Characterization of genetic diversity in the nematode *Pristionchus pacificus* from population-scale resequencing data

Christian Rödelsperger¹, Richard A. Neher², Andreas M. Weller¹, Gabi Eberhardt¹, Hanh Witte¹, Werner E. Mayer¹, Christoph Dieterich³, Ralf J. Sommer¹

¹Department for Evolutionary Biology, Max Planck Institute for Developmental Biology, 72076, Tübingen, Germany, ²Evolutionary Dynamics and Biophysics Group, Max Planck Institute for Developmental Biology, 72076, Tübingen, Germany, ³Berlin Institute for Medical Systems Biology (BIMSB) at the Max Delbrück Center for Molecular Medicine Berlin, Robert-Rößle-Str. 10, 13125 Berlin, Germany

(Dated: January 10, 2014)

The hermaphroditic nematode *Pristionchus pacificus* is an established model system for comparative studies with *Caenorhabditis elegans* in developmental biology, ecology and population genetics. In this study, we present whole genome sequencing data of 104 *P. pacificus* strains and the draft assembly of the obligate outcrossing sister species *P. exspectatus*. We characterize genetic diversity within *P. pacificus* and investigate the population genetic processes shaping this diversity. *P. pacificus* is ten times more diverse than *C. elegans* and exhibits substantial population structure that allows us to probe its evolution on multiple time scales. Consistent with reduced effective recombination in this self-fertilizing species, we find haplotype blocks that span several megabases. Using the *P. exspectatus* genome as outgroup, we polarized variation in *P. pacificus* and found a site frequency spectrum (SFS) that decays more rapidly than expected in neutral models. The SFS at putatively neutral sites is U-shaped, which is a characteristic feature of pervasive linked selection. Based on the additional findings i) that the majority of nonsynonymous variation is eliminated over time scales on the order of the separation between clades, ii) that diversity is reduced in gene rich regions, and iii) that highly differentiated clades show very similar patterns of diversity, we conclude that purifying selection on many mutations with weak effects is a major force shaping genetic diversity in *P. pacificus*.

Genetic variation originates from mutation and recombination and is removed by selection and drift. Many hermaphroditic nematode species reproduce by self-fertilization with occasional outcrossing events via males, rather than obligatory outcrossing such as in most animals. Similarly, many plant species self-fertilize or reproduce vegetatively. Self-fertilization reduces effective recombination and can result in a decrease of neutral diversity through selection against linked deleterious variants (background selection) (Charlesworth et al., 1993) or “hitchhiking” with linked beneficial variants (Gillespie, 2000; Maynard Smith and Haigh, 1974). In addition, linkage between loci with non-neutral variation leads to Hill-Robertson interference (Hill and Robertson, 1966) and more generally to a reduced efficacy of selection (Barton and Charlesworth, 1998).

Such random associations of variants with genetic backgrounds of different fitness generate a stochastic force – genetic drift – distinct from common genetic drift (Gillespie, 2000). In addition to reducing neutral genetic diversity, drift results in different patterns of variation that can be used to disentangle the roles of drift and draft. Even though the term draft was introduced in the context of recurrent hitch-hiking, we use it here more generally to refer to the effect of linked selection of any kind as it has been shown that the nature of the selected variation is immaterial as long as many loci contribute to fitness diversity (Neher and Hallatschek, 2013). The prediction that neutral diversity should be negatively correlated with the local recombination rate has been verified in populations of *Drosophila* and *Caenorhabditis briggsae* (Begun and Aquadro, 1992; Begun et al., 2007; Cutter and Choi, 2010). A comprehensive review of the theoretical concepts of linked selection and the empirical evidence for different scenarios can be found in (Cutter and Payneur, 2013; Neher, 2013).

The hermaphroditic nematode *Pristionchus pacificus* has a global distribution and a clearly defined ecological niche, i.e. it is found with scarab beetles in a necromenic association: nematodes infest beetles as arrested dauer larvae, wait for the insect’s death to resume development and feed on the growing microbes on the carcass of the beetle. The *P. pacificus* system was introduced as a satellite model to the well established *C. elegans* multiple comparative studies have revealed divergent patterns in vulva formation, dauer development, and feeding behavior (Bumbarger et al., 2013; Kienle and Sommer, 2013; Sinha et al., 2012).

Similar to *C. elegans*, *P. pacificus* has six chromosomes, which were shown to be largely macrosyntetic to their *C. elegans* counterparts (Lee et al., 2003; Srinivasan et al., 2002). Nonetheless, the genomes exhibit substantial differences in terms of size, gene content, and repeat patterns (Dieterich et al., 2008; Molnar et al., 2012). Furthermore, the generation of a genetic linkage map (Srinivasan et al., 2002) and a combined physical map (Srinivasan et al., 2003) revealed large differences to *C. elegans* with regard to the observed recombination patterns. There is limited interference in *P. pacificus* and multiple crossover events were observed in all 48 in-
individuals of the meiotic mapping panel that the genetic linkage map was built upon. Also, in contrast to *C. elegans*, the *P. pacificus* recombination patterns along the chromosomes do not show an obvious distinction between central regions with high gene density and low recombination rates as opposed to chromosomal arms with low gene density and high recombination rate (Srinivasan et al., 2003).

Phylogenetic studies have shown that hermaphroditism has evolved multiple times within the genera *Caenorhabditis* and *Pristionchus* (Denver et al., 2011). Recently, a very closely related outcrossing sister species, *P. exspectatus*, has been isolated from stag beetles in Japan that still forms viable but sterile F1 hybrids with *P. pacificus* (Kanzaki et al., 2012). Similar to the comparative studies of selfing and outcrossing species in plants (Foxe et al., 2009; Wright et al., 2002), the close phylogenetic relationship between *P. pacificus* and *P. exspectatus* provides a powerful framework for studying genome evolution and associated population genetic processes in predominantly self-fertilizing nematodes.

