nature genetics

The *Pristionchus pacificus* genome provides a unique perspective on nematode lifestyle and parasitism

Christoph Dieterich[1], Sandra W Clifton[2], Lisa N Schuster[1], Asif Chinwalla[2], Kimberly Delehaunty[2], Iris Dinkelacker[1], Lucinda Fulton[2], Robert Fulton[2], Jennifer Godfrey[2], Pat Minx[2], Makedonka Mitreva[2], Waltraud Roeseler[1], Huiyu Tian[1], Hanh Witte[1], Shiaw-Pyng Yang[2], Richard K Wilson[2] & Ralf J Sommer[1]

Here we present a draft genome sequence of the nematode *Pristionchus pacificus*, a species that is associated with beetles and is used as a model system in evolutionary biology. With 169 Mb and 23,500 predicted protein-coding genes, the *P. pacificus* genome is larger than those of *Caenorhabditis elegans* and the human parasite *Brugia malayi*. Compared to *C. elegans*, the *P. pacificus* genome has more genes encoding cytochrome P450 enzymes, glucosyltransferases, sulfotransferases and ABC transporters, many of which were experimentally validated. The *P. pacificus* genome contains genes encoding cellulase and diapausin, and cellulase activity is found in *P. pacificus* secretions, indicating that cellulases can be found in nematodes beyond plant parasites. The relatively higher number of detoxification and degradation enzymes in *P. pacificus* is consistent with its necromenic lifestyle and might represent a preadaptation for parasitism. Thus, comparative genomics analysis of three ecologically distinct nematodes offers a unique opportunity to investigate the association between genome structure and lifestyle.

Nematodes occupy a wide range of ecological niches, from free-living microbivores or predators to parasites. *C. elegans*, a soil-dwelling bacteriovorous nematode often found in compost heaps, was the first metazoan to have its genome sequenced[1,2]. The genome of the filarian parasite *B. malayi* was recently sequenced, revealing a slightly lower number of protein-coding genes compared to *C. elegans*[3]. Here we present the first analysis of the genome of the necromenic nematode *P. pacificus*, a species that lives in association with beetles. We predicted the *P. pacificus* genome to be 169 Mb in size and, thus, substantially larger than that of *C. elegans*. We also predicted at least 23,500 protein-coding genes in the 142 Mb of robustly assembled sequence, indicating that there are considerable differences in nematode genomes depending on their lifestyle.

P. pacificus resembles *C. elegans* in many traits, including a short generation time, hermaphroditic propagation and simple laboratory culture, all of which allow the use of forward and reverse genetic tools[4,5]. *P. pacificus* is a model system in evolutionary developmental biology, and analysis of vulva formation has revealed substantial differences between *P. pacificus* and *C. elegans*[6–9]. The ecological niche occupied by *P. pacificus* is completely different from that of *C. elegans*. *Pristionchus* nematodes live in close association with beetles in a nearly species-specific manner. *P. pacificus*, for example, is associated with the oriental beetle *Exomala orientalis* in the United States and Japan[10]. At the beginning of this association, known as necromeny, nonfeeding dauer larvae actively invade the beetle. The larvae remain arrested in the dauer stage until the death of the insect and resume development by feeding on bacteria, fungi and nematodes that grow on the insect's carcass[2]. Necromenic species are thus permanently exposed to a diversity of microbes and the xenobiotic compounds produced by these organisms, which may require specific defense mechanisms.

RESULTS

Sequencing and assembly

Four different *Pristionchus* genomes were sequenced at different coverage levels using the whole-genome shotgun method (**Fig. 1a**). The *P. pacificus* California strain was chosen as the reference genome, as most experimental and genetic studies so far have been carried out with this strain. The *P. pacificus* Washington strain was selected for low-coverage sequencing, as it is known to be polymorphic compared to the California strain and has been used to generate a genetic linkage map[11]. *Pristionchus maupasi* and *Pristionchus entomophagus* were sampled because their evolutionary distance from *P. pacificus* would allow phylogenetic footprinting[12].

Our draft assembly of the *P. pacificus* California reference genome attained nine-fold coverage from plasmid, fosmid and BAC libraries and was divided into 2,894 scaffolds (see Methods). The calculated genome size of this draft assembly was around 169 Mb, with the major contigs (>1 kb) having an N50 value of 18.1 kb (N50 is the supercontig size above which 50% of the assembled sequence can be

[1]Max-Planck Institute for Developmental Biology, Spemannstrasse 37, 72076 Tübingen, Germany. [2]Genome Sequencing Center, Washington University School of Medicine, Campus Box 8501, 4444 Forest Park Boulevard, St. Louis, Missouri 63108, USA. Correspondence should be addressed to R.K.W. (rwilson@wustl.edu) or R.J.S. (ralf.sommer@tuebingen.mpg.de).

