The genome of a subterrestrial nematode reveals adaptations to heat

Deborah J. Weinstein1, Sarah E. Allen1,7, Maggie C.Y. Lau2,3, Mariana Erasmus4, Kathryn C. Asalone1, Kathryn Walters-Conte1, Gintaras Deikus5, Robert Sebra5, Gaetan Borgonie6, Esta van Heerden4,8, Tullis C. Onstott2 & John R. Bracht1*

The nematode *Halicephalobus mephisto* was originally discovered inhabiting a deep terrestrial aquifer 1.3 km underground. *H. mephisto* can thrive under conditions of abiotic stress including heat and minimal oxygen, where it feeds on a community of both chemolithotrophic and heterotrophic prokaryotes in an unusual ecosystem isolated from the surface biosphere. Here we report the comprehensive genome and transcriptome of this organism, identifying a signature of adaptation: an expanded repertoire of 70 kilodalton heat-shock proteins (Hsp70) and avrRpt2 induced gene 1 (AIG1) proteins. The expanded Hsp70 genes are transcriptionally induced upon growth under heat stress, and we find that positive selection is detectable in several members of this family. We further show that AIG1 may have been acquired by horizontal gene transfer (HGT) from a rhizobial fungus. Over one-third of the genes of *H. mephisto* are novel, highlighting the divergence of this nematode from other sequenced organisms. This work sheds light on the genomic basis of heat tolerance in a complete subterrestrial eukaryotic genome.
**H. mephisto** was discovered inhabiting a fluid-filled aquifer accessed from the Beatrix Gold Mine in South Africa at 1.3 km below the surface. Radiocarbon dating indicates the aquifer water is over 6000 years old, and the lack of surface 3H infiltration, a remnant of atmospheric atomic testing, highlights its isolation from the surface biosphere. The water is warm (37 °C), alkaline (pH 7.9), hypoxic (0.42–2.3 mg/L dissolved O2), and rich in biogenic methane (CH4)3–5. In spite of these challenging conditions, a thriving, complex microbial community exists in this extreme environment, including chemolithoautotrophic organisms that extract energy from the subterrestrial rock and fix inorganic carbon24,25. Syntrophic relationships link sulfur-oxidizing denitrifying bacteria, sulfate reducers, methanogens, and anaerobic methane oxidizing organisms into a complex mutually reinforcing microbial food web that supports a rich assemblage of eukaryotic opportunistic predators, including nematodes, rotifers, and protists5. With the exception of *H. mephisto* none of these eukaryotic organisms have been cultured in the laboratory, and none have had their genomes sequenced and analyzed until now. Nematodes encode small, remarkably dynamic genomes well suited to studies of adaptation6,7. Among the most abundant animals on earth, nematodes have adapted to an incredibly diverse set of environments: from hot springs to polar ice, soil, fresh, and saltwater8, acid seeps9, and the deep terrestrial subsurface1, with a wealth of comparative genomic data available. Dynamic gene family expansion8,10 and shrinkage11 have proven good signatures of evolutionary adaptive selection in crown eukaryotes, including nematoda12. Here, we perform comprehensive genomic and transcriptomic studies in *H. mephisto*, revealing the evolutionary adaptive response to a subterrestrial environment, including expanded gene families and patterns of expression under heat stress.

**Results**

**Assembly, annotation, and phylogenetic analysis.** De novo DNA sequence assembly with Illumina data and scaffolding with PacBio reads, yielded a small complete assembly of 61.4 Mb comprising 880 scaffolds with N50 of 313 kb, though 90% of the sequence is encoded on just 193 scaffolds. The longest scaffold is just under 2.55 Mb (Table 1). Several lines of evidence suggest that this is a highly complete genome. The Core Eukaryotic Genes Mapping Approach (CEGMA)13,14 identified 240 of 248 core eukaryotic genes for a completeness of 97%, and tRNAscan-SE15 identified 352 tRNAs encoding all 20 amino acids plus selenocysteine. Benchmarking Universal Single Copy Orthologs (BUSCO)16 estimated the completeness as 81.4%, but manual inspection of its output shows that 79 apparently missing genes are actually detected, with good e-values (median 6e-19). Given that BUSCO’s thresholds are established from eight nematode genes from Clades I, III, and V17, it may not be well suited to divergent Clade IV nematodes such as *H. mephisto*, since it also scored the *P. redivivus* genome (98% complete)18 as only 82.1% complete. Therefore, considering these divergent matches to be valid, we conclude that for *H. mephisto* BUSCO detected 946/982 orthologous genes, for a completeness of 96%, consistent with CEGMA. Reinforcing the completeness of the *H. mephisto* assembly and annotation, a quantitative comparison of 3,252 protein domains shows a strong correlation with *Caenorhabditis elegans* (Fig. 1b).

