Characterization of Genetic Diversity in the Nematode *Pristionchus pacificus* from Population-Scale Resequencing Data

Christian Rödelsperger,* Richard A. Neher,† Andreas M. Weller,* Gabi Eberhardt,* Hanh Witte,* Werner E. Mayer,* Christoph Dieterich,‡ and Ralf J. Sommer*,[1]

*Department for Evolutionary Biology and †Evolutionary Dynamics and Biophysics Group, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany, and ‡Berlin Institute for Medical Systems Biology at the Max Delbrück Center for Molecular Medicine Berlin, 13125 Berlin, Germany

ABSTRACT 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 10 times more diverse than *C. elegans* and exhibits substantial population structure that allows us to probe its evolution on multiple timescales. Consistent with reduced effective recombination in this self-fertilizing species, we find haplotype blocks that span several megabases. Using the *P. exspectatus* genome as an 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 timescales 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 (Maynard Smith and Haigh 1974; Gillespie 2000). In addition, linkage between loci with nonneutral 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 draft—distinct from common genetic drift (Gillespie 2000). In addition to reducing neutral genetic diversity, draft 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 hitchhiking, 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 Payseur (2013) and Neher (2013).

doi: 10.1534/genetics.113.159855
Manuscript received November 17, 2013; accepted for publication January 16, 2014; published Early Online January 17, 2014.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.159855/-/DC1.
Sequence data from this article have been deposited under GenBank accession no. SRA057279.

[1]Corresponding author: Max Planck Institute for Developmental Biology, Spemannstrasse 37, BW 72076 Tübingen, Germany. E-mail: ralf.sommer@tuebingen.mpg.de

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* system and multiple comparative studies have revealed divergent patterns in vulva formation, dauer development, and feeding behavior (Sinha *et al.* 2012; Bumbarger *et al.* 2013; Kienle and Sommer 2013).

Similar to *C. elegans*, *P. pacificus* has six chromosomes, which were shown to be largely macrosyntenic to their *C. elegans* counterparts (Srinivasan *et al.* 2002; Lee *et al.* 2003). 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 from *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 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 rates (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 F_1 hybrids with *P. pacificus* (Kanzaki *et al.* 2012). Similar to that in the comparative studies of selfing and outcrossing species in plants (Wright *et al.* 2002; Foxe *et al.* 2009), 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 an 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.

Materials and 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 min, gel slices ranging in size from 400 to 500 bp were excised, resulting in an insert size of ~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% elution buffer Tween and pooled to 1-, 4-, 8-, and 12-plexes. For mate pair libraries, clean genomic DNA was prepared using the Genomic-tip 100/G Kit from QIAGEN (Valencia, CA). DNA was quantified by Qubit measurement, diluted with Tris–HCl to 150 ng/μl in a total volume of 70 μl, and sheared with the given settings (using SC 13) with the Hydroshear. For *P. exspectatus*, 3- and 5-kb Mate Pair libraries were generated using the Mate Pair Library v2 Kit for 2- to 5-kb libraries from Illumina.

Genome assembly

We sequenced five 150-bp paired-end libraries with insert sizes between 400 and 600 bp of an inbred *P. exspectatus* strain (10 generations of full-sibling inbreeding) on the Illumina Genome Analyzer II. 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 noninbred 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 = 170.3 Mb) 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 150-bp mate pair library reads by iterative alignment (bwa version 0.5.9-r16) (Li and Durbin 2009) and 3′ truncation. From two libraries (3 and 5 kb), only pairs for which both ends could unambiguously be mapped to the preliminary assembly were kept. This method yielded 13,019,927 3-kb mate pairs from a set of 25,685,920 raw pairs and 15,919,381 5-kb 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 RNA library of 100-bp reads for spliced reads that showed nonoverlapping alignments to one scaffold spanning potential introns or nonoverlapping alignments to different scaffolds by blat (version 34) (Kent 2002), which contain further scaffolding information. In the next iteration we used the allpath-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 ~100 kb and an iterative local assembly procedure to close intrascaffold gaps was applied (Li *et al.* 2010a; Boetzer *et al.* 2011). A contamination check against 37 *Escherichia coli* genomes downloaded from Ensembl Bacteria (release 12) identified 6 scaffolds spanning 4.7 Mb as contamination (blastn *e*-value $<10^{-30}$). The final assembly contained 167 Mb of assembled sequence in 4412 scaffolds spanning 177.6 Mb with an N50 value of 142 kb. The largest scaffold encompasses 1285.6 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 IIx (76-bp paired end): one conventional mRNA library, one SL1 enriched mRNA library, and one SL2 enriched mRNA library (A. Sinha and C. Dieterich, unpublished data). 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* to the *P. exspectatus* genome, using BLAT (Kent 2002). The resulting gene structures were used to guide the spliced alignment of single-end 100-nt RNA-seq reads from a mixed-stage *P. exspectatus* library with cufflinks. The splice donor and acceptor positions were trained for iterative gene finder training, using the SNAP gene finder (Korf 2004). The new model predicts both *trans*-spliced genes and normal genes. All available splice junction data were provided as external evidence in the gene prediction step. For *P. exspectatus*, the newly trained gene model of *P. pacificus* was used to create an initial set of candidate genes with the help of splice site information from the *P. exspectatus* genome annotation. This gene set was used to retrain the gene model to adjust it to the new target species. We predicted the final gene set by rerunning the gene finder with the new gene model and all available splice site data. This procedure yielded 28,666 *P. pacificus* gene models and 24,642 *P. exspectatus* gene models. One-to-one orthologous pairs between *P. pacificus* and *P. exspectatus* were defined by first running the program CYNTENATOR (Rödelsperger and Dieterich 2010) to compute a set of unambiguously assignable conserved gene orders that include at least two conserved genes per species and are unique with respect to the criteria that every gene can occur at most once in a gene order alignment. Subsequently, best-reciprocal hits were identified locally within individual gene order alignments and were complemented by global best-reciprocal hits, computed without synteny information.

Alignment and variant calling

Low-quality bases in the first 36 bp of raw reads with a quality <20 (error probability = 1%) were masked and reads were trimmed at the first occurrence of a low-quality (<20) base 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). Single-nucleotide variants (SNVs) and small indels were called using samtools (version 0.1.18) (Li *et al.* 2009), excluding positions with >100× coverage. We excluded all variant calls with a quality score <20, coverage <2, 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 platforms. The Sanger assembly was aligned to the PS312 Hybrid1 reference genome, using mugsy (version v1r2.2) (Angiuoli and Salzberg 2011). Based on 474,285 homozygous 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 (1×) of the Sanger data. Seventy-nine percent 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 a Sanger-based indel in up to 10 bp distance.

Population genetic analysis

With the 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 1 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 were not covered in all strains (Supporting Information, Figure S7), 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 nonoverlapping windows of 100 kb such that each window contains between 100 and 1000 differences. Chromosome-wide plots of diversity were generated by concatenating Contigs (>1 Mb) with respect to marker positions on the genetic map (Srinivasan *et al.* 2003). Contigs that lack any genetic marker or that were <1 Mb 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 1 Mb for ease of presentation.

Linkage disequilibrium

For analysis of linkage disequilibrium across all 104 strains and within clades A_1, A_2, and C, we used biallelic SNVs that could be reliably genotyped in all analyzed strains and with a minor allele frequency ≥5% to calculate r^2 values. With the exception of clade C, the set of all strains as well as clade A_1 and A_2 strains show extensive background linkage disequilibrium (LD) across chromosomes (Figure 5, A and B), which reflects the strong geographic separation across most strains. To visualize patterns of LD within clade C at distances >100 kb, we calculated mean r^2 values of 10 randomly chosen SNV pairs for every pair of 100-kb 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 an SNV as synonymous, nonsynonymous, or noncoding and for calculation of numbers of synonymous and nonsynonymous substitutions for all pairwise comparisons of strains. To calculate δ_{si}, δ_{ns} for any pair of strains, the total number of substitutions across all genes was normalized by the total number of synonymous and nonsynonymous sites of all genes calculated from the *P. pacificus* gene models.

Ancestral state inference and site frequency spectrum

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 an 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 nonoverlapping windows of 20 kb (10 kb and 50 kb 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 <95% confidence were excluded. Site frequency spectra (SFS) were calculated separately for synonymous, nonsynonymous, and noncoding SNVs as well as for high- and low-diversity regions, defined by a nonoverlapping 100-kb window across the *P. pacificus* genome.

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/.

Results

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 *Materials and Methods* for details). The genome draft contains 167 Mb of assembled sequence in 4412 scaffolds spanning a total length of 177.6 Mb with an N50 value of 142 kb (Table 1). Realignment of the sequencing data placed 96–97% of reads onto the assembly, and 95% of mapped pairs were in correct orientation.

Despite full-sibling inbreeding for 10 generations we observed ~120,000 positions with ambiguous base calls from the realignments of sequencing data of the inbred *P. exspectatus* strain. To test whether these positions represent remaining heterozygosity not eliminated by inbreeding, we sequenced the noninbred *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 $h_t \approx 1.17h_0 \times (0.809)^t$, where h_0 denotes the initial heterozygosity and t is the number of inbred generations (Naglyaki 1992). After 10 generations, the expected heterozygosity h_{10} should be 14%—much lower then 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 ~60% of all ambiguous positions, spanning ~15% of the genome assembly (Figure S1). In further support for unresolved repetitive and recently duplicated regions as a potential source for the ambiguous positions, we even found triallelic positions (Figure S1). In contrast to a previous analysis of assemblies of outcrossing nematodes (Barrière

Table 1 Features of the *P. pacificus* and *P. exspectatus* genome sequences

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

Gene features are given as median and IQR.