In this study, we present the draft genome of *P. exspectatus* and whole-genome sequencing data of 104 *P. pacificus* strains. Using *P. exspectatus* as outgroup, we investigated population structure and genetic diversity of *P. pacificus*. Our findings suggest that genetic diversity in *P. pacificus* is shaped by linked selection such as background selection on many weakly deleterious polymorphisms.

I. RESULTS

A. The draft genome of *P. exspectatus*

We sequenced an inbred strain of *P. exspectatus* to an approximate coverage of 97× on the Illumina platform using libraries with different insert sizes (see Methods for details). The genome draft contains 167Mb of assembled sequence in 4,412 scaffolds spanning a total length of 177.6Mb with an N50 value of 142kb (Table 1). Realignment of the sequencing data placed 96-97% of reads onto the assembly, 95% of mapped pairs were in correct orientation.

Despite full-sibling inbreeding for ten generations we observed 120,000 positions with ambiguous base calls from the realignments of sequencing data of the inbred *P. exspectatus* strain. In order to test, whether these positions represent remaining heterozygosity not eliminated by inbreeding, we sequenced the non-inbred *P. exspectatus* strain and identified 330,000 ambiguous positions indicating that the degree of ambiguous base calls was reduced to 38% during the inbreeding process. The theoretical reduction of heterozygosity during a full-sibling inbreeding process can be approximated by

\[ h_t = 1.17 h_0 \times (0.809)^t \]

where \( h_0 \) denotes the initial heterozygosity and \( t \) is the number of inbred generations (Naglyaki, 1992). After ten generations, the expected heterozygosity \( h_{10} \) should be 14% much lower than the observed 38%. Although we cannot rule out a decreased efficacy of inbreeding due to recessive lethal alleles or inversions, these results suggest that up to 57% of the observed ambiguous positions do not represent remaining heterozygosity. We hypothesized that these ambiguous positions may be due to overcompressed repetitive sequences or recently duplicated regions that could not be resolved by the genome assembler. In accordance with this, we found a consistently higher read coverage in regions that cover roughly 60% of all ambiguous positions and spanning around 15% of the genome assembly (Supplemental Fig. 1). In further support for unresolved repetitive and recently duplicated regions as a potential source for the ambiguous positions, we even found triallelic positions (Supplemental Fig. 1). In contrast to a previous analysis of assemblies of outcrossing nematodes (Barrière et al., 2009), we found ambiguous regions not only on autosomes but also on the sex chromosome. Thus, we estimate an actual genome size of *P. exspectatus* of around 200Mb. Whole genome alignments between *P. pacificus* and *P. exspectatus* covered 79.2 and 78.9Mb of uniquely alignable sequence, respectively, and revealed 5.2 million substitutions and 1.2 million indels indicating a sequence divergence of approximately 10% that is distributed uniformly across the chromosomes.

Using RNA-seq derived gene models as training set, we predicted 24,642 complete gene models (28,236 including partial models). Excluding genes with ambiguous gene structures due to alternative isoforms, evaluation of single-transcript RNA-seq gene models showed that 80% of expressed exons overlapped predicted exons on the same strand. For 58% of these exons, start and end positions were predicted correctly at nucleotide resolution. However, 44% of predicted exons showed no evidence for expression indicating either false predictions or constitutively low and/or highly spatiotemporally restricted expression. Protein domain annotation using PFAM showed strong correlations between the two species (Spearman’s \( \rho = 0.77, P < 10^{-15} \)). Only two domain families (PF01498 and PF01359) showed a more than 20-fold enrichment in *P. pacificus* relative to *P. exspectatus* (\( P < 10^{-15} \), Fisher’s exact test); these domains correspond to DNA transposons of the mariner family suggesting that this family has been active in the *P. pacificus* lineage following speciation. However, based on inferred structural variations such as duplication and deletions (see below), we found no evidence that this transposon family is still active in *P. pacificus*.

In this study, we use the *P. exspectatus* genome sequence as outgroup for population genetic analysis of *P. pacificus*. A detailed comparison to *P. pacificus* and another closely related species, *P. arc anus*, will be presented elsewhere.
B. Sequencing of 104 natural isolates

To investigate genetic diversity and the underlying population genetic processes, we selected 104 strains of *P. pacificus*, including the reference strain PS312 for second-generation sequencing. Strains were selected based on biogeography, beetle association, and microsatellite patterning (Morgan et al., 2012) (Fig. 1A). Sixty-one of the 104 sequenced strains are from the Island of La Réunion in the Indian Ocean, which represents a hot-spot of *P. pacificus* biodiversity and has been the focus of recent population genetic studies on *P. pacificus* (Morgan et al., 2012). All strains were inbred for at least ten generations and known to be largely homozygous at microsatellite markers. We sequenced genomic DNA of these strains on the Illumina platform with mean coverage ranging from 6 to 37×. Most strains were sequenced to a coverage of 9× (median, interquartile range (IQR): 8 – 12), while a few strains were sequenced much deeper.

Individual strains showed between 23,000 and 1.4 million single nucleotide variants (SNVs) relative to the reference genome (Supplementary Table 1). In total, we identified 7.1 million SNVs and 2.1 million indel (< 100 bp) positions. We assessed the quality of variant calls by comparing the SNVs derived from short read alignments with the SNVs derived from whole genome alignments of the Sanger sequenced *P. pacificus* mapping strain PS1843 from Washington (Dieterich et al., 2008). Out of 470,000 SNV calls that are covered by both platforms, 98% were in agreement, indel calls showed an agreement of 79-96% (see Methods).