The repetitive component of the *H. mephisto* genome is highly divergent. Using a custom RepeatModeler repeat library, RepeatMasker masked 24.3% of the genome, denoting 21.1% as interspersed repeats, 87.3% of which are unknown (Table 2). Evaluation of these sequences with nhmmer19 and the DFAM database positively identified 44.3% of these as helitrons, with 8.8% retrotransposons and 23.6% DNA transposons (Table 2). However, consistent with the genomic divergence of *H. mephisto*, this repetitive element repertoire appears extremely different from known elements, including many unique or novel repeat families needing further characterization, and a significant 23.3% remain unclassified by either algorithm (Table 2).

The *H. mephisto* nuclear genome was annotated with Maker20, TopHat1, StringTie22, and TransDecoder23, and transcripts were clustered with gffcompare to define a total of 16,186 protein-coding loci. An additional 1,023 loci were identified that do not encode proteins over 50 amino acids and are candidate noncoding transcripts, for a total of n = 17,209 transcribed loci in the *H. mephisto* genome, producing 34,605 transcripts for an average of 2.0 transcripts per locus (Table 1). Intron lengths averaged 473 compared to 320 bp for *C. elegans* (Table 1).

We used single-copy orthologous proteins (SCOOGs) to build a phylogenetic tree placing *H. mephisto* as a distant relative of free-living *Panagrolaimus redivivus*, the nearest fully sequenced nematode relative, within Clade IV (Fig. 1a). The comparison of domain counts between *C. elegans* and *H. mephisto* uncovered two domains strikingly enriched in *H. mephisto*: avrRpt2-induced gene, AIG1 (84 domains vs. 0) and 70-kD heat-shock protein, Hsp70 (126 domains vs. 15) (Fig. 1b). The most over-represented domains in *C. elegans* are the Meprin And TRAF-Homology (MATH) domain, Fog-2 Homology Domain (FTH) and F-box associated (FBA_2) domains, all apparent lineage-specific expansions in *Caenorhabditis*.24,25 (Fig. 1b). We used OrthoVenn26 to identify orthologous genes between *D. melanogaster*, *C. elegans*, *P. redivivus*, and the 16,186 *H. mephisto* proteins, identifying a set of 5,397 shared among all nematodes, with 3,233 shared among all four invertebrates. Among the 417 *H. mephisto*-specific orthologous groups, the largest cluster was Hsp70, with 107 proteins.

**Hsp70 expansion.** The expansion of Hsp70 is particularly evo-

cative because these well-studied heat-activated chaperones refold proteins denatured by heat27–30 as part of a coordinated response to heat-shock31–34. Even more intriguing, Hsp70 are expanded in organisms adapted to environmental thermal stress, but not in nematodes parasitic on endothermic hosts, suggesting the expansion of Hsp70 may be a general strategy for adaptation to environmental, not parasitic, heat (Fig. 1c). Bayesian phylogenetic analysis of Hsp70 proteins recovered known paralogs specific to cellular compartments including mitochondrial and endoplasmic reticulum (ER)35 along with a cluster grouping human, mouse, and nematode genes (Cluster I, Fig. 2a). The recovered Hsp70 gene tree topology is robust, given that the same structure was recovered by maximum likelihood (Supplementary Fig. 1). The human sequences in Cluster I, include well-characterized Hsp70 sequences36. Cluster II is a new 37-member *H. mephisto*-only group, which surprisingly is most closely related to another novel *Diploscapter* cluster with 59 genes (Cluster III, Fig. 2a). These data suggest that the Hsp70 gene family has undergone significant amplification within the *Diploscapter* and *Halicephalobus* lineages which, owing to their evolutionary distance (Fig. 1a), most likely did not inherit these expanded gene families from a common ancestor. Instead, we propose these genes underwent independent expansions in both lineages under shared evolutionary pressure to adapt to heat stress: *Diploscapter pachys* is thermotolerant37, while *Diploscapter coronatus* is a facultative parasite of humans38 and a member of the genus has been found in thermal waters39.