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 ~200 Mb. Whole-genome alignments between *P. pacificus* and *P. exspectatus* covered 79.2 and 78.9 Mb of uniquely alignable sequence, respectively, and revealed 5.2 million substitutions and 1.2 million indels, indicating a sequence divergence of ~10% that is distributed uniformly across the chromosomes.

Using RNA-seq-derived gene models as a 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 those 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 >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 an outgroup for population genetic analysis of *P. pacificus*. A detailed comparison to *P. pacificus* and another closely related species, *P. arcanus*, will be presented elsewhere.

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) (Figure 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 10 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 SNVs relative to the reference genome (Table S1). 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). Of 470,000 SNV calls that are covered by both platforms, 98% were in agreement, and indel calls showed an agreement of 79–96% (see *Materials and Methods*).

Despite extensive inbreeding (at least 10 generations), ~6% (median, IQR: 5.2–7.3%) of SNVs for all strains were called as heterozygous (Table S1). 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 >2 kb (Table S1), 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 >30 generations shows similar levels of heterozygosity to those of the other strains. Third, the isolate RS5410 from La Réunion (Figure 1B), which was previously identified as an admixed strain (Morgan *et al.* 2012) and should therefore show a higher level of heterozygosity, exhibits only 4.4% of ambiguous variant calls, of which 27% fall into large duplicated regions. 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 *Materials and Methods*).

As mentioned above, we detected structural variants and copy number variations between 2 kb and up to 1 Mb, using a method [cnv-seq (Xie and Tammi 2009)] that compares differences in read depth relative to resequencing data of the reference strain (Figure S2). 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 were predicted as deleted and duplicated, respectively. We verified predicted deletions with PCR amplification experiments for three cellulase genes in 24 strains and found perfect agreement (Mayer *et al.* 2011).

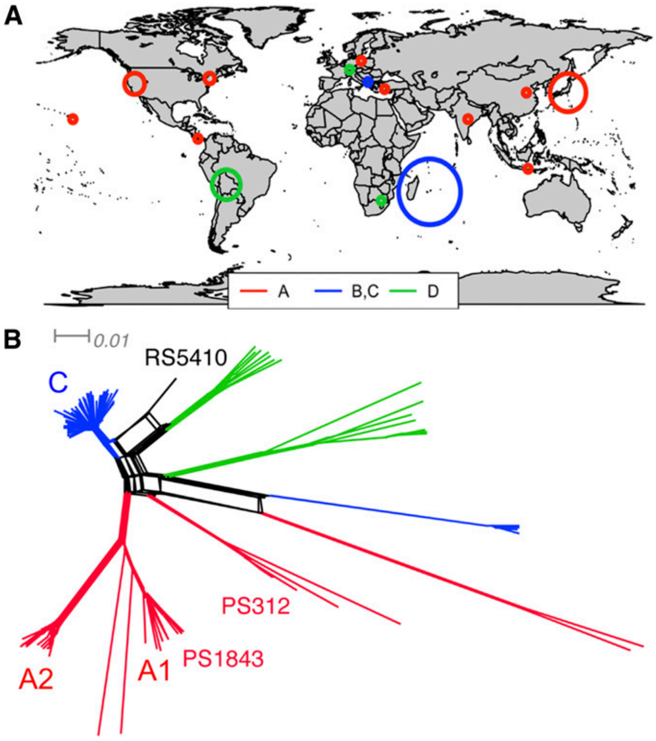

Figure 1 (A) Worldwide sampling of *P. pacificus* strains. Circle sizes indicate the numbers of sequenced strains. Colors denote the predominant mitochondrial and microsatellite clade per region (Morgan *et al.* 2012). (B) Population structure of *P. pacificus* natural isolates visualized as a split network based on average number of substitutions within 1 million variable sites that were genotyped in all strains. The three most deeply sampled clades are used for further analysis. Clade C (*N* = 44) is almost exclusively sampled from la Réunion, A_1 (*N* = 15) contains strains from North America and Asia, and A_2 (*N* = 16) consists of South and Central American strains as well as strains from the Indian Ocean. Colors indicate the predominant mitochondrial and microsatellite clade per branch (Morgan *et al.* 2012).

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 A_1, A_2, 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 Figure 1B (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 (Figure S3). The clades A_1, A_2, 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 timescales. 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 and represent genetic diversity. Comparisons between clades have features of diversity and divergence.

Figure 2 shows the nucleotide diversity within clades A_2 and C (see Figure S4 for comparisons including clade A_1), 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 intraclade diversity in 100-kb windows is 1.9 and 2.2 $\times 10^{-3}$ for clades A_2 and C, respectively. Strains from clade A_2 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 A_2 and C in 100-kb windows). Similarly, distance between the clades is correlated with diversity within clades C and A_2 (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 $\log_{10}$ distance values in 100-kb windows is 0.21 as opposed to ~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 <20% (windows of 100 kb, standard deviation of $\log_{10} d = 0.07$, Figure 2).

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 (Maynard Smith and Haigh 1974; Hudson and Kaplan 1995; Nordborg *et al.* 1996). Even though the diversity patterns in different clades are very similar, most intraclade diversity is due to private variation, rather than diversity shared in the common ancestral population (3% and 7% of polymorphic positions in clades C and A_2 are polymorphic in both; the joint SFS of different clades are shown in Figure S5). 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 a similar manner 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

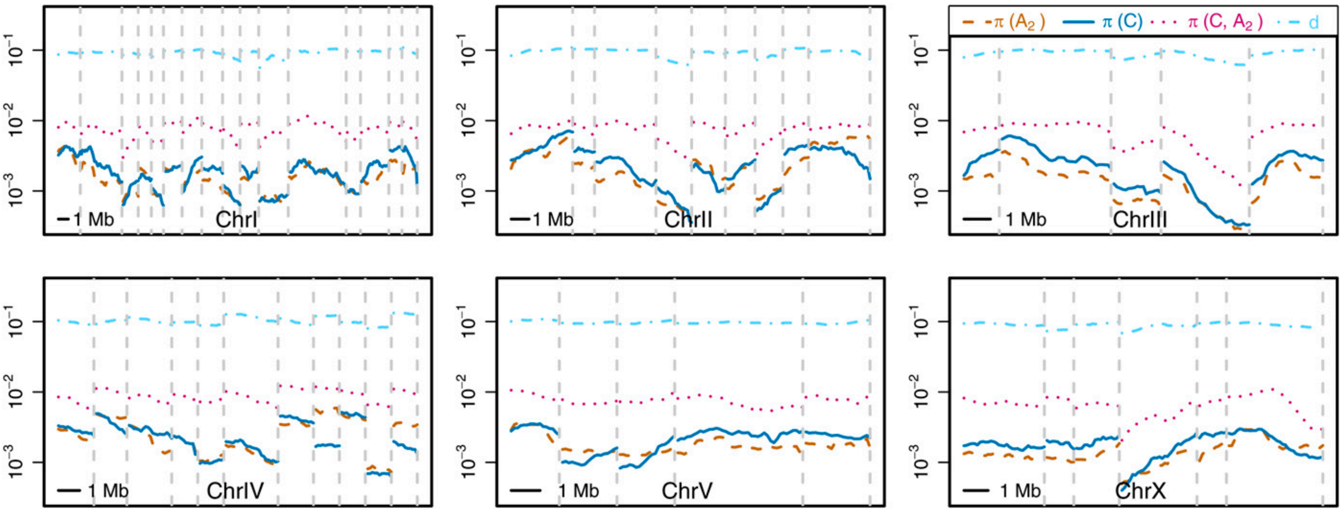

Figure 2 Genome-wide divergence (*d*) to the sister species *P. exspectatus* and nucleotide diversity π within and between clades C and A_2 are plotted along the chromosomes. π values were calculated in 100-kb windows and smoothed with a running average of 1 Mb for ease of presentation. Supercontigs (>1 Mb) with markers on the genetic map (Srinivasan *et al.* 2003) were concatenated to visualize the chromosomal distribution. The dashed lines denote supercontig boundaries with unknown physical distance. Diversity within different clades is very well correlated, suggesting that similar population genetic factors shape this variation.

of global selective sweeps seems unlikely, as we do not find long haplotypes that are shared among clades (Figure S6). 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 that of the present-day within-clade diversity. Together with homogeneous accumulation of divergence, we expect the interclade distance to be correlated with the intraclade diversity, but with a smaller relative variability. Both of these expectations are consistent with our observations.

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 $p_N/p_S \approx 0.32$ suggests widespread purifying selection that results in pruning of two of three nonsynonymous mutations. Note, however, that these genome-wide measures of synonymous and nonsynonymous diversity are dominated by interclade comparisons. Restricted to clade C, we find an average p_N/p_S ratio of 0.39, which indicates that a substantial fraction of nonsynonymous mutations were pruned only at the interclade level and segregate within clades.