Received 28 January; accepted 7 July; published online 21 September 2008; doi:10.1038/ng.227

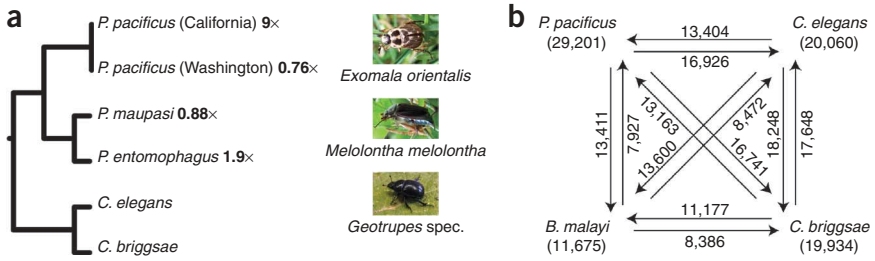

Figure 1 Phylogeny and comparison of genomic features in nematodes. (**a**) Schematic representation of the phylogenetic relationship of the four sequenced *Pristionchus* genomes and *C. elegans*. The *P. pacificus* California strain PS312 was the reference genome. The *P. pacificus* Washington strain PS1843 was used to generate a genetic linkage map and was highly polymorphic to the California strain. *P. maupasi* and *P. entomophagus* are close relatives of *P. pacificus* and were especially suited for phylogenetic footprinting studies. Scarab beetle images to the right of each *Pristionchus* species highlight the specific nematode-beetle association of the respective *Pristionchus* species. (**b**) Pairwise comparison scheme showing homology relations between the predicted proteomes of *P. pacificus*, *C. elegans*, *C. briggsae* and *B. malayi*. Numbers represent best BLASTP matches (score ≥ 50 bits) for bidirectional similarity searches.

found). The genome of the Washington strain differed from the California strain in 4.3% of all ungapped positions in a whole-genome alignment. We found 40,682 indels larger than 10 bp. Simple sequence-length polymorphism markers were generated in the largest scaffolds to link them to the genetic linkage map, which was successful for 90% of the assembled nucleotides. Together, the *P. pacificus* California and Washington genomes showed the highest rates of polymorphism in any known nematode species. The other two *Pristionchus* genomes were almost equally diverged from *P. pacificus* at the primary sequence level and support the detection of evolutionarily constrained genomic elements. As in *C. elegans,* the 1*n* chromosome number was 5+1 for all three *Pristionchus* species. The average overall GC content was slightly higher in *Pristionchus* genomes (~42%) than in *Caenorhabditis* genomes (~38%). Data from this study are freely available from the Washington University Genome Sequencing Center and the Department of Evolutionary Biology at the Max-Planck Institute for Developmental Biology and are currently incorporated into WormBase (see Methods for URLs).

Annotation of the *P. pacificus* genome

We assessed the repeat content of the *P. pacificus* genome and applied the RECON algorithm with a subsequent assembly step. This procedure identified 512 complex repetitive elements with a minimal copy number of 10 (**Supplementary Table 1** online). Complex and other repeats covered ~17% of the sequenced genome. The difference in genome size from 100 Mb in *C. elegans* to roughly 169 Mb in *P. pacificus* cannot be fully explained by this greater repeat content, although it certainly contributes.

To characterize the protein-coding portion of the *P. pacificus* genome, we began by identifying protein-coding genes. We generated an in-house database of 7,266 SL1 and 4,821 SL2 *trans*-spliced ESTs. Both types of ESTs were separately clustered and aligned onto the genome. This process yielded 2,156 unique SL1 and 1,112 unique SL2 transcript structures. We used three distinct gene prediction algorithms—Augustus, SNAP and GlimmerHMM—all of which were trained on the SL1-spliced EST dataset. The algorithms delivered 19,878, 29,201 and 33,769 gene predictions, respectively, for the *P. pacificus* genome (**Table 1** and **Supplementary Table 2** online)[13–15].

To assess the accuracy of the available gene prediction sets, we tested them using the curated dataset of SL2-spliced ESTs and by experimental validation of selected gene families. Both types of analyses strongly supported the results of the SNAP dataset. Specifically, SNAP attained 95% specificity and 77% sensitivity at the nucleotide level, whereas Augustus attained 96% specificity but only 67% sensitivity (**Supplementary Table 2**). At the gene level, SNAP predicted 944 (85%) of 1,112 SL2-spliced genes, whereas Augustus predicted only 845 (76%) of the genes. We did not further consider the GlimmerHMM prediction set, as it did not perform comparably to the two others. Similarly, experimental validation by RT-PCR of 367 gene predictions encoding cytochrome P450, sulfotransferases, trypsins and other enzymes showed a success rate of 81% and 70% for SNAP and Augustus, respectively. Given the superior performance of SNAP on the *P. pacificus* genome assembly, we used this gene prediction set for all further analyses. Of the 29,201 genes predicted by SNAP, 6,788 were unique and did not overlap with Augustus predictions. In validation experiments, 42 (76%) of 55 genes of this class were confirmed, indicating that the majority of these unique gene predictions were real. With this estimate, we calculated a final minimal gene content of approximately 23,500 genes in the 84% of the *P. pacificus* genome present in the current assembly (**Table 1** and **Supplementary Table 2**). *P. pacificus* gene predictions were not biased to either merging or splitting genes, given that 400 randomly sampled 1:1 orthologs of *P. pacificus* and *C. elegans* showed similar cDNA length distributions (**Supplementary Fig. 1** online). Overall, the genomic length distribution of *P. pacificus* transcripts was similar to that of *C. elegans* (**Supplementary Fig. 2** online). The greater number of transcripts therefore provides a partial explanation for the increase in genome size, whereas we saw no evidence for segmental genome duplications.