A signature of adaptive evolutionary change is positive selection, in which mutations altering amino acids (dN) are
enriched relative to (presumably neutral) synonymous changes (dS), giving a dN/dS ratio (ω) greater than 1. A value of ω less than 1 indicates elimination of mutations that alter the amino acid sequence, also known as purifying selection, and occurs when protein function is preserved by evolution.

Given the long branch lengths, gene family expansions, and potential for selection acting on Hsp70 gene function, we used PAML to test for statistical evidence of positive selection. To facilitate this analysis, we created a new Bayesian phylogeny using a subset of genes from H. mephisto, the Diploscapter species, and C. elegans outgroups, resulting in a well-resolved Bayesian phylogeny (Fig. 2b). We ran a branch-sites test in PAML, which estimates two ω parameters for different codons on a preselected foreground branch of the phylogenetic tree. The ω1 parameter accounts for pervasive purifying selection at specific sites (codons) but a second ω2 measures values greater or equal to 1 (from neutral to positive selection) at other sites. Because ω2 is estimated from the data, it is an indicator of positive selection if it is reported to be above 1. Furthermore, by performing the test twice, once with ω2 fixed at 1 (neutrality, a null model) and once allowing it to be freely estimated from the data, a likelihood ratio test (LRT) can be used to derive a p value quantifying the strength of positive selection on a particular branch. In our analysis, positive selection was detected along the long branch leading to the H. mephisto cluster: ω2 of 197 on 89% of amino acids, p = 0.04 (Table 3), however, it was not robust to Bonferroni correction for multiple hypothesis testing, which may be too conservative. Given that this branch has by far the most sites (89%) under positive selection, we have made it semi-bold (Fig. 2b). Values of ω2 greater than 1 were detected on ten of eleven branches from clusters II and III (Table 3) with four giving a p value that is significant after correction for multiple hypothesis testing (lines strongly bolded in Fig. 2b).

In particular, the root of the Diploscapter cluster (Fig. 2b, branch K) showed evidence of strong positive selection (Table 3). As controls, we tested a Cluster I short branch (L) and mitochondrial gene (M), which showed no evidence of positive selection (Fig. 2b, Table 3). Together these data suggest that positive selection is detectable in specific Hsp70 lineages in both Halicephalobus and Diploscapter.

AIG1 expansion. AIG1 was originally identified as a pathogen response gene in plants, but is also involved in survival of T cells in mammals, where they were named immune-associated nucleotide binding proteins (IANs), or GTase of immunity-associated proteins (GIMAPs). These proteins function as GT- binding molecular switches controlling cell fate. AIG1 was originally reported to be completely absent from nematodes, but by relaxing the statistical stringency we find a single copy of the domain (Y67D2.4) in C. elegans annotated as a homolog of human mitochondrial ribosome associated GTase 1 (MTG1), suggesting a possible divergence and expansion of the GTase superfamily in H. mephisto and other nematodes. Consistent with this, blastp against the nr database and HMMPER search of Uniprot reference proteomes identified hundreds of matches with low percent identity (~30%) to the AIG1s identified in H. mephisto. Nonetheless, the matches range from 250 to 300 amino acids in length and with e-values from 1e-20 to 1e-30; they come from species as diverse as nematodes, fungi, arthropods, and the parasitic protist Giardia intestinalis. These sequences are generally annotated as uncharacterized or hypothetical proteins, though some are annotated as p-loop containing nucleoside triphosphate hydrolases, suggesting a large GTase protein family, previously uncharacterized, resides within eukaryotes.