To quantify this dependence of purifying selection on the timescale of separation, we calculated the density of synonymous differences, δ_{si}, and nonsynonymous differences, δ_{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 δ_{ns}/δ_{si} rather than p_N/p_S or d_N/d_S. Figure 3 shows the dependence of this ratio δ_{ns}/δ_{si} on the separation measured by the synonymous distance δ_{si} (stochastic fluctuations in δ_{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 δ_{ns}/δ_{si} of up to 0.5. With increasing distance, δ_{ns}/δ_{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 $d_N/d_S = 0.14$ for the comparison of 1:1 orthologs betweeen *P. pacificus* and *P. exspectatus*. Note that as we go from comparisons of strains within one clade to interclade comparisons and eventually to interspecies comparisons, the interpretation of δ_{ns}/δ_{si} gradually changes from p_N/p_S to d_N/d_S. For a more in-depth discussion of time-dependent d_N/d_S, see Rocha *et al.* (2006) and Mugal *et al.* (2014).

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 timescales and segregate as high-frequency polymorphisms. Below, we complement these observations for genome-wide ratios of nonsynonymous and synonymous differences with an analysis of the 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 (Srinivasan *et al.* 2003) has insufficient resolution to be

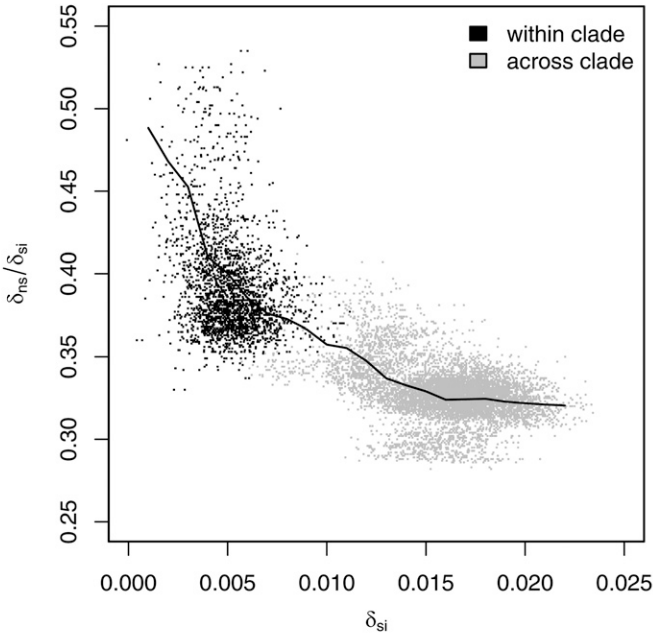

Figure 3 Individual dots represent δ_{ns}/δ_{si} ratios of all pairwise comparisons of the 104 *P. pacificus* strains and were plotted across different timescales as measured in δ_{si}. The line represents a smoothed average δ_{ns}/δ_{si} ratio of all measurements at a given δ_{si}. δ_{ns}/δ_{si} ratios decrease with distance (δ_{si}) from 0.5 to 0.3, suggesting that ~50% of nonsynonymous mutations have been selected against at very short time periods and were not observed as variable. In addition, ~20% of nonsynonymous variations are pruned with increasing distance. Assuming that this drop is due to purifying selection, we conclude that ~20% of nonsynonymous variation are weakly deleterious and segregate as high-frequency polymorphisms.

useful for this purpose. While the observed correlation on its own could also be the result of recurrent hitchhiking in gene-rich regions, the strong correlation of the diversity patterns across clades suggests purifying selection as a likely cause; see discussion above.

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 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 20-kb 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 20 kb 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 *Materials and Methods*). Variations of this strategy (different block length, different substitution models, different tree-building algorithms, and 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 mutations 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. Figure 4A shows the ratio of nonsynonymous to synonymous mutations at frequencies above a threshold ν. 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 $\nu \rightarrow 1$.

Figure 4B shows the polarized SFS of clade C on a double-logarithmic scale such that power laws show as straight lines. At frequencies <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% and 40%, the slope of the SFS is compatible with $1/k^2$, indicated as a dashed line in Figure 4B. The $1/k^2$ behavior and the nonmonotonicity are 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 Shraiman 2011; Neher and Hallatschek 2013).

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 (Figure 4C). The relative abundance of synonymous polymorphisms increases only by a factor of 2 as we go from low to high frequency (see Figure 4A). This similarity suggests that the dynamics of polymorphisms are 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 >0.6, consistent with the gradual manifestation of purifying selection on old and frequent alleles ($P < 10^{-6}$, Fisher's exact test, Figure 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 (Figure 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 Figure 3), we used the Web server DFE-α (distribution of fitness effects) (Keightley and Eyre-Walker 2007) to infer parameters of the

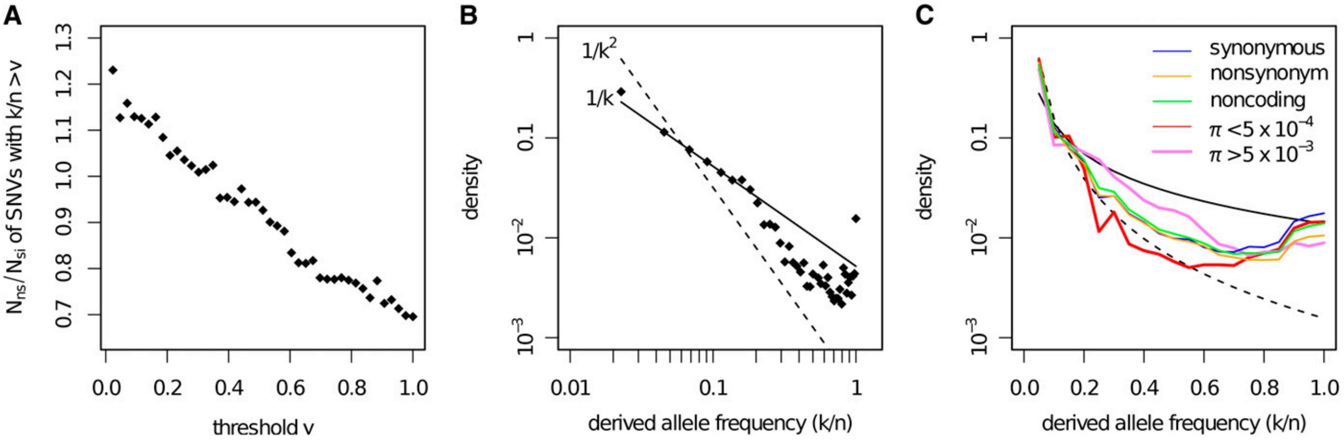

Figure 4 (A) Monotonic decrease of the ratio of the numbers of nonsynonymous (N_{ns}) and synonymous (N_{si}) SNVs at derived allele frequencies (*k*/*n*) above a threshold ν. The ratios represent unnormalized ratios of raw counts of nonsynonymous and synonymous SNVs in the site frequency spectrum (SFS). 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 that would manifest as an uptick at values of $\nu \rightarrow 1$. (B) SFS of clade C is shown on a double-logarithmic plot. If neutral diversity is shaped by constant genetic drift, we expect that the number of derived alleles present in *k* of *n* strains decays as 1/*k*, indicated as straight solid line. The dashed line indicates the corresponding expectation, $1/k^2$, under a genetic draft model (for rare alleles). At frequencies <10%, the SFS is proportional to 1/*k*, and at intermediate frequencies between 10% and 50%, the SFS is steeper and compatible with $1/k^2$. (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 SFS. Only at frequencies above $k/n = 0.6$ do we see a systematic excess of synonymous over nonsynonymous mutations ($P < 10^{-6}$, Fisher's exact test).

DFE, using the synonymous and nonsynonymous SFS within clade C. DFE-α preferred a model with a 4.5-fold population size increase ($N_2 = 4.5N_1$) $\sim N_2$ generations in the past over a model with a single population size (log-likelihood difference 325). For the two-epoch model, DFE-α estimated that 45% of nonsynonymous mutations have deleterious effects in excess of $N_2|s| = 100$, consistent with our previous results. Another 37% of nonsynonymous mutations are estimated to be approximately neutral with $N_2|s| < 1$, which is compatible with the observed δ_{ns}/δ_{si} at larger timescales (Figure 3).

The shape of the SFS is sensitive to past demography, direct selection, linked selection, and genetic drift. Disentangling the effects of these process 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 noncoding 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 $1/k^2$ decay of the SFS at low and intermediate frequencies (Neher and Shraiman 2011). The SFS we observe, however, are nonmonotonous and it is impossible to obtain nonmonotonous 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 nonmonotonous 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 (Figure 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 (Walczak *et al.* 2012; Neher and Hallatschek 2013). 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 a subject of future investigation.