Gene structure

We observed a clear trend toward higher exon numbers in *P. pacificus* transcripts compared to *C. elegans* (**Table 1**). A linear regression on the Q-Q plot of exon numbers revealed roughly twice (2.129) as many exons in *P. pacificus* genes than in *C. elegans* genes. The increase in exon numbers was strongly associated on a gene-by-gene basis in 1:1 orthologs ($P < 10^{-16}$; **Supplementary Fig. 3** online). Median exon and intron sizes also differed substantially between the two species (**Table 1**). *P. pacificus* exons were 42% shorter and its introns 59%

Table 1 Comparison of *P. pacificus* PS312 and *C. elegans* N2 gene features

Gene feature	*C. elegans*	*P. pacificus*
Predicted gene number[a]	—	29,201
Final minimal gene number[b]	20,060	23,500
Median number of coding exons per transcript[c]	6	11
Median exon length (bp)	147	85
Median intron length (bp)	69	110
Median genomic transcript length (bp)	1,941	1,675
Median coding sequence length (bp)	1,029	618

[a]SNAP gene predictions. [b]Protein-coding genes in WormBase 160 and gene number estimate. [c]Based on 1:1 orthologs.

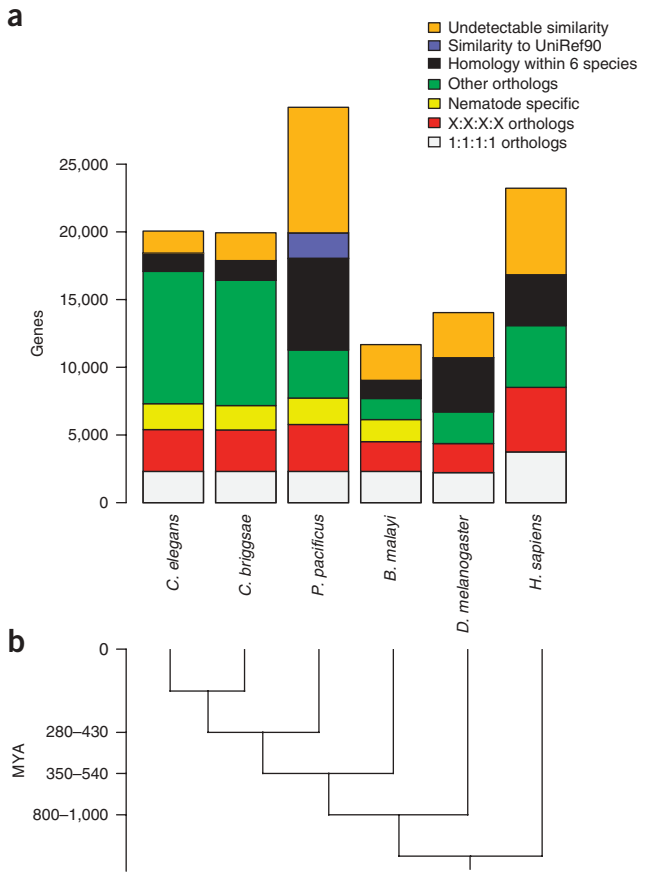

Figure 2 Orthology assignment in nematodes and comparison to non-nematode species. (**a**) Comparison of proteomes of four nematodes and two additional metazoan genomes separated by increasing phylogenetic distances from *P. pacificus*. All proteomes were divided into subsets according to the inferred homology relations. The 1:1:1:1 set denotes single-copy orthologs that were present in all four nematode species. X:X:X:X indicates orthologs that were possibly present in multiple copies in individual species. Nematode-specific orthologs did not have a counterpart in either *D. melanogaster* or *H. sapiens*. Orthologs that did not fall into any of the previous categories were classified as "other orthologs." Protein similarity searches (score $\geq$ 50 bits) within the six proteomes delivered additional putative homologs. At the extremes, $\sim$11,000 predicted proteins of *P. pacificus* had no counterpart in any of the five other organisms. When homology searches to UniRef90 (UniProt reference clusters; all known proteomes grouped into protein clusters sharing >90% sequence identity; E-value $\leq 10^{-2}$) were included, this number was reduced by 1,887 predicted proteins. In total, $\sim$20,000 *P. pacificus* predicted proteins showed detectable sequence similarity to other organisms. (**b**) Divergence times were estimated by analysis of 574 clocklike trees of 1:1:1 orthologs between *B. malayi*, *P. pacificus* and *C. elegans*. Nematodes diverged from arthropods 800–1,000 million years ago (MYA). See **Supplementary Figure 6** for details.

longer than their *C. elegans* counterparts. Genomic extent and length of *P. pacificus* coding regions were 14% and 40% lower, respectively, than in *C. elegans*. As outlined below, further analysis indicated that *P. pacificus*–specific genes with no orthologs in *C. elegans* were on average shorter than conserved genes.