Despite extensive inbreeding (at least 10 generations), around 6% (median, IQR: 5.2 – 7.3%) of SNVs for all strains were called as heterozygous (Supplemental Table 1). As in the case of *P. exspectatus*, several lines of evidence suggest that a large portion of the observed heterozygous signal is due to recent duplications. First, 29% (median, IQR: 26 – 31%) of these heterozygous SNVs fall into duplications that are larger than two kilobases (Supplemental Table 1), which are the smallest duplicated regions that can be reliably detected based on read coverage. Second, a subset of 20 strains that was inbred for more than 30 generations shows similar levels of heterozygosity than the other strains. Third, the isolate RS5410 from La Réunion (Fig.1B), which was previously identified as an admixed strain (Morgan et al., 2012) and should therefore show a higher level of heterozygosity, only exhibits 4.4% of ambiguous variant calls, of which 27% fall into large duplicated regions. In order to minimize the effect of coverage fluctuations and apparent heterozygosity, most population genetic analysis presented below is restricted to positions that were sufficiently covered in all analyzed strains and did not show any signal of heterozygosity (see Methods).

As mentioned above, we detected structural variants and copy number variations between two kilobase up to a megabase using a method (cnv-seq, Xie and Tammi (2009)) that compares differences in read depth relative

<table>
<thead>
<tr>
<th>Feature</th>
<th>Unit</th>
<th><em>P. pacificus</em></th>
<th><em>P. exspectatus</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembly size</td>
<td>[Mb]</td>
<td>172.5</td>
<td>177.6</td>
</tr>
<tr>
<td>Assembled sequence</td>
<td>[Mb]</td>
<td>153.2</td>
<td>167.0</td>
</tr>
<tr>
<td>N50 scaffold size</td>
<td>[Mb]</td>
<td>1.25</td>
<td>0.14</td>
</tr>
<tr>
<td>GC-content</td>
<td></td>
<td>42.8%</td>
<td>42.8%</td>
</tr>
<tr>
<td>Predicted genes</td>
<td></td>
<td>28,666</td>
<td>24,642</td>
</tr>
<tr>
<td>Coding sequence</td>
<td>[Mb]</td>
<td>27.5</td>
<td>27.1</td>
</tr>
<tr>
<td>Gene length</td>
<td>[kb]</td>
<td>1.9 (1.0-3.5)</td>
<td>2.5 (1.3-4.5)</td>
</tr>
<tr>
<td>Transcript length</td>
<td>[kb]</td>
<td>0.7 (0.4-1.2)</td>
<td>0.8 (0.4-1.4)</td>
</tr>
<tr>
<td>Exons per gene</td>
<td></td>
<td>7 (4-12)</td>
<td>8 (5-14)</td>
</tr>
<tr>
<td>Exon length</td>
<td>[bp]</td>
<td>87 (65-115)</td>
<td>86 (63-112)</td>
</tr>
<tr>
<td>Intron length</td>
<td>[bp]</td>
<td>119 (58-242)</td>
<td>128 (63-253)</td>
</tr>
</tbody>
</table>

Table 1 Features of the *P. pacificus* and *P. exspectatus* genome sequences. Gene features are given as median and IQR.
to resequencing data of the reference strain (Supplementary Fig. 2). Compared to the reference genome, 7.2% (median, IQR: 6.8–7.4%) and 2.4% (median, IQR: 2.3–2.6%) of the genome was predicted as deleted and duplicated, respectively. We verified predicted deletions with PCR amplification experiments for three cellulase genes and microsatellite data indicated by different colors in the figure (Morgan et al., 2012). The strong population subdivision is in contrast to the lack of population structure in C. elegans (Phillips, 2006) but mirrors findings in C. briggsae (Cutter et al., 2006). Consistent with the strong population structure, the first two principal components (PCs) of the SNV data explain 22.9% and 16.3% of the global variability, respectively (Supplemental Fig. 3). The clades A1, A2, and C are clearly separated along the first two PCs. Further analysis of clade C, the most deeply sampled clade with a total of 44 strains, did not reveal additional clusters but only a slight signal reflecting the local geography on La Réunion Island. We focus our population genetic analyses below on this homogeneous clade C to minimize the influence of population structure. Nevertheless, the highly structured population of *P. pacificus* and its outgroup *P. exspectatus* allows us to investigate patterns of sequence evolution over a wide range of time scales. The differences that accumulated between *P. pacificus* and *P. exspectatus* clearly contribute to divergence between species, while mutations that segregate within each of the clades are relatively young genetic diversity. Comparisons between clades have features of diversity and divergence.

C. Genome-wide levels of nucleotide diversity

Figure 1B shows the phylogenetic relationship of the 104 sequenced strains as a split-network. The populations fall into a number of clearly separated clades and we denote the most deeply sampled clades with at least 15 strains each as A1, A2, and C. The clades suggested by diversity in the nuclear genome largely agree with the previous designation based on the mitochondrial genome and microsatellite data indicated by different colors in the figure (Morgan et al., 2012). The strong population subdivision is in contrast to the lack of population structure in *C. elegans* (Phillips, 2006) but mirrors findings in *C. briggsae* (Cutter et al., 2006). Consistent with the strong population structure, the first two principal components (PCs) of the SNV data explain 22.9% and 16.3% of the global variability, respectively (Supplemental Fig. 3). The clades A1, A2, and C are clearly separated along the first two PCs. Further analysis of clade C, the most deeply sampled clade with a total of 44 strains, did not reveal additional clusters but only a slight signal reflecting the local geography on La Réunion Island. We focus our population genetic analyses below on this homogeneous clade C to minimize the influence of population structure. Nevertheless, the highly structured population of *P. pacificus* and its outgroup *P. exspectatus* allows us to investigate patterns of sequence evolution over a wide range of time scales. The differences that accumulated between *P. pacificus* and *P. exspectatus* clearly contribute to divergence between species, while mutations that segregate within each of the clades are relatively young genetic diversity. Comparisons between clades have features of diversity and divergence.