Linkage disequilibrium and haplotype structure

Finally, we investigated the pattern of LD in *P. pacificus*. Figure 5, A and B, shows the average LD between pairs of loci measured as r^2 as a function of the physical separation between the loci for the global sample of 104 strains and clades C, A_1, and A_2. In all cases, LD drops over a distance of ~20 kb (which is the scale used to infer the ancestral state above). With exception of clade C, all data sets show substantial LD even across chromosomes ($r^2 = 0.12-0.18$), which we attribute to the strong geographic separation of strains even within clades (see Table S1). For the more closely related and geographically restricted clade C strains, we find very little LD across chromosomes ($r^2 = 0.04$), but blocks of local LD, the largest of which spans several megabases (Figure 5C). The lack of LD across chromosomes supports our assertion that clade C is the only well-mixed population suitable for detailed population genetic analysis. A complementary analysis of haplotype blocks (see *Materials and Methods*) also revealed shared haplotype blocks of megabase size between clade C strains (Figure S6). While long haplotype blocks conserved across continents were found in *C. elegans* (Andersen *et al.* 2012), most of shared haplotype blocks in *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 *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 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).

Discussion

By sequencing 104 strains of *P. pacificus* and a closely related outgroup, *P. exspectatus*, we have shown that *P. pacificus* contains extensive population structure and overall levels of diversity are one order of magnitude higher than in *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 *P. pacificus* and its sister species *P. exspectatus* is homogenous across the genome. The latter observation is consistent with results of the accompanying article in this issue (Weller *et al.* 2014), which presents a mutation accumulation experiment that shows that the *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. 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. Fifty percent 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 timescales on the order of the separation between clades, which corresponds to ~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 100-kb windows, we conclude that background selection plays an important role in shaping *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 genealogies manifests itself in an SFS with a steeper decay and an uptick at high derived allele frequencies, both of which are characteristic signatures of genetic draft or linked selection (Neher and Shraiman 2011; Neher and Hallatschek 2013). 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 (Rockman and Kruglyak 2009; Cutter and Choi 2010; Andersen *et al.* 2012). 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 hitchhiking (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 population and a close outgroup has allowed us to study evolution on a variety of timescales spanning distances from 0.1 to 10% and revealed a consistent decreasing trend in the ratio of nonsynonymous to synonymous mutations (Rocha *et al.* 2006; Mugal *et al.* 2014). This trend implies an abundance of weakly deleterious mutations—at least when their effect is averaged over larger and larger timescales. 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* that was caused by recent migration patterns and strong selective sweeps might be linked to human dispersal within the last few centuries (Phillips 2006; Andersen *et al.* 2012). 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 to studies of self-fertilizing nematodes and plants (Nordborg *et al.* 2005; Cutter *et al.* 2006; Kim *et al.* 2007; Cutter and Choi 2010), our presentation of diversity within *P. pacificus* highlights the need to study multiple satellite model systems to better understand common features and distinct patterns of genome evolution

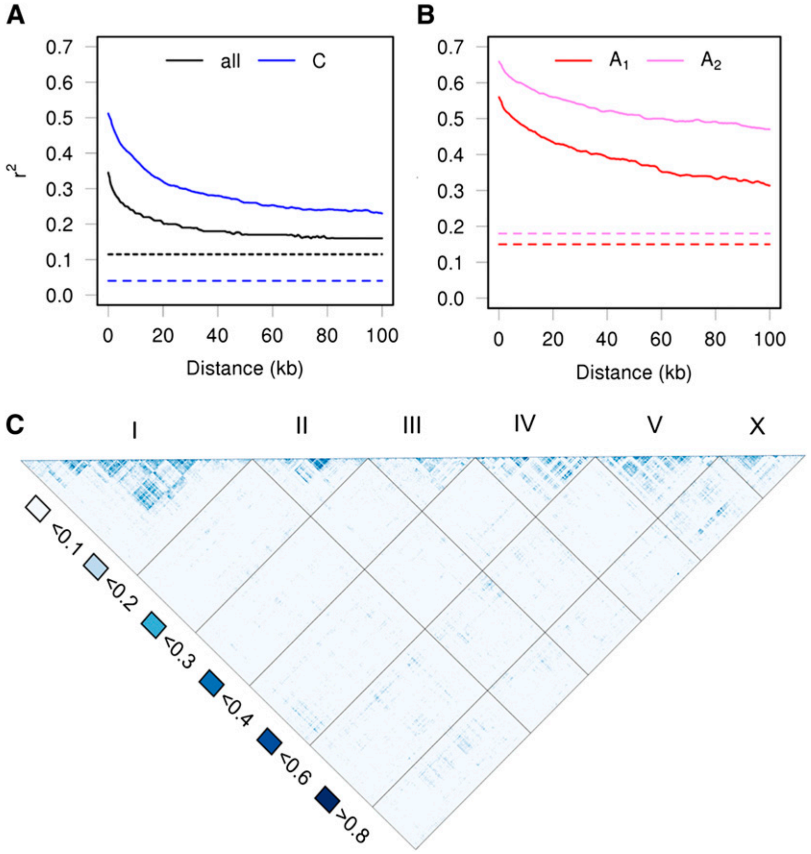

Figure 5 (A) Average r^2 values between pairs of loci for all 104 strains and clade C. LD drops over the first 20 kb but stays at a relatively high level across longer distances (solid lines) and even across chromosomes (dashed lines). We attribute this high level of background LD to the strong differentiation between subclades as a result of geographic separation. Restricted to clade C, local LD is higher, while interchromosome LD is low. (B) Average r^2 values for clades A_1 and A_2. Unlike clade C, both clades show substantial background LD, which can also be explained by geographic separation within these clades; *i.e.*, individual strains within the same clade were collected on different continents. (C) Average LD between all 100-kb windows for clade C across the genome was calculated as average r^2 for 10 pairs of biallelic SNVs with 5–95% allele frequency. LD across chromosomes is virtually absent, indicating the effective reshuffling of chromosomes in outcrossing events. However, clade C shows several blocks spanning megabases in strong LD on most chromosomes.

of self-fertilizing species. Finally, our catalog of natural variation will form the basis for further studies associating phenotypic variability to genetic variation within *P. pacificus*.

Acknowledgments

We thank Dr. Cameron Weadick for critically reading the manuscript. This project was funded by the Max-Planck Society.

Literature Cited

Andersen, E. C., J. P. Gerke, J. A. Shapiro, J. R. Crissman, R. Ghosh *et al.*, 2012 Chromosome-scale selective sweeps shape *Caenorhabditis elegans* genomic diversity. Nat. Genet. 44: 285–290.

Andolfatto, P., 2007 Hitchhiking effects of recurrent beneficial amino acid substitutions in the *Drosophila melanogaster* genome. Genome Res. 17: 1755–1762.

Angiuoli, S. V., and S. L. Salzberg, 2011 Mugsy: fast multiple alignment of closely related whole genomes. Bioinformatics 27: 334–342.

Barrière, A., S.-P. Yang, E. Pekarek, C. G. Thomas, E. S. Haag *et al.*, 2009 Detecting heterozygosity in shotgun genome assemblies: lessons from obligately outcrossing nematodes. Genome Res. 19: 470–480.

Barton, N. H., and B. Charlesworth, 1998 Why sex and recombination? Science 281: 1986–1990.

Begun, D. J., and C. F. Aquadro, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster.* Nature 356: 519–520.

Begun, D. J., A. K. Holloway, K. Stevens, L. W. Hillier, Y.-P. Poh *et al.*, 2007 Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLoS Biol. 5: 2534–2559.

Boetzer, M., C. V. Henkel, H. J. Jansen, D. Butler, and W. Pirovano, 2011 Scaffolding pre-assembled contigs using sspace. Bioinformatics 27: 578–579.

Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley, and W. Stephan, 1995 The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 140: 783–796.

Bumbarger, D. J., M. Riebesell, C. Rödelsperger, and R. J. Sommer, 2013 System-wide rewiring underlies behavioral differences in predatory and bacterial-feeding nematodes. Cell 152: 109–119.

Cao, J., K. Schneeberger, S. Ossowski, T. Günther, S. Bender *et al.*, 2011 Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat. Genet. 43: 956–963.

Charlesworth, J., and A. Eyre-Walker, 2008 The McDonald-Kreitman test and slightly deleterious mutations. Mol. Biol. Evol. 25: 1007–1015.

Charlesworth, B., M. T. Morgan, and D. Charlesworth, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics 134: 1289–1303.

Cutter, A. D., and J. Y. Choi, 2010 Natural selection shapes nucleotide polymorphism across the genome of the nematode *Caenorhabditis briggsae*. Genome Res. 20: 1103–1111.

Cutter, A. D., and B. A. Payseur, 2013 Genomic signatures of selection at linked sites: unifying the disparity among species. Nat. Rev. Genet. 14: 262–274.

Cutter, A. D., M.-A. Félix, A. Barrière, and D. Charlesworth, 2006 Patterns of nucleotide polymorphism distinguish temperate and tropical wild isolates of *Caenorhabditis briggsae*. Genetics 173: 2021–2031.

Denver, D. R., K. A. Clark, and M. J. Raboin, 2011 Reproductive mode evolution in nematodes: insights from molecular phylogenies and recently discovered species. Mol. Phylogenet. Evol. 61: 584–592.

Dieterich, C., S. W. Clifton, L. N. Schuster, A. Chinwalla, K. Delehaunty *et al.*, 2008 The *Pristionchus pacificus* genome provides a unique perspective on nematode lifestyle and parasitism. Nat. Genet. 40: 1193–1198.

Foxe, J. P., T. Slotte, E. A. Stahl, B. Neuffer, H. Hurka *et al.*, 2009 Recent speciation associated with the evolution of selfing in *Capsella*. Proc. Natl. Acad. Sci. USA 106: 5241–5245.