Gene content

To characterize the protein-coding genes, we explored homology relationships among the proteomes of the nematodes *B. malayi*, *C. elegans*, *Caenorhabditis briggsae* and *P. pacificus*. We inferred putative groups of orthologs with the MultiParanoid[16] approach; we determined possible protein functions with the PFAM[17] and KEGG[18] annotations.

BLASTP similarity searches (homology cutoff of 50 bits, which is equivalent to an alignment of approximately nine consecutive matches) revealed that 58% of all *P. pacificus* gene predictions showed substantial similarity to 68% of the *C. elegans* proteome (**Fig. 1b**). Similar proportions were observed for the comparison of *P. pacificus* and *C. briggsae* (57% and 66%, respectively). About 46% of the *P. pacificus* proteome matched 68% of the available *B. malayi* proteome. Likewise, 73% of the *B. malayi* proteome matched either of the two *Caenorhabditis* genomes.

We further analyzed protein-level similarities by defining orthologous gene groups for a set of six species: *P. pacificus, C. elegans, C. briggsae, B. malayi, Drosophila melanogaster* and *Homo sapiens* (**Fig. 2**). A core set of 4,700 orthologous gene groups is common to all four nematode species and possibly existed in a common ancestor. A gene ontology analysis of the associated biological processes revealed enrichment in core metabolic processes, such as nucleotide, nucleic acid and protein metabolic processes. Some orthology relations were exclusively established between pairs of nematodes: 197 gene sets between *B. malayi* and *P. pacificus* and 1,163 between *P. pacificus* and the two *Caenorhabditis* species. Orthology relations also existed between a single nematode and non-nematode species (38 genes between *P. pacificus* and the two non-nematode species).

We next independently grouped the proteomes of *C. elegans* and *P. pacificus* into families using a Markov cluster algorithm[19]. The two proteomes were separately grouped into 10,192 and 16,944 clusters, respectively; 1,011 of the *C. elegans* and 3,253 of the *P. pacificus* gene groups contained more than one gene (data not shown). Notably, many of the *P. pacificus*–specific genes were grouped into gene families by this approach (see below). Transcription factors and signal transduction pathways, which are important for the evolutionary analysis of developmental processes, were generally conserved between *P. pacificus* and *C. elegans*, although copy-number differences were found, such as for the GATA transcription factors (6 copies in *P. pacificus*, 12 copies in *C. elegans*).

We observed notable differences in protein domain distributions of the *C. elegans* and *P. pacificus* proteomes. **Figure 3** shows the four PFAM domains and KEGG pathways with the greatest absolute increase and decrease in copy numbers relative to *C. elegans*. Three of the top four increased PFAM classes in *P. pacificus* (cytochrome P450, UDP-glycosyltransferases and carboxylesterases) have a key role in the metabolism of xenobiotics. The detoxification of xenobiotics is divided into two phases[20], functionalization (phase I) and conjugation (phase II; **Fig. 4**). In phase I, the water solubility of xenobiotics is improved by the addition of electrophilic or nucleophilic groups. In phase II, hydrophilic conjugation results in compounds that can readily be shuttled out of the cell through active transport mechanisms, such as ABC transporters. In the *P. pacificus* genome, we observed elevated copy numbers of almost all gene types involved in the metabolism of xenobiotics (**Fig. 4**). Most pronounced were the increases in cytochrome P450 enzymes in phase I and sulfotransferase and UDP-glucuronosyltransferase enzymes in phase II. To test the validity of the cytochrome P450 and sulfotransferase gene predictions, we used cDNA verification and confirmed 80% (160 of 198) and 100% (17 of 17) of the gene predictions, respectively, under standard laboratory conditions (**Fig. 4**). We also observed several other gene family expansions in *P. pacificus*, including an enigmatic increase in

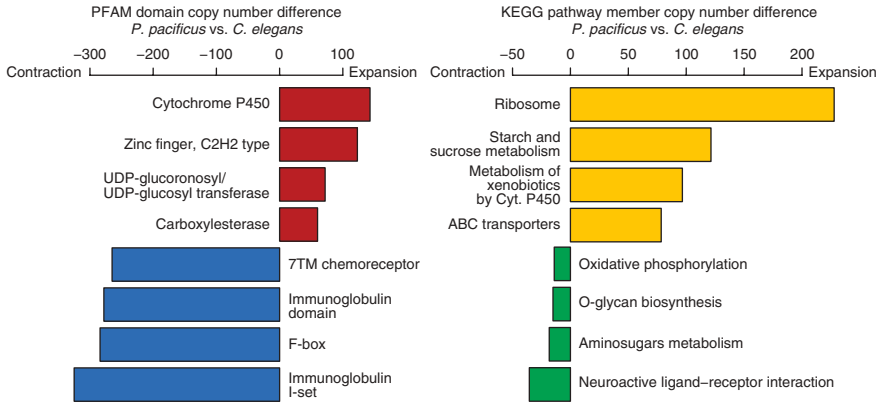

Figure 3 Examples of *P. pacificus*–specific expansions. Shown are the top four prominent expansions and contractions of PFAM protein domain families and KEGG pathway annotations. Positive numbers indicate an increase in *P. pacificus* copy number relative to *C. elegans*; negative numbers indicate a decrease.