Figure 2 shows the nucleotide diversity within clades A2 and C (see Supplemental Fig. 4 for comparisons including clade A1), the average distances between these two clades, and the distance between *P. pacificus* and *P. exspectatus* for each of the six chromosomes. The average genome-wide intra-clade diversity in 100kb windows is 1.9 and 2.2 \( \times 10^{-3} \) for clade A2 and C, respectively. Strains from clade A2 and C differ at an average fraction of 7.5 \( \times 10^{-3} \) of all sites. The genome-wide patterns of diversity within clades are well correlated between clades (Spearman’s \( \rho = 0.75 \), \( P < 10^{-15} \) for clades A2 and C in 100kb windows). Similarly, distance between the clades is correlated with diversity within clades C and A2 (Spearman’s \( \rho = 0.63 \) and 0.66, \( P < 10^{-15} \), respectively). However, the distances between clades fluctuate less in relative terms than diversity within clades (standard deviation of log10 distance values in 100kb windows is 0.21 as opposed to approx. 0.30 for fluctuations of within clade diversity). The average divergence between *P. pacificus* and *P. exspectatus* is 9.3% with a typical fluctuation of less than 20% (windows of 100kb, standard deviation of
This homogeneous divergence between the sister species suggests a constant accumulation of mutations and the absence of large-scale mutation rate variation. Hence the strongly fluctuating diversity patterns within each clade are likely due to population genetic processes that differentially affect genetic diversity in different regions of the genome such as selective sweeps or background selection (Hudson and Kaplan, 1995; Maynard Smith and Haigh, 1974; Nordborg et al., 1996). Even though the diversity patterns in different clades are very similar, most intra-clade diversity is due to private variation, rather than diversity shared in the common ancestral population (3% and 7% of polymorphic positions in clade C and A2 are polymorphic in both, the joint SFS of different clades are shown in Supplementary Figure 5). This suggests that genetic diversity is shaped by processes that act similarly but independently in different clades. The most plausible explanation is background selection, which depends primarily on the local deleterious mutation rate and recombination (Hudson and Kaplan, 1995; Nordborg et al., 1996). These two processes are not expected to be strongly influenced by the environment and should affect diversity in similar matter independent of location. Local adaptation and the associated selective sweeps, in contrast, should result in different diversity patterns within different clades (Kawecki and Ebert, 2004). The alternative explanation of global selective sweeps seems unlikely, as we do not find long haplotypes that are shared among clades (Supplementary Fig. 6). The correlation between across-clade and within-clade diversity likely stems from ancestral diversity. This ancestral diversity is expected to have had a chromosomal profile similar to the present day within-clade diversity. Together with homogeneous accumulation of divergence, we expect the inter-clade distance to be correlated with the intra-clade diversity, but with a smaller relative variability. Both of these expectations are consistent with our observations.

D. Deleterious variants at high frequencies

Restricting the analysis of the global P. pacificus diversity to coding regions, we find a mean silent site diversity $\pi_{si} = 1.4 \times 10^{-2}$ and a nonsynonymous site diversity of $\pi_{ns} = 4.5 \times 10^{-3}$. The resulting $pN/pS \approx 0.32$ suggests widespread purifying selection that results in pruning of two out of three nonsynonymous mutations. Note, however, that these genome-wide measures of synonymous and nonsynonymous diversity are dominated by inter-clade comparisons. Restricted to clade C, we find an average $pN/pS$ ratio of 0.39, which indicates that a substantial fraction of nonsynonymous mutations were only pruned at the inter-clade level and segregate within clades.

To quantify this dependence of purifying selection on the time scale of separation, we calculated the density of synonymous differences, $\delta_{si}$, and nonsynonymous differences, $\delta_{ns}$, for all pairwise comparisons of the 104 strains. As comparisons across clades have features of diversity and divergence we have chosen to represent the ratio between nonsynonymous and synonymous differences as $\delta_{ns}/\delta_{si}$ rather than $pN/pS$ or $dN/dS$. Fig. 3 shows the dependence of this ratio $\delta_{ns}/\delta_{si}$ on the separation measured by the synonymous distance $\delta_{si}$ (stochastic fluctuations in $\delta_{si}$ are negligible since we are comparing entire genomes and the total number of differences in each comparison is large). The most closely related strains within one clade show $\delta_{ns}/\delta_{si}$ of up to 0.5. With increasing distance, $\delta_{ns}/\delta_{si}$ drops from 0.5 to 0.3 when comparing strains between clades. In other words, we find that nonsynonymous differences accumulate with a decreasing rate as we go to larger degrees of separation. Finally, we find $dN/dS=0.14$ for the comparison of 1:1 orthologs between P. pacificus and P. exspectatus. Note that as we go from comparisons of strains within one clade to inter-clade comparisons and eventually to inter-species comparison, the interpretation of $\delta_{ns}/\delta_{si}$ gradually changes from $pN/pS$ to $dN/dS$. For a more indepth discussion of
time dependent $dN/dS$, see (Mugal et al., 2014; Rocha et al., 2006).

From the observation $\delta_{ns}/\delta_{si} \approx 0.5$ for closely related strains, we conclude that 50% of nonsynonymous substitutions are so deleterious that they are rarely seen in wild isolates and are eliminated quickly. In contrast, another 20% of nonsynonymous substitutions are weakly deleterious such that they are pruned only on longer time scales and segregate as high frequency polymorphisms. Below, we will complement these observations for genome wide ratios of nonsynonymous and synonymous differences with an analysis of the site frequency spectra (SFS) of synonymous and nonsynonymous mutations.

The most immediate consequence of abundant deleterious mutations is a reduction of diversity by background selection (Charlesworth et al., 1993). Indeed, we find a strong negative correlation between the fraction of coding sequences with diversity (Spearman’s $\rho = -0.35, P < 10^{-15}$ for 100 kb windows in clade C), a finding that has also been observed in the selfing plant Arabidopsis thaliana (Nordborg et al., 2005). Similarly, we expect a positive correlation between genetic diversity and recombination rates (Begun et al., 2007; Cutter and Choi, 2010), but the only available genetic map (Shinvasan et al., 2003) has insufficient resolution to be useful for this purpose. While the observed correlation on its own could also be the result of recurrent hitch-hiking in gene rich regions, the strong correlation of the diversity patterns across clades suggest purifying selection as a likely cause, see discussion above.