Gillespie, J. H., 2000 Genetic drift in an infinite population. The pseudohitchhiking model. Genetics 155: 909–919.

Gnerre, S., I. Maccallum, D. Przybylski, F. J. Ribeiro, J. N. Burton *et al.*, 2011 High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. USA 108: 1513–1518.

Gouzy, J., S. Carrere, and T. Schiex, 2009 FrameDP: sensitive peptide detection on noisy matured sequences. Bioinformatics 25: 670–671.

Gusev, A., J. K. Lowe, M. Stoffel, M. J. Daly, D. Altshuler *et al.*, 2009 Whole population, genome-wide mapping of hidden relatedness. Genome Res. 19: 318–326.

Hill, W. G., and A. Robertson, 1966 The effect of linkage on limits to artificial selection. Genet. Res. 8: 269–294.

Hudson, R. R., and N. L. Kaplan, 1995 Deleterious background selection with recombination. Genetics 141: 1605–1617.

Huson, D. H., and D. Bryant, 2006 Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23: 254–267.

Kanzaki, N., E. J. Ragsdale, M. Herrmann, W. E. Mayer, and R. J. Sommer, 2012 Description of three *Pristionchus* species (nematoda: Diplogastridae) from Japan that form a cryptic species complex with the model organism *P. pacificus*. Zoolog. Sci. 29: 403–417.

Kawecki, T. J., and D. Ebert, 2004 Conceptual issues in local adaptation. Ecol. Lett. 7: 1225–1241.

Keightley, P. D., and A. Eyre-Walker, 2007 Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics 177: 2251–2261.

Kelley, D. R., M. C. Schatz, and S. L. Salzberg, 2010 Quake: quality-aware detection and correction of sequencing errors. Genome Biol. 11: R116.

Kent, W. J., 2002 Blat—the blast-like alignment tool. Genome Res. 12: 656–664.

Kienle, S., and R. J. Sommer, 2013 Cryptic variation in vulva development by cis-regulatory evolution of a hairy-binding site. Nat. Commun. 4: 1714.

Kim, S., V. Plagnol, T. T. Hu, C. Toomajian, R. M. Clark *et al.*, 2007 Recombination and linkage disequilibrium in *Arabidopsis thaliana*. Nat. Genet. 39: 1151–1155.

Kimura, M., and T. Ohta, 1973 The age of a neutral mutant persisting in a finite population. Genetics 75: 199–212.

Korf, I., 2004 Gene finding in novel genomes. BMC Bioinformatics 5: 59.

Krawitz, P., C. Rödelsperger, M. Jger, L. Jostins, S. Bauer *et al.*, 2010 Microindel detection in short-read sequence data. Bioinformatics 26: 722–729.

Lee, K.-Z., A. Eizinger, R. Nandakumar, S. C. Schuster, and R. J. Sommer, 2003 Limited microsynteny between the genomes of *Pristionchus pacificus* and *Caenorhabditis elegans*. Nucleic Acids Res. 31: 2553–2560.

Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics 25: 1754–1760.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The sequence alignment/map format and samtools. Bioinformatics 25: 2078–2079.

Li, R., W. Fan, G. Tian, H. Zhu, L. He *et al.*, 2010a The sequence and de novo assembly of the giant panda genome. Nature 463: 311–317.

Li, R., H. Zhu, J. Ruan, W. Qian, X. Fang *et al.*, 2010b De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 20: 265–272.

Lunter, G., and M. Goodson, 2011 Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. 21: 936–939.

Macpherson, J. M., G. Sella, J. C. Davis, and D. A. Petrov, 2007 Genomewide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in *Drosophila*. Genetics 177: 2083–2099.

Mayer, W. E., L. N. Schuster, G. Bartelmes, C. Dieterich, and R. J. Sommer, 2011 Horizontal gene transfer of microbial cellulases into nematode genomes is associated with functional assimilation and gene turnover. BMC Evol. Biol. 11: 13.

Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. 23: 23–35.

McDonald, J. H., and M. Kreitman, 1991 Adaptive protein evolution at the Adh locus in *Drosophila*. Nature 351: 652–654.

Messer, P. W., and D. A. Petrov, 2012 The McDonald-Kreitman test and its extensions under frequent adaptation: Problems and solutions. arXiv: 1211.0060.

Molnar, R. I., H. Witte, I. Dinkelacker, L. Villate, and R. J. Sommer, 2012 Tandem-repeat patterns and mutation rates in microsatellites of the nematode model organism *Pristionchus pacificus*. G3 2: 1027–1034.

Morgan, K., A. McGaughran, L. Villate, M. Herrmann, H. Witte *et al.*, 2012 Multi locus analysis of *Pristionchus pacificus* on La Réunion island reveals an evolutionary history shaped by multiple introductions, constrained dispersal events and rare out-crossing. Mol. Ecol. 21: 250–266.

Mugal, C. F., J. B. W. Wolf, and I. Kaj, 2014 Why time matters: codon evolution and the temporal dynamics of dN/dS. Mol. Biol. Evol. 31: 212–231.

Naglyaki, T., 1992 *Introduction to Theoretical Population Genetics*. Springer New York, USA.

Neher, R. A., 2013 Genetic draft, selective interference, and population genetics of rapid adaptation. Annu. Rev. Ecol. Evol. Syst. 44: 195–215.

Neher, R. A., and O. Hallatschek, 2013 Genealogies of rapidly adapting populations. Proc. Natl. Acad. Sci. USA 110: 437–442.

Neher, R. A., and B. I. Shraiman, 2011 Genetic draft and quasi-neutrality in large facultatively sexual populations. Genetics 188: 975–996.

Neher, R. A., T. A. Kessinger, and B. I. Shraiman, 2013 Coalescence and genetic diversity in sexual populations under selection. Proc. Natl. Acad. Sci. USA 110: 15836–15841.

Nordborg, M., B. Charlesworth, and D. Charlesworth, 1996 The effect of recombination on background selection. Genet. Res. 67: 159–174.

Nordborg, M., T. T. Hu, Y. Ishino, J. Jhaveri, C. Toomajian *et al.*, 2005 The pattern of polymorphism in *Arabidopsis thaliana*. PLoS Biol. 3: e196.

Patterson, N., A. L. Price, and D. Reich, 2006 Population structure and eigenanalysis. PLoS Genet. 2: e190.

Phillips, P. C., 2006 One perfect worm. Trends Genet. 22: 405–407.

Price, M. N., P. S. Dehal, and A. P. Arkin, 2009 Fasttree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol. Biol. Evol. 26: 1641–1650.

Pupko, T., I. Pe, R. Shamir, and D. Graur, 2000 A fast algorithm for joint reconstruction of ancestral amino acid sequences. Mol. Biol. Evol. 17: 890–896.

Rae, R., H. Witte, C. Rödelsperger, and R. J. Sommer, 2012 The importance of being regular: *Caenorhabditis elegans* and *Pristionchus* *pacificus* defecation mutants are hypersusceptible to bacterial pathogens. Int. J. Parasitol. 42: 747–753.

Rocha, E. P., J. M. Smith, L. D. Hurst, M. T. Holden, J. E. Cooper *et al.*, 2006 Comparisons of dN/dS are time dependent for closely related bacterial genomes. J. Theor. Biol. 239: 226–235.

Rockman, M. V., and L. Kruglyak, 2009 Recombinational landscape and population genomics of C*aenorhabditis elegans*. PLoS Genet. 5: e1000419.

Rödelsperger, C., and C. Dieterich, 2010 Cyntenator: progressive gene order alignment of 17 vertebrate genomes. PLoS ONE 5: e8861.

Sargsyan, O., and J. Wakeley, 2008 A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. Theor. Popul. Biol. 74: 104–114.

Seger, J., W. Smith, J. Perry, J. Hunn, Z. Kaliszewska *et al.*, 2010 Gene genealogies strongly distorted by weakly interfering mutations in constant environments. Genetics 184: 529.

Sella, G., D. A. Petrov, M. Przeworski, and P. Andolfatto, 2009 Pervasive natural selection in the *Drosophila* genome? PLoS Genet. 5: e1000495.

Sinha, A., R. J. Sommer, and C. Dieterich, 2012 Divergent gene expression in the conserved dauer stage of the nematodes *Pristionchus pacificus* and *Caenorhabditis elegans*. BMC Genomics 13: 254.

Srinivasan, J., W. Sinz, C. Lanz, A. Brand, R. Nandakumar *et al.*, 2002 A bacterial artificial chromosome-based genetic linkage map of the nematode *Pristionchus pacificus*. Genetics 162: 129–134.

Srinivasan, J., W. Sinz, T. Jesse, L. Wiggers-Perebolte, K. Jansen *et al.*, 2003 An integrated physical and genetic map of the nematode *Pristionchus pacificus*. Mol. Genet. Genomics 269: 715–722.

Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan *et al.*, 2010 Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28: 511–515.

Wakeley, J., 2008 *Coalescent Theory.* Roberts & Co. Greenwood Village, Colorado, USA.

Walczak, A. M., L. E. Nicolaisen, J. B. Plotkin, and M. M. Desai, 2012 The structure of genealogies in the presence of purifying selection: a fitness-class coalescent. Genetics 190: 753–779.