ribosomal proteins (especially S19, L37 and S5e). Likewise, gene family reductions relative to *C. elegans* were common—for example, genes containing the receptor L domain (3 genes in *P. pacificus* versus 63 genes in *C. elegans*). Notably, the number of seven-transmembrane-receptor genes was also lower in *P. pacificus* compared to *C. elegans* (69 versus 527 copies for PF1604 and PF1461). We speculated that the large number of nuclear hormone receptors in *C. elegans* is an adaptation to the less defined environment of this nematode.

A substantial number of the *P. pacificus* gene predictions may constitute novel genes with unique functions in the biology and ecology of *P. pacificus* (**Fig. 2a**). Many of these gene predictions were supported by intraspecies protein similarities, conserved coding regions and local gene clustering. More than two-thirds of a randomly selected test set of hypothetical genes were confirmed by cDNA expression (data not shown). Evidence from EST data also supported some of these specific genes. We tried to infer possible annotations for *P. pacificus*–specific proteins by protein similarity searches against protein sequence and domain databases (UniRef90 and PFAM). Such searches expanded the set of homologous genes by 1,887 that matched the UniRef90 database at an E-value cutoff of 10^{-2}, but not any of the five aforementioned genomes, including *C. elegans* (**Fig. 2a**).

Among the 1,887 genes that matched the UniRef90 database were seven putative cellulase (EC 3.2.1.4) genes, which had previously only been reported from plant-parasitic nematodes[21,22]. Six of these seven gene predictions were confirmed by cDNA expression, and all had the required active-site residues (**Supplementary Fig. 4** online). All *P. pacificus* cellulases belonged to glycosyl hydrolase family 5, and sequence comparison revealed that they were of common origin (**Supplementary Fig. 5** online). *P. pacificus* cellulases were most similar to cellulases of the bacterium *Xylella fastidiosa*, *Pyrococcus* archaea and the slime mold *Dictyostelium* (**Supplementary Fig. 5**). A Congo red–polysaccharide interaction assay clearly detected cellulase activity in the supernatant of mixed-stage *P. pacificus* liquid cultures (**Fig. 5**). The cellulase *Ppa-cel-1* was expressed in the posterior pharynx and the anterior part of the intestine and might therefore be involved in energy metabolism (**Fig. 5**). The cellulases of the plant-parasitic nematodes *Globodera rostochiensis* and *Heterodera glycines* have been suggested to result from lateral gene transfer (LGT)[21,23]. The finding of diverse cellulases in the genomes of *P. pacificus* and plant-parasitic nematodes supported their ancient and independent origin and indicated that cellulases are not restricted to parasitic nematodes. Indeed, it has been suggested that microbes in the digestive systems of nonparasitic nematodes are the source of cellulase and other genes and that LGT events had a crucial role in the evolutionary transition toward parasitism[23]. *P. pacificus* might therefore represent an important intermediate stage with a nonparasitic lifestyle.

The presence of several other gene families also supported an involvement of LGT in the acquisition of new genes in *P. pacificus*. We identified and experimentally confirmed four of five *P. pacificus* diapausin genes (data not shown). The diapausin protein was originally isolated from the leaf beetle *Gastrophysa atrocyanea* and contains several disulfide bridges[24]. Diapausin is thought to protect the dormant beetle from fungal infections by acting as an N-type voltage-gated calcium channel blocker. It has been suggested that the beetle diapausin gene was acquired by LGT from iridoviruses, which contain putative peptides with sequence similarity to diapausin[24]. Given the association of *Pristionchus* with hibernating beetles, the finding of diapausin genes in the *P. pacificus* genome might indicate complex genomic interactions of organisms sharing the same ecological niche. However, with multiple bacterial and viral genomes also being present in the same habitat, the putative donors and recipients of these potential LGTs can only be identified in future studies involving all potential partners. Furthermore, we found 16 EST clones that were much more similar to genes from soil bacteria than from eukaryotes. Of these 16 genes, 13 were independently verified by cDNA expression (data not shown). *P. pacificus* might therefore have acquired a substantial fraction of its new genes by LGT and related mechanisms. Although LGT from bacteria to eukaryotes is still controversial, the cellulase example of *P. pacificus* is supported by longevity and integration into host biology, two notable features of LGT[25,26].