E. Site frequency spectra and linked selection

Rare SNVs tend to be young while common SNVs are typically old (Kimura and Ohta, 1973). Hence the histogram of SNVs present in $k$ out of $n$ strains provides a rich summary of genetic diversity that is informative about the demographic and evolutionary history of the population. Deciding on whether an allele is common or rare requires polarization, i.e., an inference of the ancestral state at the locus. To infer the ancestral state of clade C, we broke the genome into 20kb intervals, built maximum likelihood trees on these blocks using fasttree (Price et al., 2009) and rooted these trees with the genome of $P$. exspectatus. Blocks of 20kb are in substantial LD and contain sufficient SNVs to resolve most of the genealogical relationships between strains. Next, we used a probabilistic model to infer the ancestral state at each internal node of the tree (see Methods). Variations of this strategy (different block length, different substitution models, different tree building algorithms, different ancestral reconstruction software) yielded similar results.

There are many more rare derived mutations than common ones. Hence if even a small fraction of rare mutation are incorrectly polarized, the number of high frequency derived mutations is overestimated substantially. For this reason, many authors “fold” the SFS and consider only the minor allele frequency irrespective of polarization. However, valuable information is lost by “folding”. We therefore decided to present the unfolded SFS and we provide independent evidence that our polarization is sufficiently accurate to allow a faithful characterization of the SFS at high frequencies. Fig. 4A shows the ratio of nonsynonymous to synonymous mutations at frequencies above a threshold $\nu$. The ratio of nonsynonymous to synonymous SNVs is monotonically decreasing as expected if purifying selection prunes the majority of amino-acid substitutions. If most of the SNVs inferred to be at high frequency were wrongly polarized rare alleles, this curve should rise again with $\nu \rightarrow 1$.

Fig. 4B shows the polarized SFS of clade C on a double logarithmic scale such that power-laws show as straight lines. At frequencies below 20%, the observed SFS is compatible with a $1/k$ decay as expected in neutrally evolving populations of constant size (Wakeley, 2008). At higher frequencies, however, the SFS decreases much more rapidly, before increasing again for alleles close to fixation. Similar U-shaped SFS have been observed in plants (Cao et al., 2011). At intermediate derived allele frequencies between 20-40%, the slope of the SFS is compatible with a $1/k^2$ indicated as dashed line. The $1/k^2$ behavior and the non-monotonicity is expected if the dominant force changing allele frequencies is selection at linked (genetic draft) loci regardless of whether this variation is positively or negatively selected (Braverman et al., 1995; Neher and Hallatschek, 2013; Neher and Shraiman, 2011).

Compared to the almost 100 fold variation in the SFS between singletons and alleles at intermediate frequencies, synonymous, nonsynonymous and noncoding polymorphisms all follow very similar distributions (Fig. 4C). The relative abundance of synonymous polymorphisms increases only by a factor of two as we go from low to high frequency (see Fig. 4A). This similarity suggests that the dynamics of polymorphisms is quasi-neutral in the sense that their fate is largely determined by their genetic background rather than their own effect on fitness. We observe a systematic excess of synonymous over nonsynonymous mutations at frequencies above 0.6 consistent with the gradual manifestation of purifying selection on old and frequent alleles ($P < 10^{-6}$, Fisher’s exact test, Fig. 4C). While the SFS is quite insensitive to the type of mutation, stratifying the SFS by the overall nucleotide diversity in surrounding regions reveals a strong dependence of the shape of the SFS on coalescence time (Fig. 4C) as predicted by models of abundant linked selection in recombining populations (Neher et al., 2013).

To complement our inference of mutational effects based on pairwise comparisons of strains (see above and Fig. 3), we used the web-server DFE-$\alpha$ (Keightley and Eyre-Walker, 2007) to infer parameters of the DFE using the synonymous and nonsynonymous SFS within clade C. DFE-$\alpha$ preferred a model with a 4.5 fold population size increase ($N_2 = 4.5 N_1$) roughly $N_2$ generations in the past over a model with a single population size
Figure 4  (A) Monotonic decrease of the ratio of the numbers of nonsynonymous (Nns) and synonymous (Nsyn) SNVs at derived allele frequencies (k/n) above a threshold ν. The ratios represent unnormalized ratios of raw counts of nonsynonymous and synonymous SNVs in the site frequency spectrum (SFS). As these sites are called in all 104 P. pacificus strains and are uniquely alignable to the outgroup P. exspectatus, this subset of sites is likely under stronger purifying selection than the genome-wide average. The monotonic decrease with increasing ν is consistent with the effect of purifying selection. There is no signal of polarization errors which would manifest as an uptick at values of ν → 1. (B) SFS of clade C is shown on a double logarithmic plot. If neutral diversity is shaped by constant genetic drift, one expects that the number of derived alleles present in k out of n strains decays as 1/k, indicated as straight solid line. The dashed line indicates the corresponding expectation, 1/k², under a genetic draft model (for rare alleles). At frequencies below 10%, the SFS is proportional to 1/k, at intermediate frequencies between 10% and 50%, the SFS is steeper and compatible with 1/k². (C) The SFS decays less rapidly in regions of high than of low diversity (100 kb windows). In contrast, distinct functional categories have almost indistinguishable SFSs. Only at frequencies above k/n = 0.6, do we see a systematic excess of synonymous over nonsynonymous mutations (P < 10⁻⁵, Fisher’s exact test).

(log-likelihood difference 325). For the two-epoch model, DFE-α estimated that 45% of nonsynonymous mutations have deleterious effects in excess of N2|s| = 100, consistent with our previous results. Another 37% of nonsynonymous mutations are estimated to be approximately neutral with N2|s| < 1, which is compatible with the observed δns/δs at larger time scales (Fig. 3).