Consistent with the relatively low percent identity of these genes, most did not align well with the H. mephisto sequences, and if they did align, they did not show good bootstrap support in phylogenetic trees. This suggests the eukaryotic superfamily of GTases is comprised of distinct and divergent subfamilies, only one of which is the AIG1 domain expanded in H. mephisto. We ultimately were able to obtain good alignments and well-supported phylogenetic trees from only 17 nematode species: an AIG1-like cluster including H. mephisto, Diploscapter coronatus, D. pachys, A. suum, and P. redivivus, while a separate MTG1-related cluster includes the Caenorhabditis sequence and other nematodes (Fig. 3a). Notably, the two distinct clusters resolve with 100% bootstrap support, and while the MTG1-related cluster includes members of all sequenced nematode clades (I, III, IV, and V; no Clade II genomes are currently available), the AIG1-like group includes only members of clades III–V, suggesting a potential origin in the Chromadorea and extensive amplification in H. mephisto (Fig. 3a).

In order to relate these nematode GTase subfamilies with the previously identified AIG1/IAN/GIMAP sequences from vertebrates and plants, we built a tree including a broader taxonomic sampling beyond nematodes (Fig. 3b). To simplify the tree, we included only six nematodes in this analysis, five of which we previously characterized as having the AIG1-like genes: H. mephisto, P. redivivus, A. suum, D. coronatus, and D. pachys, and the C. elegans representative of the MTG1-related proteins. In this tree the original IAN/GIMAP human–plant cluster was recovered but deeply rooted from the MTG1-related side of the phylogeny with good branch support (Fig. 3b).

We found two sequences from Rhizophagus irregularis, a plant–root associated fungus, resolve cleanly into the newly discovered AIG1-like group including H. mephisto (Fig. 3b). The R. irregularis sequences are from two single-nucleus genome

| Table 1 Genome and transcriptome data for H. mephisto and comparison to C. elegans, M. hapla, and P. redivivus |
|-------------------------------------------------|-------------------------------------------------|-------------------------------------------------|-------------------------------------------------|
| **H. mephisto** | **C. elegans** | **M. hapla** | **P. redivivus** |
| Assembly size (Mb) | 61.4 | 100 | 53.0 | 64.4 |
| NSO (kb) | 313 | 17,494 | 38 | 262 |
| # of scaffolds (kb) | 880 | 7 | 3,452 | 940 |
| Longest (kb) | 2,546 | 20,924 | 360 | 2,280 |
| Shortest (kb) | 1 | 13.8 | 0.7 | 0.4 |
| Protein-coding loci (nonredundant) | 16,186 | 19,922 | 14,420 | 24,249 |
| Potential noncoding loci | 1,023 | 703 | - | - |
| Average intron length (bp) | 473 | 320 | 153 | 163 |
| Average exon length (bp) | 332 | 202 | 171 | 288 |
| Transcripts | 34,605 | 33,303 | 16,676 | 26,372 |
| Repetitive content (%) | 24.3 | 12.7 | 18.3 | 7.1 |
| GC content (%) | 32.1 | 35.4 | 27.4 | 44.3 |