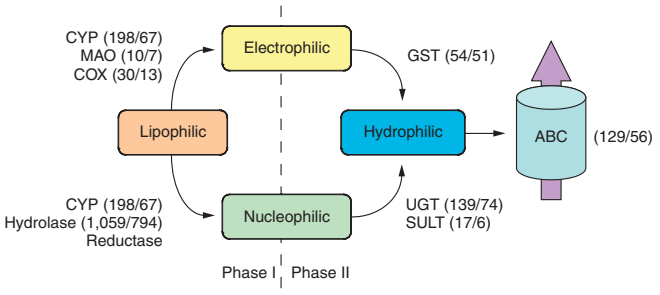

Figure 4 Schematic representation of the metabolism of xenobiotics. Lipophilic xenobiotics are modified to form electrophilic or nucleophilic substances (phase I) by the activity of several enzyme groups, such as cytochrome P450 monooxygenases (CYP), monoamine oxidases (MAO), cyclooxygenases (COX), hydrolases and reductases. In phase II, hydrophilic compounds are formed by the activity of transferases (GST, glutathione-*S*-transferase; UGT, UDP-glucuronosyltransferases; SULT, sulfotransferase). ABC transporters export these modified compounds. *P. pacificus* shows an expansion of most of these enzyme groups compared to *C. elegans*. Species-specific copy numbers are given in parentheses (*P. pacificus*/*C. elegans*) next to each enzyme class.

Figure 5 Cellulase activity in *P. pacificus*. Left, cellulase activity assay on carboxymethylcellulose agar plates. Clear halos were detected in the positive control (purified *Aspergillus* cellulase) and the supernatant of *P. pacificus* mixed-stage cultures. No halos were seen in the supernatant of mixed-stage *C. elegans* or *Escherichia coli* cultures. Right, *in situ* hybridization of the cellulase *Ppa-cel-1* revealed expression in the posterior pharynx and the anterior part of the intestine.

Estimation of nematode divergence dates

In nematodes, the estimation of divergence dates is hampered by the absence of fossils. However, whole-genome comparisons have been used to estimate the divergence of *C. elegans* and *C. briggsae* and the divergence of *Brugia* and *Caenorhabditis* using the divergence of nematodes from arthropods (800–1,000 million years ago) for calibration[3,19]. We analyzed 574 clocklike trees of 1:1:1 orthologs (*B. malayi*, *P. pacificus* and *C. elegans*) to position the divergence time of *Pristionchus* and *Caenorhabditis* (**Fig. 2b** and **Supplementary Fig. 6** online). Setting the divergence of *Brugia* and *Caenorhabditis* to 1.0 resulted in an estimate of 0.8 for the *Pristionchus-Caenorhabditis* separation. Plugging in the previously published divergence dates[3,19], we estimated that *Pristionchus* and *Caenorhabditis* separated 280–430 million years ago.

DISCUSSION

The draft genome of the necromenic nematode *P. pacificus* covered 84% of the predicted 169 Mb genome. We annotated a total of 23,500 gene predictions. This proteome complexity may be related to the ecology of this organism. With its necromenic beetle association, *P. pacificus* has an intermediate position between the microbivorous *C. elegans* and true parasites.

Nematode parasitism has evolved multiple times independently[27]. However, little is known about the genetic and genomic factors involved in the initial steps toward parasitism. Comparative genomics of ecologically distinct nematodes offers a unique opportunity to study the association between genome structure and life-style. The omnivorous feeder *P. pacificus* should have the most complex metabolic pathways for nutrition and protection against defense and prey compared to the microbivorous *C. elegans* or the animal parasite *B. malayi*, which live in host-controlled environments. We speculate that the proteome complexity of *P. pacificus* and other omnivorous nematodes represents a prerequisite for parasitism. It has indeed been argued that necromenic associations serve as preadaptations toward true parasitism[28,29]. Tolerance to low oxygen concentrations and the toxicity of host enzymes, as well as morphological characteristics such as dauer larvae and movable mouthparts, might be facilitators of the evolutionary change to parasitism[28,29]. The genome and the extensive genetic tool kit of *P. pacificus* provide a platform for the analysis of the development, behavior and ecology of this organism and might be a useful paradigm for studying the transition to nematode parasitism.

METHODS

Genome assembly and size estimates. We used a whole-genome shotgun strategy to assemble the *P. pacificus* genome. Three different DNA library types (plasmids, fosmids and BACs) with insert sizes of ~4 kb, ~40 kb and 80–150 kb were end-sequenced and submitted to the assembly process. In total, we generated 2,063,200 paired-end plasmid reads, 43,389 paired-end fosmid reads, 41,230 single-end fosmid reads and 35,419 paired-end BAC reads. The assembly of reads was carried out using the PCAP.REP software[30]. We placed 2,020,055 reads on the final assembly. The assembly size amounted to ~146 Mb of ungapped and ~169 Mb of gapped sequence. These size estimates agreed with *k*-mer count statistics and flow cytometry measurements (161.7 ± 0.7 Mb; data not shown). The total supercontig number was 5,106, of which 2,894 were larger than 1 kb. The maximal supercontig length was ~3.2 Mb, and the N50 supercontig size was 737,446 bp. We determined the average GC content from the assembly to 43%.