Weller, A. M., C. Rödelsperger, G. Eberhardt, R. I. Molnar, and R. J. Sommer, 2014 Opposing forces of A/T-biased mutations and G/C-biased gene conversions shape the genome of the nematode *Pristionchus pacificus*. Genetics 196: 1145–1152.

Wright, S. I., B. Lauga, and D. Charlesworth, 2002 Rates and patterns of molecular evolution in inbred and outbred *Arabidopsis*. Mol. Biol. Evol. 19: 1407–1420.

Xie, C., and M. T. Tammi, 2009 Cnv-seq, a new method to detect copy number variation using high-throughput sequencing. BMC Bioinformatics 10: 80.

Communicating editor: D. Begun

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.159855/-/DC1

Characterization of Genetic Diversity in the Nematode *Pristionchus pacificus* from Population-Scale Resequencing Data

Christian Rödelsperger, Richard A. Neher, Andreas M. Weller, Gabi Eberhardt, Hanh Witte, Werner E. Mayer, Christoph Dieterich, and Ralf J. Sommer

DOI: 10.1534/genetics.113.159855

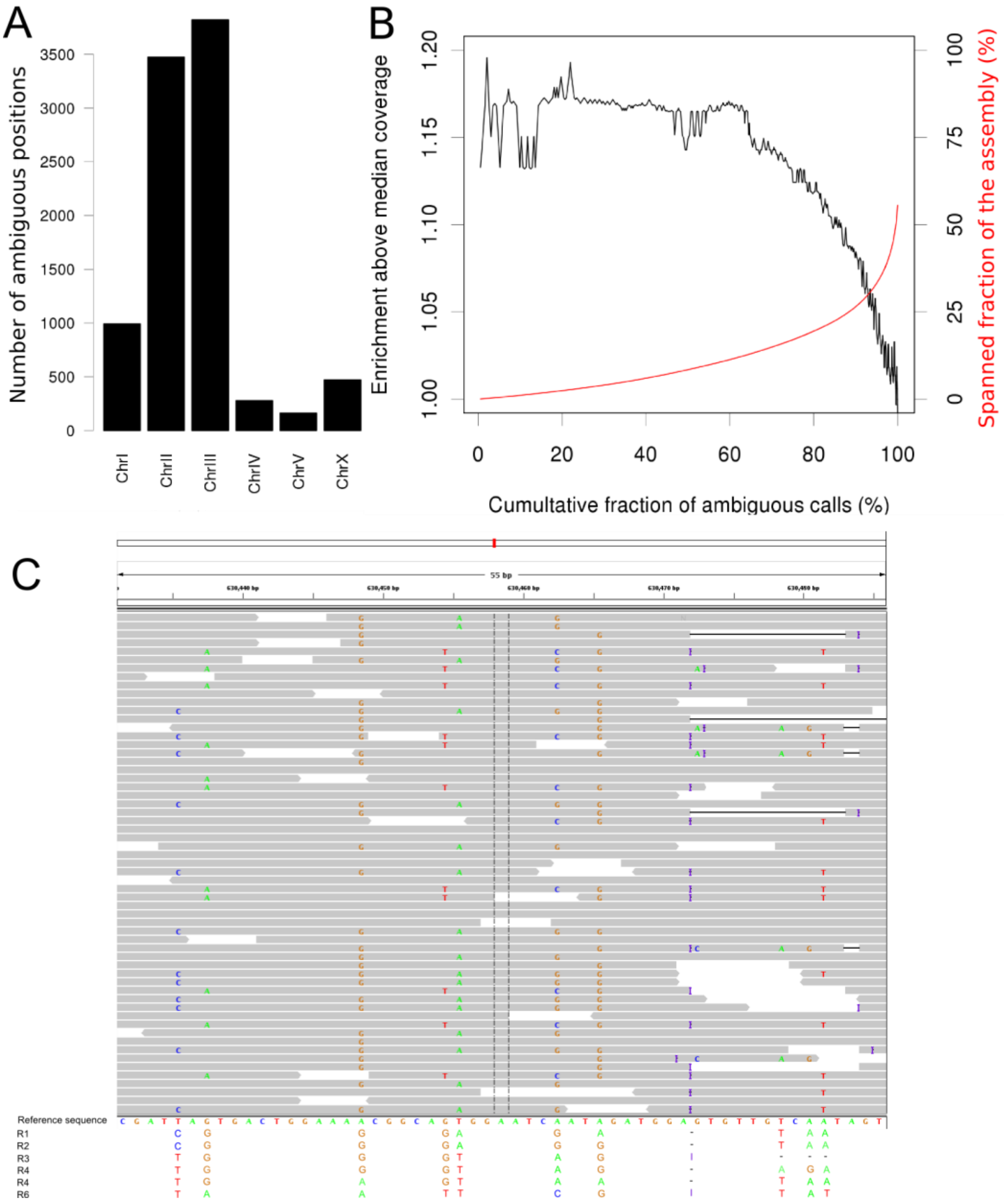

Figure S1 Heterozygosity in the *P. exspectus* assembly. **(A)** Ambiguous sites are present on autosomes and the X chromosome. Shown are the numbers of heterozygous calls per chromosome. *P. exspectatus.* Scaffolds were assigned to chromosomes by blasting *P. pacificus* genetic markers against the *P. exspectatus* assembly. **(B)** Ambiguous sites are associated with increased coverage. Non-overlapping windows of 10kb were ordered by decreasing number of ambiguous sites (x-axis) and the ratio of median coverage within the top X% above the genome-wide median coverage is plotted. At 60-70% of ambiguous sites, the ratio drops to again and reaches a value of one when approaching 100% of ambiguous sites. **(C)** Screenshot from the IGV browser showing realigned *P. exspectatus* reads. One position shows three different genotypes, which cannot be explained by remaining heterozygosity, indicating that reads must originate from highly similar paralogous regions. The alignments suggest that at least six different regions (R1-R6) are required to explain the observed pattern.

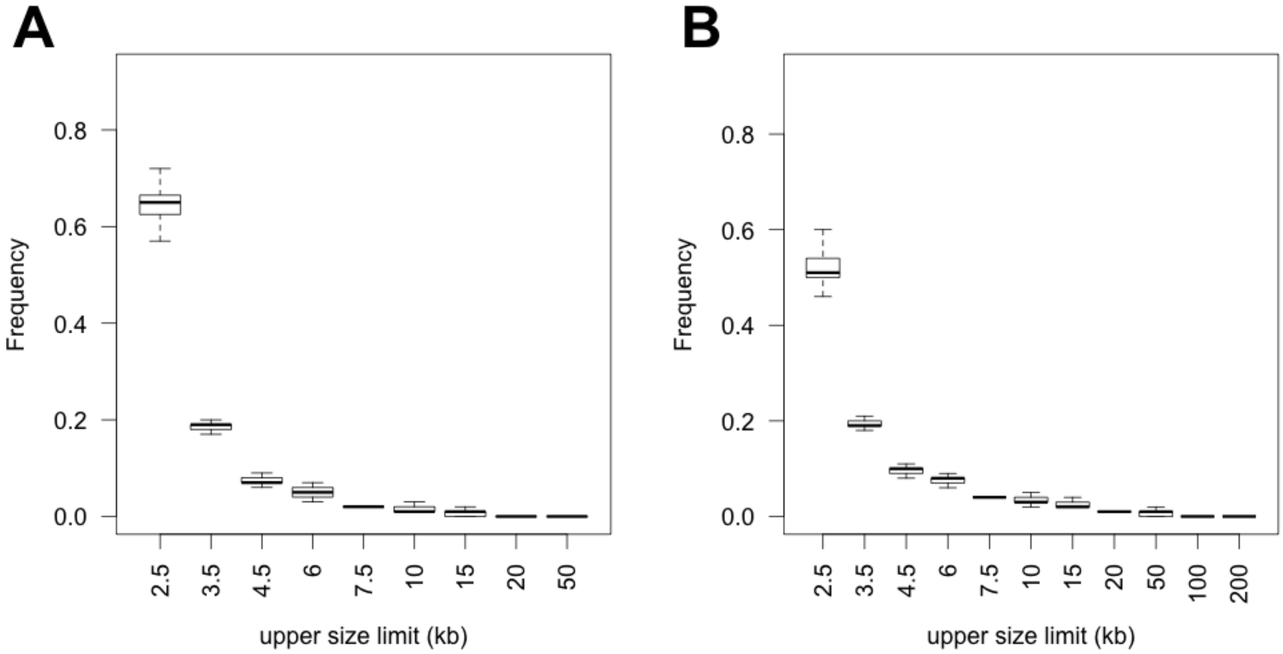

Figure S2 Size distribution of detected duplications and deletions. Duplications and deletions (≥ 2kb) were detected by comparing the read coverage in each of the 103 strains to the coverage of the resequenced reference strain (PS312) using the program cnv-seq (*P*<0.01). The boxplots show median and interquartile range of the size distribution of duplications **(A)** and deletions **(B)** across all 103 strains.