Data for C. elegans and P. redivivus from Srinivasan et al. and for M. hapla from Foth et al.
Fig. 1 Genomic comparison of H. mephisto protein-coding genes. a Multilocus phylogeny of H. mephisto using 99 single-copy orthologous genes (SCOOGS). b Pfam domain comparison of nonredundant protein domain content in C. elegans versus H. mephisto at an e-value cutoff of 1e-10. Expanded AIG1 and Hsp70 domain families are marked, along with the MATH, FTH, and FBA_2 domains known C. elegans-specific expansions. c Examination of Hsp70 family expansion across species, quantifying proteins, not domains, identified using e-values as in (b). d Venn diagram comparing orthologous gene clusters between D. melanogaster, C. elegans, P. redivivus, and H. mephisto. Abbreviations: MATH meprin and TRAF-homology domain, FTH Fog-2 homology domain, FBA_2 F-box associated domain
isoform switch, with a longer version, MSTRG.2841.1, expressed cuticle components and ornithine-oxo-acid transaminase. Their data suggest that, in contrast to reports of AIG1 being absent in a member of Spirurina and separated from the Rhabditina, 80 have been controversial, transfer from fungus to fungus sequences, accessions EXX68593.1 and EXX68414.49. These data suggest that, in contrast to reports of AIG1 being absent in invertebrates,45,50 Clade III-V nematodes do have a previously unknown member of this protein family potentially derived from an ancient horizontal gene transfer from a rhizobial fungus. D. pachys has been reported to inhabit the rhizobial zone of plant roots,37,31 where it is ideally situated to acquire horizontally transferred genes from cohabiting fungus. Given that we found five nematode species host orthologs of the AIG1-like domain in their genomes (Fig. 3a) the HGT event must have been in a Chromadorean ancestor lineage, not in contemporary D. pachys, though H. mephisto has undergone a spectacular expansion of this gene family that is not present in the other lineages, most of which have only a few copies (Fig. 3a). The divergence of suborders Rhabditina (C. elegans, D. pachys, D. coronatus) and Tylenchina (P. redivivus and H. mephisto) has been estimated at 22 million years ago using molecular clock methods.52. A. suum is a member of Spirurina and separated from the Rhabditina 80 million years ago,53 and the putative HGT event must predate these divergences. While inter-domain HGT into eukaryotes has been controversial, transfer from fungus to C. elegans has been documented, setting a precedent for this hypothesized HGT event.54 Preliminary analysis of codon bias to test for HGT was documented, setting a precedent for this hypothesized HGT event54. Therefore, under hot conditions it may be vital for H. mephisto to co-express Hsp70 and ARMET, particularly if they synergize in responding to the damage due to heat in different cellular compartments, Hsp70 in cytosol and ARMET in ER. It appears that H. mephisto has developed different genomic strategies for upregulating these two genes: mild upregulation of many paralogous genes (Hsp70) versus extremely highly induced expression of a single-copy gene (ARMET). Regardless, ARMET is a strong marker of ER stress with cytoprotective roles.58,60,61.

Another upregulated transcript was Bax Inhibitor-1 (BI-1), which was upregulated fourfold (Fig. 4b). BI-1 is a conserved antiaptopotic protein that prevents ER stress-induced apoptosis.62 BI-1 interacts, binds and suppresses IRE1α activity by canceling its endoribonuclease and kinase activity to promote cell survival.63 As an ER stress pro-survival factor,62 BI-1 potentially compensates for the lack of heat induction of AIG1, which plays a similar role inhibiting apoptosis,63 but may be more specifically tuned to the UPR response.

Table 2 Repetitive element composition of the H. mephisto genome

| Class          | RepeatMasker | nhmmer |
|----------------|--------------|--------|
| Retrotansposons| 3.8          | 8.8    |
| DNA transposons| 8.8          | 23.6   |
| Helitrons      | 0.0          | 44.3   |
| Unclassified   | 87.3         | 23.3   |

The percentage of repeats belonging to each class is shown.

Genes differentially expressed under heat stress. We performed a comprehensive differential expression analysis of H. mephisto transcripts whose expression changes under heat stress, and found that while Hsp70 transcripts are statistically upregulated on exposure to heat, AIG1 transcripts are not (Fig. 4a). Analyzing all transcripts, we identified 285 heat-upregulated and 675 heat-downregulated (Fig. 4b). Because many upregulated transcripts were unknown, and the upregulated set is small, Gene Ontology analysis identified no enriched categories. However, the downregulated set of transcripts was enriched in peptides, as well as cuticle components and ornithine-oxo-acid transaminase (Table 4 and Fig. 4b). The downregulation of cuticular component is entirely driven by 13 collagens (out of 89 genes in the genome).