The shape of the SFS is sensitive to past demography, direct selection, linked selection, and genetic drift. Disentangling the effects of these processes is challenging as they have overlapping characteristics. We argue that the dominant features of the SFS observed in clade C are the result of linked selection. If direct selection was a dominant force, the SFS of synonymous, nonsynonymous, and non-coding sites would have qualitatively different shape. We observe the effect of purifying selection against nonsynonymous mutations at high frequencies, but the difference is small. Effects of demography are harder to discount. Both exponential expansion and linked selection result in a 1/k² decay of the SFS at low and intermediate frequencies (Neher and Shraiman, 2011). The SFS we observe, however, are non-monotonous and it is impossible to obtain non-monotonous SFS in any neutral model of one population irrespective of past population size changes (see appendix B in Sargsyan and Wakeley (2008)). Models of linked selection or genetic draft predict precisely this type of non-monotonous SFS (Neher and Hallatschek, 2013). Furthermore, the shape of the SFS differs in regions of high and low diversity. Signatures of draft are more pronounced in low diversity regions (Fig. 4C). This observation is consistent with the expectation that draft reduces coalescence time and hence diversity and that this can differ along the genome depending on the local mutational input of selected diversity and recombination rate (Neher et al., 2013). Young alleles at the rare end of the frequency spectrum are compatible with genetic drift and the bump in the SFS around 15% might reflect residual population structure within clade C. Overall, however, P. pacificus is not compatible with a neutral model. Strong distortions of genealogies by purifying selection on many weak effect alleles are a likely explanation for the observed SFS (Neher and Hallatschek, 2013; Walczak et al., 2012). However, our understanding of migration between the subpopulations of P. pacificus remains limited and to what extent reintroduction of ancestral alleles by migrations from different subpopulations contributes to the observed patterns remains subject of future investigation.

F. Linkage disequilibrium and haplotype structure

Finally, we investigated the pattern of linkage disequilibrium (LD) in P. pacificus. Figures 5A and B show the average LD between pairs of loci measured as r² as a function of the physical separation between the loci for the global sample of 104 strains and clades C, A1, and A2. In all cases, LD drops over a distance of about
most of shared haplotype blocks in \textit{P. pacificus} are clade-specific. Consistent with a reduction of neutral diversity through linked selection, we find a moderate negative correlation between LD and diversity (Spearman’s \( \rho = -0.25, P < 10^{-15} \) for 100 kb windows).

Interpretation of the observed haplotype patterns is hampered by the absence of a high resolution genetic map for \textit{P. pacificus}. Whether the long-range LD simply reflects absence of recombination or whether other population genetic forces are responsible for maintaining this LD needs to be addressed in future work. However, we would like to point out that the previous low-resolution genetic map did not find an obvious absence of crossovers in the central regions of the chromosomes (Srinivasan et al., 2003).

\section*{II. DISCUSSION}

By sequencing 104 strains of \textit{P. pacificus} and a closely related outgroup, \textit{P. exspectatus}, we have shown that \textit{P. pacificus} contains extensive population structure and overall levels of diversity are one order of magnitude higher than in \textit{C. elegans} (Andersen et al., 2012). However, individual clades show similar levels of nucleotide diversity with a genome-wide profile that is highly correlated between clades, suggesting that background selection is a major force shaping genetic diversity. The accumulation of divergence between \textit{P. pacificus} and its sister species \textit{P. exspectatus} is homogenous across the genome. The latter observation is consistent with results of the accompanying manuscript presenting a mutation accumulation experiment that shows that the \textit{P. pacificus} genome does not exhibit strong mutation rate variation on large physical scales and that synonymous and nonsynonymous mutations occur at their expected frequency (Weller et al. 2014). By analyzing the ratio of synonymous and nonsynonymous differences between strains at various distances and mutations at different frequencies, we characterized the strength of purifying selection on coding regions. 50\% of nonsynonymous mutations are so deleterious that they are not found in a typical population sample. Of those nonsynonymous mutations that are observed, again roughly half are weakly selected against and pruned over time scales on the order of the separation between clades, which corresponds to roughly 1-2\% divergence at silent sites. Combined with our finding that nucleotide diversity in genomic windows anticorrelates strongly with the fraction of coding sequence in 100kb windows, we conclude that background selection plays an important role in shaping \textit{P. pacificus} diversity.

Recent theoretical work has shown that background selection cannot be fully described by a reduced effective population size but results in substantial distortions of genealogies in particular when many weakly deleterious mutations segregate at high frequency (Seger et al., 2010; Walczak et al., 2012). This distortion of genealo-
gies manifests itself in an SFS with a steeper decay and an up-tick at high derived allele frequencies, both of which are characteristic signatures of genetic draft or linked selection (Neiher and Hallatschek, 2013; Neiher and Shraiman, 2011). The signature of genetic draft in the SFS of *P. pacificus* is consistent with previous studies that proposed selection at linked sites as one important factor shaping genomic diversity of self-fertilizing nematodes (Andersen et al., 2012; Cutter and Choi, 2010; Rockman and Kruglyak, 2009). In addition to distorted SFS, we also find megabase scale haplotype blocks in strong linkage disequilibrium, which provide a strong opportunity for linked selection.

Whether on top of purifying selection adaptive substitutions or fluctuating selection play a prominent role remains currently unclear. McDonald-Kreitman type tests (McDonald and Kreitman, 1991) that have been used to quantify adaptive evolution in *Drosophila* (Sella et al., 2009) are very vulnerable to segregating deleterious mutations (Charlesworth and Eyre-Walker, 2008; Messer and Petrov, 2012) and hence not suitable for *P. pacificus*. Methods used in *Drosophila* to detect more recent adaptations based on signatures of hitch-hiking (Andolfatto, 2007; Macpherson et al., 2007) are also inapplicable in organisms with long range LD.