Repeat analysis. Repeats covered 17% of the *P. pacificus* genome[31]. Repeat coverage can be subdivided into repeat classes that may overlap on the genome level. Low-complexity regions covered 11%, tandem repeats 3% and complex repeats 14% of the genome, respectively. In *C. elegans* and *C. briggsae*, 16.5% and 22% of the genomes are repetitive, respectively[19]. The increase in genome size from 100 Mb in *C. elegans* to roughly 169 Mb in *P. pacificus* therefore cannot be explained by an increase in repetitive elements. Fifty SNAP gene predictions contained an identifiable transposase domain, and 42 SNAP gene predictions may encode a reverse transcriptase domain (E-value $< 10^{-2}$).

Gene finding. We generated an in-house database of 7,266 SL1 and 4,821 SL2 *trans*-spliced ESTs. Both types of EST sequences were separately clustered with CAP3 (ref. 32) and aligned onto the genome with Exonerate[33]. This process yielded 2,156 unique SL1 and 1,112 unique SL2 transcript structures. The SL1 ESTs served as training data for the gene prediction algorithms SNAP[14] and AUGUSTUS[13]. We had to infer the correct reading frame for our training set from homologous *C. elegans* proteins, given the incomplete EST information. The SNAP gene model was configured to predict the coding region of *trans*-spliced genes (additional information is available at http://www.pristionchus.org). Performance of the gene modeler and the gene combiner (JIGSAW[34]) is reported in **Supplementary Table 2**.

Computation of orthology relationships. Pairwise homology and orthology relationships were inferred from whole-genome BLASTP similarity searches using the InParanoid software[35]. MultiParanoid[16] was then used to discover orthology relationships between proteins in multiple proteomes.

Gene families. The predicted *P. pacificus* proteome was grouped into gene families using the MCL algorithm[36]. We took normalized BLASTP bit scores as a measure of unidirectional similarity. For example, the similarity of protein A to protein B was expressed as 2 × score(AB) / score(AA) + score(BB). The MCL algorithm was run using parameters –I 1.6 and –s 6. The same parameter setting was used for the analysis of *C. briggsae* gene families[19].

EST analysis. We used all publicly available *P. pacificus* ESTs from dbEST (26,262 ESTs). We also sequenced 7,266 SL1 ESTs and 4,821 SL2 ESTs. We used two alternative strategies to map the ESTs onto the genome: (i) align first, cluster later, with EST alignments using Exonerate[33]; and (ii) cluster first, align later, with all ESTs clustered using CAP3 with default parameters. The EST cluster consensus sequences were aligned to the *P. pacificus* genome using Exonerate[33].

Divergence time estimates. We compiled a set of ~1,051 strictly single-copy orthologs for *B. malayi*, *P. pacificus*, *C. elegans*, *C. briggsae* and *D. melanogaster*. Two phylogenetic trees were computed for each set of nematode proteins using a maximum-likelihood method (PHYLIP, Jones-Taylor-Thornton model of protein evolution with or without the assumption of a molecular clock). We considered a protein set to evolve clocklike if a likelihood ratio test confirmed

that both likelihoods did not differ significantly ($P > 0.05$). Actual divergence time estimates were computed for each clocklike tree using the nonparametric rate-smoothing algorithm described in Sanderson[37].

URLs. Washington University Genome Sequencing Center, http://genome.wustl.edu/; Department of Evolutionary Biology, Max-Planck Institute for Developmental Biology, http://www.pristionchus.org; WormBase, http://www.wormbase.org/; dbEST, http://www.ncbi.nlm.nih.gov/projects/dbEST/.

Database accession numbers. The dbEST dataset reported in this study contained the accession numbers FG094704–FG106788.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank P.W. Sternberg, J. Srinivasan and members of the Sommer lab for discussion and helpful comments on the manuscript. This work was funded by National Human Genome Research Institute grant U54HG003079 and the Max-Planck Society.

AUTHOR CONTRIBUTIONS

C.D. carried out most of the bioinformatics analysis; the Genome Sequencing Center team at Washington University (S.W.C., A.C., K.D., L.F., R.F., J.G., P.M., M.M., S.-P.Y., R.K.W.) conducted the genome sequencing project; L.N.S. and H.T. did the experimental gene confirmation; I.D., W.R. and H.W. experimentally linked the genetic linkage map to the *P. pacificus* genome; and C.D., S.W.C., R.K.W. and R.J.S. designed these studies and contributed to the writing of this paper.