As noted, Hsp70 transcripts respond to heat (Fig. 4a) but with two exceptions they do not make the twofold cutoff, and one is actually statistically inhibited on heat stress (Fig. 4b). One AIG1 transcript is included in the statistically downregulated set (Fig. 4b). An intriguing finding is SAX-2 (Sensory AXon guidance 2), a protein involved in neuronal development in C. elegans,56, and whose inactivation causes a variety of developmental, cellular, and behavioral phenotypes.57. In H. mephisto heat stress causes an isoform switch, with a longer version, MSTRG.2841.1, expressed exclusively at 25°C and a shorter version, MSTRG.2841.2, exclusively expressed at 38–40°C. Thus, sax-2 is identified as a transcript both heat-upregulated (12.9-fold) and heat-downregulated (7.0-fold) on the volcano plot (Fig. 4b). The long isoform is 9,286 bp and encodes a 3,016 amino acid (aa) protein, with the alternative isoform being 132 basepairs shorter owing to an 84 bp change in transcriptional start site and alternate 3’ splice site choice for exon 10 (of 16 exons total). These isoform differences only result in a 16 aa deletion from the predicted protein. Given that the protein is ~3,000 aa, and the 16 aa change does not significantly alter the encoded domains (identified by HMMER as MOR2_PAG1 N, mid, and C-terminal domains), the functional implication of the transcriptional shift and alternative splicing of SAX-2 remains to be investigated in future work.

Discussion

As the founding genome sequence of the Halicephalobus genus, H. mephisto illuminates previously unexplored territory. H. mephisto separated from Caenorhabditis at least 22 million years ago,52,64 and likely over 100 million years ago,65 though calibrating the nematode molecular clock is difficult because of their poor fossil record.64. Our data suggest that nematodes access the deep subsurface from surface waters facilitated by seismic activity.66. This transition from surface to deep subsurface would be expected to exert strong selective pressures on their genomes, which in nematodes are particularly evolutionarily dynamic.6,7,10–12. Therefore, we speculate that the H. mephisto divergence may reflect selection more than neutral genetic drift, making it a particularly informative genome. Consistent with this, the expanded Hsp70 and AIG1 gene families are extremely divergent from earlier exemplars, residing on extended branches in phylogenetic analysis (Figs. 2 and 3), and positive selection was detected along several lineages of the Hsp70 phylogeny (Fig. 2b).

We investigated the novelty of H. mephisto’s 16,186 protein-coding genes by a combination of domain search (HMMER against the Pfam-A database), blastp against the manually curated high-confidence Uniprot-Swissprot database, blastp against the combined proteomes of 28 nematode species (listed in Methods), and Interproscan 5 analysis. At an e-value of 1e-4 we found 10,567 proteins identified by one or more of these methods, leaving 5,619 unknown genes (34.7% of 16,186) lacking domains.
Fig. 2 Analysis of Hsp70. a Bayesian phylogenetic tree of Hsp70. H. mephisto sequences marked with an asterisk (*) and D. pachys with arrows. Branch numbers indicate posterior probabilities; scale bar represents substitutions per site. b Hsp70 protein Bayesian tree used for dN/dS analysis of coding sequence. Branch letters A–M correspond to Table 3 for $\omega$ (dN/dS) value. Bold branches indicate statistical significance of dN/dS $p$ value after correcting for multiple hypothesis testing, and the long branch (F) is semi-bold because it is statistically significant prior to multiple testing correction with 89% of sites under positive selection (Table 3). Branch numbers indicate posterior probabilities; scale bar represents substitutions per site. The sequences used in both trees are indicated with Wormbase (nematode) or Genbank Accessions (other species).
or any recognizable homology. Nevertheless, 3,599 (64.0%) of the unknown genes are expressed (defined as having at least 5 FPKM across 12 replicates) increasing our confidence they are real genes. Genes whose expression was not detected in our analysis may be expressed at extremely low levels (below 5 FPKM across all replicates) or be expressed under different environmental conditions than the laboratory culture we employed. We conclude that these 5,619 completely unknown genes, and particularly their 3,599 expressed subset, are intriguing candidates for functional adaptation to the deep terrestrial subsurface.