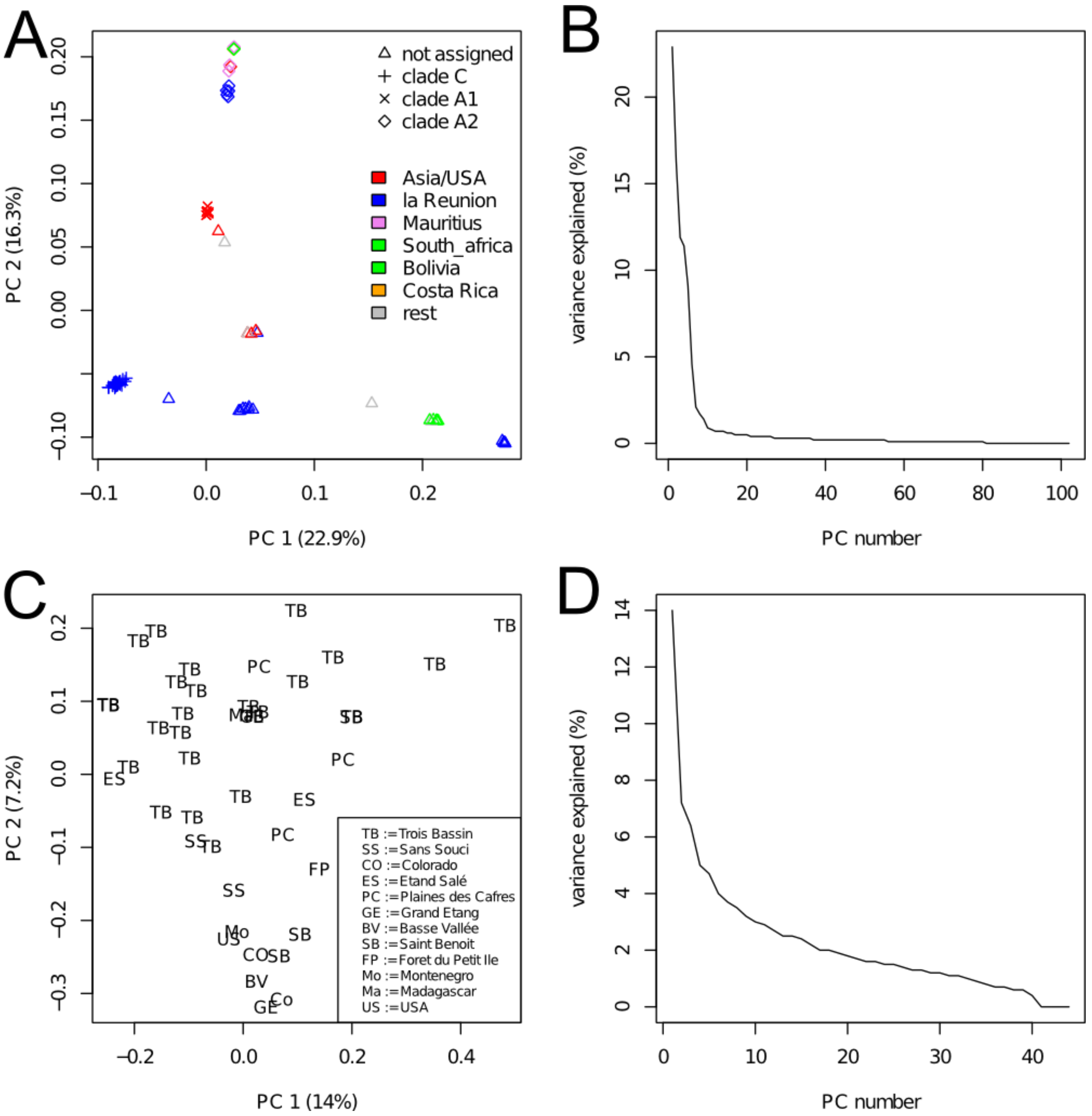

Figure S3 Principal component analysis of SNV data. In order to reduce bias by short range LD, principal component analysis was performed by EIGENSOFT 3.0, using one biallelic SNV with 5-95% allele frequency per 50kb window. **(A)** First two principal components (PC) for all strains. Both PCs were significant (*P<0.001*) according to Tracy-Widom statistics. **(B)** Variance, explained by the individual principal components. **(C)** First two PCs for strains from clade C (*P<0.001*, according to Tracy-Widom statistics). The first two PCs reveal a separation between strains sampled from Trois Bassins as opposed to most other locations. **(D)** Variance, explained by the individual PCs for clade C.

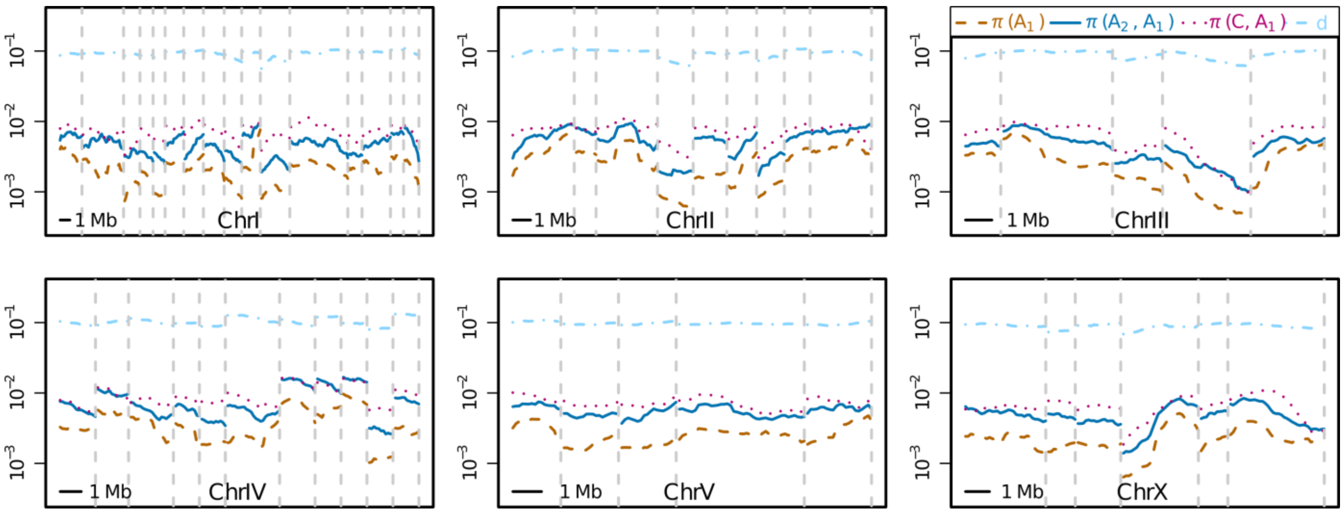

Figure S4 Clade A_1 diversity. Divergence (d) between *P. pacificus* and *P. exspectatus*, nucleotide diversity π within clade A_1 and between A_1 and clades C and A_2 are shown. π values represent averages over 1Mb windows. Supercontigs (>1Mb) with markers on the genetic map were concatenated to visualize the chromosomal distribution. The dashed lines denote supercontig boundaries with unknown physical distance.