In summary, the combination of high-diversity in a structured populations and a close outgroup has allowed us to study evolution on a variety of times scales spanning distances from 0.1% to 10% and revealed a consistent decreasing trend in the ratio of nonsynonymous to synonymous mutations (Mugal et al., 2014; Rocha et al., 2006). This trend implies an abundance of weakly deleterious mutations — at least when their effect is averaged over larger and larger time scales. With respect to the comparison to *C. elegans*, our genome-wide analysis of *P. pacificus* populations describes a complementary picture that may reflect the substantially different natural histories of both nematodes. *P. pacificus* and *C. elegans* are both predominantly self-fertilizing species, but they occupy distinct ecological niches and it has been speculated that the reduction of diversity observed in *C. elegans* which was caused by recent migration patterns and strong selective sweeps, might be linked to human dispersal within the few last centuries (Andersen et al., 2012; Phillips, 2006). This scenario is unlikely for the beetle associated *P. pacificus*. Despite the differences, both species reveal strong evidence for linked selection shaping genetic diversity within their genomes. Thus, in addition with studies of self-fertilizing nematodes and plants (Cutter and Choi, 2010; Cutter et al., 2006; Kim et al., 2007; Nordborg et al., 2005), our presentation of diversity within *P. pacificus* highlights the need to study multiple satellite model systems in order to better understand common features and distinct patterns of genome evolution of self-fertilizing species. Finally, our catalogue of natural variation will form the basis for further studies associating phenotypic variability to genetic variation within *P. pacificus*.

**III. MATERIALS & METHODS**

### Genomic library preparation

For preparation of genomic DNA, the MasterPure DNA purification Kit from Epicentre was used, resulting in high yields of clean DNA. DNA was quantified by Qubit measurement and diluted with TE to 20 ng/µl in a total volume of 55 µl. Genomic libraries were generated using the TruSeq DNA Sample Preparation Kit / v2 from Illumina. DNA was sheared with the default settings using the Covaris S2 System. Following the protocol, end repair, adenylation and index adapter ligation were performed. After running samples on a 2% agarose gel for 90 minutes, gel slices ranging in size from 400-500 bp were excised resulting in an insert size of approximately 300-400 bp. After fragments were amplified by PCR, libraries were validated on a Bioanalyzer DNA 1000 chip. All libraries were diluted to a concentration of 10 nM in 0.1% EB-Tween and pooled to 1, 4, 8, and 16-plexes. For mate pair libraries, clean genomic DNA was prepared using the Genomic-tip 100/G Kit from Qiagen. DNA was quantified by Qubit measurement, diluted with Tris-HCl to 150 ng/µl in a total volume of 70 µl and was sheared with the given settings (using SC 13) with the Hydroshear. For *P. exspectatus*, 3 kb and 5 kb Mate Pair libraries were generated using the Mate Pair Library v2 Kit for 2-5 kb Libraries from Illumina.

### Genome assembly

We sequenced five 150bp paired-end libraries with insert sizes between 400-600bp of an inbred *P. exspectatus* strain (10 generations of full-sibling inbreeding) on the Illumina Genome Analyzer II. In order to quantify the degree of ambiguous base calls vs. remaining heterozygosity after the 10 rounds of full sibling inbreeding, we also sequenced one library of the non-inbred wild-type strain of *P. exspectatus*, which was not used for the *de novo* assembly. Reads were error corrected using quake with k-mer size 19 (version 0.2.2) (Kelley et al., 2010). The same quality masking and trimming procedure as for the genome sequencing data of 104 strains was applied. The assembly process was carried out in several steps. A preliminary assembly (k=49, N50=6.8 kb, assembled sequence=170.3Mb) using only paired end libraries was created with the SOAP de novo assembler (version 1.05) (Li et al., 2010b). The preliminary assembly was used to remove ligation sites within 150bp mate pair library reads by iterative alignment (bwa version 0.5.9-r16) (Li and Durbin, 2009) and 3’ truncation. From two libraries (3kb and 5kb), only pairs for which both ends could unambiguously be mapped to the preliminary assembly were kept. This method yielded 13,019,927 3kb mate pairs from a set of 25,685,920 raw pairs and 15,919,381 5kb mate pairs from a set of 31,112,885 raw pairs. We ran the SOAP assembler with paired-end libraries in the
assembly step and the mate-pair libraries as additional scaffolding information resulting in a second assembly (k=63). To further improve the contiguity of the genome assembly we screened a single end library of 100bp reads for spliced reads that showed non-overlapping alignments to one scaffold spanning potential introns or non-overlapping alignments to different scaffolds by blat (version 34) (KENT, 2002) which contain further scaffolding information. In the next iteration we used the all-paths-LG assembler (version 42069) (GNERRE et al., 2011) with paired-end reads, mate pairs, spliced reads, and one overlapping library of 20 million read pairs resampled from the previous assembly. A final scaffolding step including information from 2937 commercially obtained BACs spanning approximately 100kb and an iterative local assembly procedure to close intra-scaffold gaps was applied (BOETZER et al., 2011; LI et al., 2010a). A contamination check against 37 E. coli genomes as downloaded from Ensembl Bacteria (release 12) identified 6 scaffolds spanning 4.7Mb as contamination (blastn e-value< $10^{-30}$). The final assembly contained 167Mb of assembled sequence in 4,412 scaffolds spanning 177.6 Mb with an N50 value of 142kb. The largest scaffold encompasses 1,285.6 kb and genome-wide GC content is 43%. The final assembly of *P. exspectatus* was aligned against the Hybrid1 assembly of *P. pacificus* using the whole genome alignment tool mugsy (ANGIUOLI and SALZBERG, 2011).

**Gene annotation and orthology assignment**

Three mixed stage RNA libraries from *P. pacificus* were sequenced on an Illumina Genome Analyzer Ix (76bp paired-end): one conventional mRNA library, one SL1 enriched mRNA library and one SL2 enriched mRNA library (Sinha and Dieterich; manuscript in preparation). The three libraries were used to build a reference transcriptome with cufflinks/cuffmerge (version 1.3.0) (TRAPNELL et al., 2010). Open Reading Frames (ORFs) were derived from the inferred transcript sequences with FrameDP (GOUZY et al., 2009). We collected all complete coding sequences without frameshifts and used them as a training set for gene finding in *P. pacificus*. We mapped the assembled transcripts of *P. pacificus* genome-wide GC content is 43%. The final assembly of *P. exspectatus* was aligned against the Hybrid1 assembly of *P. pacificus* using the whole genome alignment tool mugsy (ANGIUOLI and SALZBERG, 2011).

