB-49
F-box
F-box associated, type 2
(AA)
Frequency of Derived Sites
P95A
N259D
H269N
G301A
(CDS)
D
SRW-68
Out.
TM
In.
C43W
L92S
S269R
R343Q
non-Hawaiian
Hawaiian
Outgroup: XZ1516
Outgroup: ECA396
Outgroup: ECA742







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