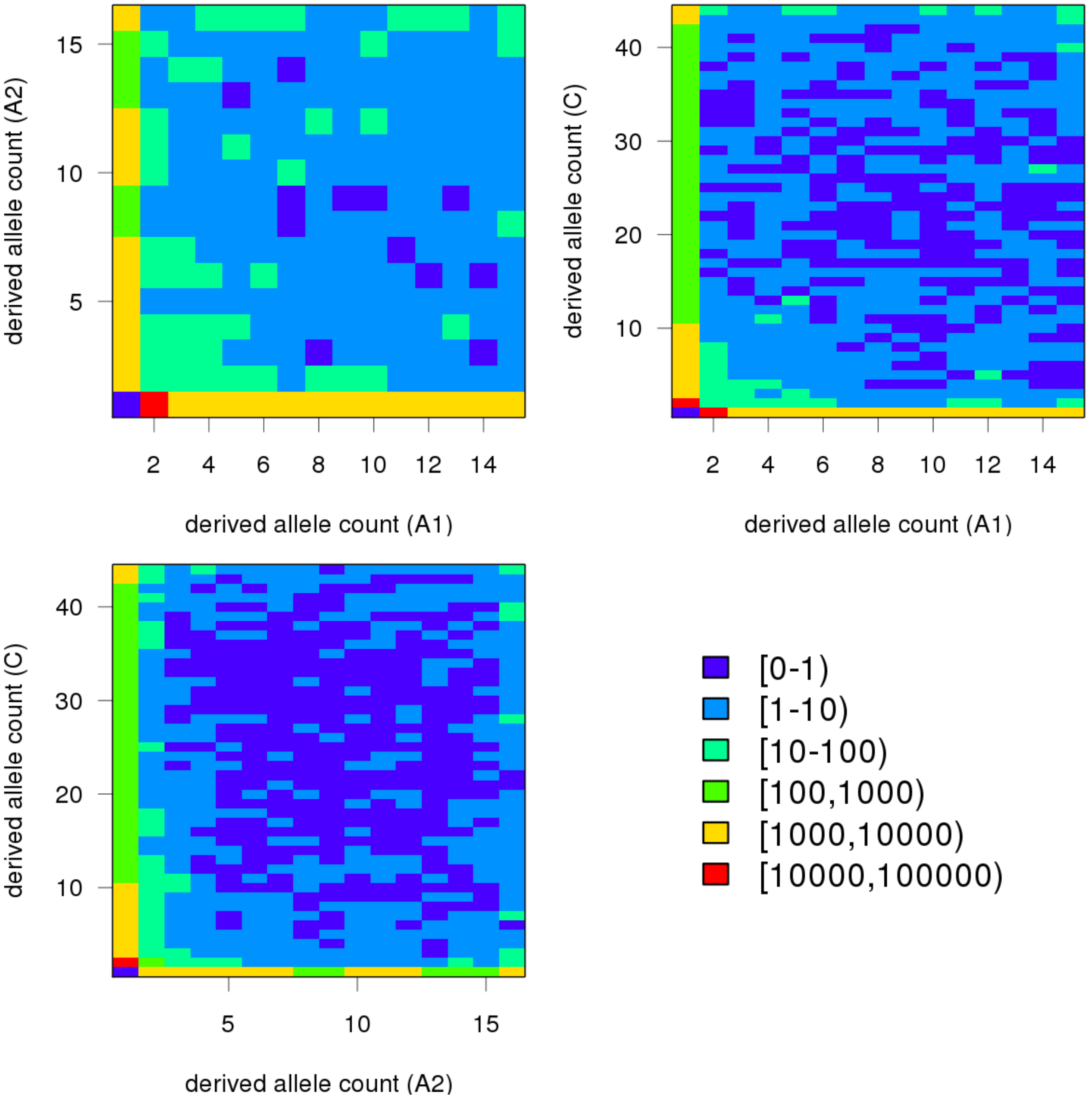

Figure S5 Joint site frequency spectra (SFS) for clades A1, A2, and C. Total numbers of derived allele counts are shown for all three pairwise comparisons of clades. The vast majority of variation is located along the x and y-axis indicating that most variation is clade-specific. Only a small percentage of total variation is shared between clades and may therefore represent ancestral diversity, introgression, or convergent evolution.

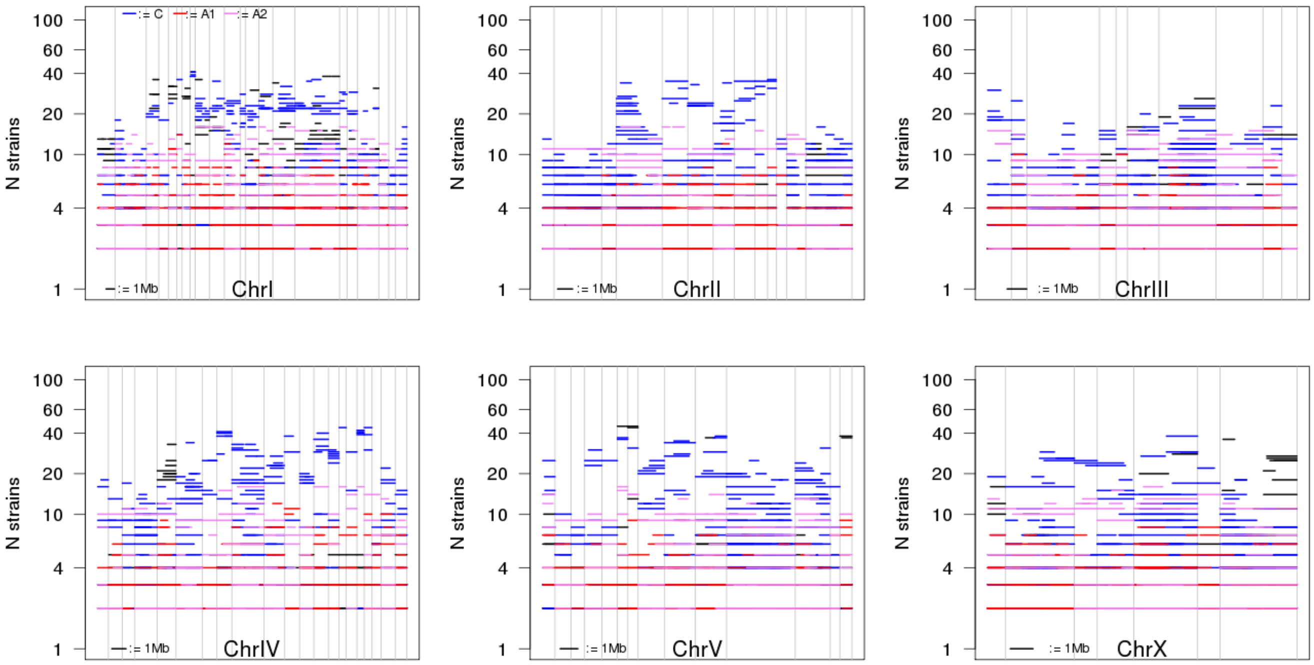

Figure S6 Length and frequency distribution of shared haplotype blocks. The x-axis denotes the relative positions of shared haplotypes separated by supercountig boundaries. All supercontigs with markers that have been mapped to the genetic mapped are shown. The haplotypes are colored with respect to a clade if all strains sharing that haplotype are member of that clade or black otherwise.

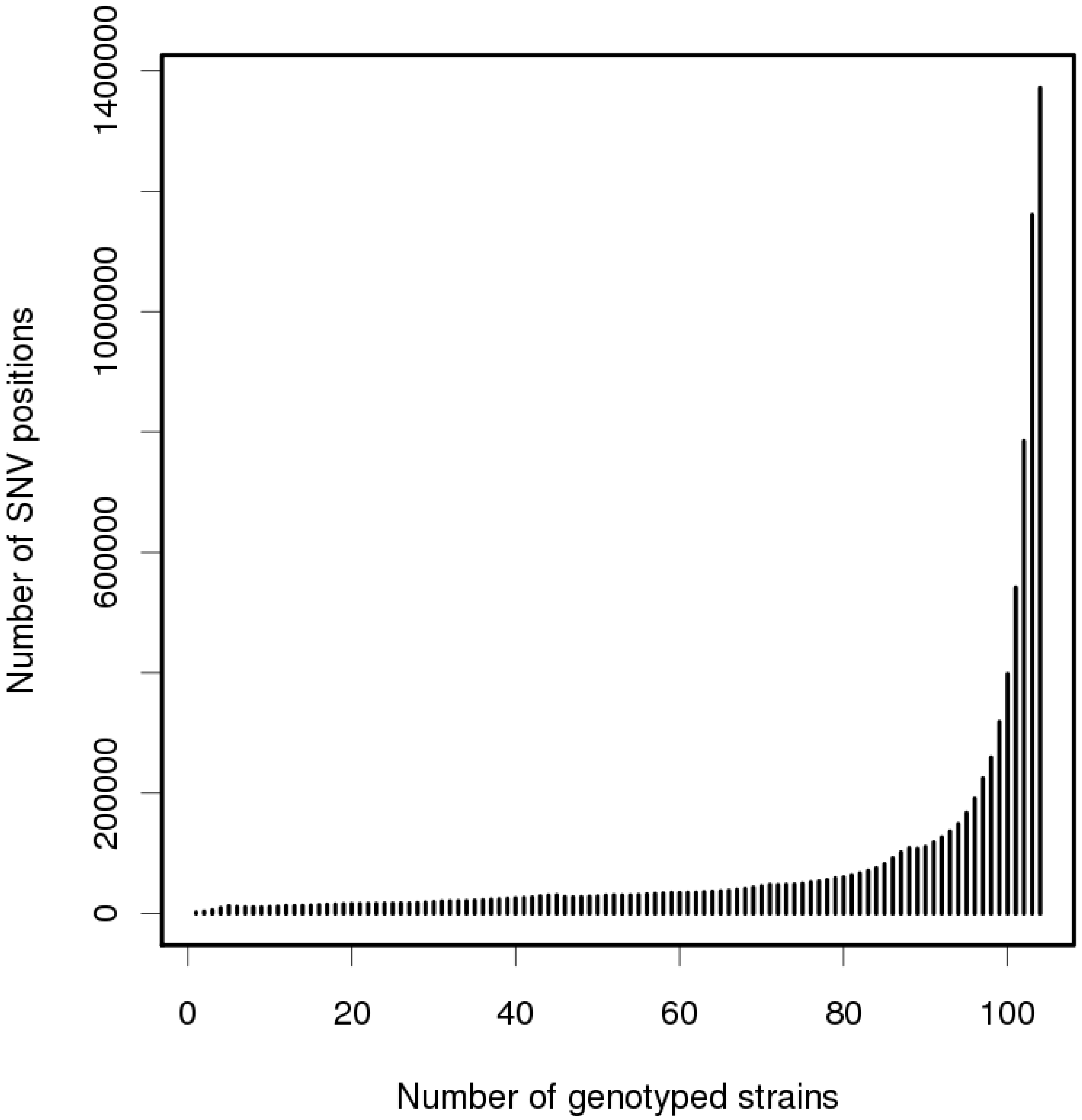

Figure S7 Number of SNV positions that could be genotyped in X strains. For all detected variable sites, we tested in how many strains these sites could be genotyped reliably (coverage ≥ 2, samtools quality score ≥ 20, no signal of heterozygosity in any of the strains). Around 1.4 million positions could reliably be genotyped in all 104 strains. For most population genetic analysis was done using only those 1.4 mio SNV positions.

Table S1 Numbers about sequenced strains

Available for download as an Excel file at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.159855/-/DC1