Published online at http://www.nature.com/naturegenetics/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282**, 2012–2018 (1998).
2. Kiontke, K. & Sudhaus, W. Ecology of *Caenorhabditis* species. in *WormBook* ed. (The *C. elegans* Research Community, WormBook, doi/10.1895/wormbook.1.37.1, 2006).
3. Ghedin, E. *et al.* Draft genome of the filarial nematode parasite *Brugia malayi*. *Science* **317**, 1756–1760 (2007).
4. Sommer, R.J., Carta, L.K., Kim, S.-Y. & Sternberg, P.W. Morphological, genetic and molecular description of *Pristionchus pacificus* sp. n. *Fundam. Appl. Nematol.* **19**, 511–521 (1996).
5. Hong, R.L. & Sommer, R.J. *Pristionchus pacificus*: a well rounded nematode. *Bioessays* **28**, 651–659 (2006).
6. Zheng, M., Messerschmidt, D., Jungblut, B. & Sommer, R.J. Conservation and diversification of Wnt signaling function during the evolution of nematode vulva development. *Nat. Genet.* **37**, 300–304 (2005).
7. Schlager, B., Röseler, W., Zheng, M., Gutierrez, A. & Sommer, R.J. HAIRY-like transcription factors and the evolution of the nematode vulva equivalence group. *Curr. Biol.* **16**, 1386–1394 (2006).
8. Yi, B. & Sommer, R.J. The *pax-3* gene is involved in vulva formation in *Pristionchus pacificus* and is a target of the Hox gene *lin-39*. *Development* **134**, 3111–3119 (2007).
9. Tian, H., Schlager, B., Xiao, H. & Sommer, R.J. Wnt signaling by differentially expressed Wnt ligands induces vulva development in *Pristionchus pacificus*. *Curr. Biol.* **18**, 142–146 (2008).
10. Herrmann, M. *et al.* The nematode *Pristionchus pacificus* is associated with the oriental beetle *Exomala orientalis* in Japan. *Zoolog. Sci.* **24**, 883–889 (2007).
11. Srinivasan, J. *et al.* A bacterial artificial chromosome-based genetic linkage map of the nematode *Pristionchus pacificus*. *Genetics* **162**, 129–134 (2002).
12. Mayer, W.E., Herrmann, M. & Sommer, R.J. Phylogeny of the nematode genus *Pristionchus* and implications for biodiversity, biogeography and the evolution of hermaphroditism. *BMC Evol. Biol.* **7**, 104 (2007).
13. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**Suppl 2, ii215–ii225 (2003).
14. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
15. Majoros, W.H., Pertea, M. & Salzberg, S.L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
16. Alexeyenko, A., Tamas, I., Liu, G. & Sonnhammer, E.L.L. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics* **22**, e9–e15 (2006).
17. Finn, R.D. *et al.* The Pfam protein family database. *Nucleic Acids Res.* **36**, D281–D288 (2008).
18. Kanehisa, M. *et al.* KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36**, D480–D484 (2008).
19. Stein, L.D. *et al.* The genome sequence of *Caenorhabditis briggsae*: A platform for comparative genomics. *PLoS Biol.* **1**, 166–192 (2003).
20. Oesch, F. & Arand, M. Xenobiotic metabolism. in *Toxicology* Marquardt, H. *et al.* (eds.) Academic Press, San Diego, pp. 83–107 (1999).
21. Smant, G. *et al.* Endogenous cellulases in animals: isolation of β-1,4-endoglucanase genes from two species of plant-parasitic cyst nematodes. *Proc. Natl. Acad. Sci. USA* **95**, 4906–4911 (1998).
22. Kikuchi, T., Jones, J.T., Aikawa, T., Kosaka, H. & Ogura, N. A family of glycosyl hydrolase family 45 cellulases from the pine wood nematode *Bursaphelenchus xylophilus*. *FEBS Lett.* **572**, 201–205 (2004).
23. Keen, N.T. & Roberts, P.A. Plant parasitic nematodes: digesting a page from the microbe book. *Proc. Natl. Acad. Sci. USA* **95**, 4789–4790 (1998).
24. Tanaka, H. *et al.* Insect diapause-specific peptide from the leaf beetle has consensus with a putative iridovirus peptide. *Peptides* **24**, 1327–1333 (2003).
25. Dunning Hotopp, J.C. *et al.* Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science* **317**, 1753–1756 (2007).
26. Blaxter, M. Symbiont genes in host genomes: fragments with a future. *Cell Host Microbe* **2**, 211–213 (2007).
27. Blaxter, M. *et al.* A molecular evolutionary framework for the phylum Nematoda. *Nature* **392**, 71–75 (1998).
28. Weischer, B. & Brown, D.J.F. *An Introduction to Nematodes* (Pensoft, Moscow, 2000).
29. Poulin, R. *Evolutionary Ecology of Parasites* (Princeton University Press, Princeton, New Jersey, 2007).
30. Huang, X. *et al.* Application of a superword array in genome assembly. *Nucleic Acids Res.* **34**, 201–205 (2006).
31. Bao, Z. & Eddy, S.R. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.* **12**, 1269–1276 (2002).
32. Huang, X. & Madan, A. CAP3: A DNA sequence assembly program. *Genome Res.* **9**, 868–877 (1999).
33. Slater, G.S.C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
34. Allen, J.E., Majoros, W.H., Pertea, M. & Salzberg, S.L. JIGSAW, GeneZilla, and GlimmerHMM: puzzling out the features of human genes in the ENCODE regions. *Genome Biol* **7**Suppl 1, S9.1–S9.13 (2006).
35. Remm, M., Storm, C.E. & Sonnhammer, E.L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **314**, 1041–1052 (2001).
36. Enright, A.J., Dongen, S.V. & Ouzounis, C.A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
37. Sanderson, M.J. A nonparametric approach to estimating divergence times in the absence of rate constancy. *Mol. Biol. Evol.* **14**, 1218–1231 (1997).