Alignment and variant calling

Low quality bases in the first 36bp of raw reads with a quality below 20 (error probability≈1%) were masked and reads were trimmed at the first occurrence of a low quality (< 20) in the rest of the read. Reads were aligned to the Hybrid1 genome assembly of the *P. pacificus* PS312 strain (California) using stampy (version 1.0.12) (LUNTER and GOODSON, 2011). Duplicate reads were removed and reads were locally realigned using GATK (version 2.1-13). SNVs and small indels were called using samtools (version 0.1.18) (LI et al., 2009) excluding positions with > 100x coverage. We excluded all variant calls with a quality score below 20, coverage below two, short indels, and heterozygous or ambiguous positions with a consensus quality score FQ > 0 as defined by the samtools mpileup command. Large structural variations were called using cnv-seq (XIE and TAMMI, 2009). The previously published low coverage Sanger assembly of the *P. pacificus* PS1843 (Washington) strain (DIETERICH et al., 2008) was used to compare variant calls between the Sanger and Illumina platform. The Sanger assembly was aligned to the PS312 Hybrid1 reference genome using mugsy (version v1r2.2) (ANGIUOLI and SALZBERG, 2011). Based on homozygous 474,285 SNV positions that were covered by Sanger data, the genotyping accuracy was defined as the fraction of SNVs called by both platforms among all Illumina variant calls. For 938,544 SNVs obtained from the whole genome alignment of the PS1843 Sanger assembly, only 53% of variant calls agreed with the Illumina data. We attribute this high number of putative false positive SNVs to the low coverage (1x) of the Sanger data. 79% of Illumina-based indel calls were found to be in agreement with the Sanger data based on comparison of the equivalent indel region (KRAWITZ et al., 2010); however, 96% of Illumina-based indel calls overlapped with an Sanger-based indel in up to 10bp distance.
Population genetic analysis

With exception of the calculation of genetic diversity, all population genetic analysis was restricted to positions sufficiently covered in all analyzed strains such that confident variant calls could be made, i.e. coverage ≥ 2, samtools quality score ≥ 20 and no signal of heterozygosity (samtools consensus quality score FQ < 0). For visualization of phylogenetic relationships one million SNVs with genotypes in all strains were randomly selected and concatenated as input for SplitsTree4 (version 4.12.6) (HUSON and BRYANT, 2006). Principal component analysis was done with EIGENSOFT (version 3.0) (PATTERSON et al., 2006). Shared haplotype blocks were identified by the program GERMLINE (version 1-5-1) (GUSEV et al., 2009). As the majority of variable sites was not covered in all strains (Supplemental Fig. 7), for pairwise comparisons, we used all variant calls against the reference genome assembly (Hybrid1) to calculate the average number of substitutions. Genetic diversity π was calculated as the average number of substitutions between all pairwise comparisons for the strains of interest. π values were calculated in non-overlapping windows of 100kb such that each window contains between 100 and 1000 differences. Chromosome-wide plots of diversity were generated by concatenating Contigs (> 1Mb) with respect to marker positions on the genetic map (SRINIVASAN et al., 2003). Contigs that lack any genetic marker or that were shorter than 1Mb were excluded from these plots (a total of ~40% of the total assembly is excluded by these criteria). The signal was further smoothed with a running average with window size of 1Mb for ease of presentation.

Linkage disequilibrium

For analysis of linkage disequilibrium across all 104 strains and within clades A1, A2, and C, we used biallelic SNVs that could be reliably genotyped in all analyzed strains and with a minor allele frequency ≥ 5% in order to calculate $r^2$ values. With exception of clade C, the set of all strains as well as clade A1 and A2 strains show extensive background LD across chromosomes (Figure 5A and B), which reflects the strong geographic separation across most strains. To visualize patterns of LD within clade C in at distances larger than 100kb, we calculated mean $r^2$ values of ten randomly chosen SNV pairs for every pair of 100kb windows across the whole P. pacificus genome (Figure 5B).

Analysis of coding regions

The quality filtered variant calls from samtools were compared to the P. pacificus gene models using a custom C++ program designed for classification of indels and nucleotide substitutions in mutant resequencing projects (RAE et al., 2012). This allowed for the classification of SNV as synonymous, nonsynonymous, or noncoding, and for calculation of numbers of synonymous and non-synonymous substitutions for all pairwise comparisons of strains. In order to calculate $\delta_{s},\delta_{ns}$ for any pair of strains, the total number of substitutions across all genes were normalized by the total number of synonymous and nonsynonymous sites of all genes as calculated from the P. pacificus gene models.

Ancestral state inference and SFS

For the inference of ancestral states, homozygous positions sufficiently covered in all strains were implanted into the reference genome (Hybrid1) to generate hypothetical haplotypes with well supported SNVs. As outgroup, we used an implanted genome that contained all 5.2 million single nucleotide substitutions between the genome sequences of P. pacificus and P. exspectatus. In regions that did not align between the two species, no ancestral state was inferred. The inference was done by building a maximum likelihood tree using fasttree (PRICE et al., 2009) of all clade C strains and the P. exspectatus sequence in non-overlapping windows of 20kb (10kb and 50kb yield similar results). We inferred the most likely ancestral sequences of all nodes of the tree using a custom python script implementing a variant of the dynamic programming algorithm for probabilistic ancestral inference (PUPKo et al., 2000). For each of the resulting trees, the most likely sequence at the root of clade C was used as the ancestral sequence to polarize variation in clade C. Positions at which the ancestral state was inferred with less than 95% confidence were excluded. SFS were calculated separately for synonymous, nonsynonymous, and noncoding SNVs as well as for high and low diversity regions, defined by a non-overlapping 100kb windows across the P. pacificus genome.

A. Data availability

All reads were submitted to the NCBI Sequence Read Archive. Variant calls, the P. exspectatus genome assembly, and the gene models (version SNAP2012) are available at http://www.pristionchus.org/variome/.

References