These data are consistent with a previous report that new nematode genomes tend to yield around 33% proteins that are unrecognizable outside their genus, given that H. mephisto is the first of its genus to be sequenced fully. As a control we examined the proteome of P. redivivus, using the identical protein-identification pipeline, finding 33.8% unknown genes, quite similar to the numbers for H. mephisto. We also tested the pine wood nematode, Bursaphelenchus xylophilus, and identified 25.6% unknown genes. Like H. mephisto, P. redivivus and B. xylophilus are the first genomes of their respective genera to be sequenced. When a within-genus comparison is available, the number of novel genes has been reported to drop to around 10%, which we tested by examining the proteins of Meloidogyne hapla, which has a within-genus match to M. incognita in the 28 nematode blast database. Consistent with predictions, we identified 8.4% unknown proteins in M. hapla. Therefore, we conclude that while the genomic plasticity we observe for H. mephisto is significant, the number of unknown genes is broadly consonant with nematode molecular systematics showing that roundworm genomes are extremely dynamic. The novelty of the H. mephisto genes reflect a combination of evolutionary adaptation and a lack of closely related comparative Halicephalobus species in databases. Supporting this, of the 1,730 genes that match nematode genome(s) only, and were not identified by Interproscan, Uniprot-Swissprot, or Pfam, 1,480 of them match P. redivivus (Supplementary Fig. 2A), the nearest sequenced nematode relative of H. mephisto (Fig. 1a). These 1730 genes are not widely conserved even among roundworms, with only 17 (1%) of them identified in all 28 species and most (511) identified in only one other species (Supplementary Fig. 2B); unsurprisingly the majority of these (399, 78%) were only found in P. redivivus.

The assembled genome of H. mephisto is smaller (61.4 Mb) than most other sequenced nematode genomes with the exception of M. hapla (54 Mb) and M. incognita (47–51 Mb), though similar in size to the most closely related species, P. redivivus (64.4 Mb). H. mephisto reproduces via parthenogenesis, while both M. hapla and M. incognita are facultatively parthenogenetic. In Caenorhabditis, loss of males has been linked to genome shrinkage. In M. incognita and D. pachys, the loss of sexual reproduction leads to functional haploidization as alleles diverge into paralogs across the genome (leading to most genes being present as duplicate, divergent copies). The conversion of a diploid into a functional haploid genome is associated with three predominant changes: (1) a high degree of heterozygosity as lack of recombination leads to high divergence between alleles, now paralogs, (2) assembly of a haploid genome, and (3) detection of two copies of most genes that are single-copy in other organisms.

The heterozygosity we observed in H. mephisto is modest: the kmer frequency distribution from Illumina reads shows two prominent peaks consistent with approximately 1% heterozygosity (Supplementary Fig. 3A) and mapping reads back to the assembly identified 707,190 snps and 55,683 indels (762,873 total variants) confirming an overall SNP heterozygosity of 1.15%. In contrast, D. pachys displays 4% heterozygosity. The H. mephisto genome assembly we obtained is largely haplotype-merged: mapping the reads back to the genome shows that 59.7 Mb of the assembled sequence (97%) is at haplotype-merged coverage (10×) and only 1.7 Mb (3%) exists as potentially diverged haplotypes at lower coverage (Supplementary Fig. 3B). In agreement, CEGMA analysis of assembled reads (Supplementary Fig. 3C) and fragments (Supplementary Fig. 3D) are predominantly single peaks at approximately 100× coverage.

Allele-to-paralog conversion results in two recognizably distinct gene copies, or paralogs, in the genome. In the D. pachys genome CEGMA found an average of 2.12 copies of each core eukaryotic ortholog yet in H. mephisto CEGMA reported an average 1.19 copies of each gene while BUSCO identified 97% single-copy genes. We therefore conclude that H. mephisto exhibits very little functional haploidization into paralogs, and it may represent an early evolutionary stage in this process.

It is important to verify that amplified Hsp70 and AIG1 gene families do not represent assembly errors. One way this could happen is if an assembler produces multiple overlapping contigs encoding the same locus, leading to artifically high gene copy numbers. In that case multiple near-identical copies of the genes would be present. To test this, we extracted the corresponding nucleotide sequences for each family and performed within-family, all-vs.-all blastn at an evalue of 1e-4 and filtered for nonself matches. These nonself matches were relatively divergent.

### Table 3 Branch-site analysis of dN/dS ratios (ω) from tree in Fig. 2b

| Branch          | ω2 (positive selection) | ω1 (purifying selection) | Fraction sites under ω2 | Fraction sites under ω1 | LRT statistic (2×lnL) | p Value |
|-----------------|-------------------------|--------------------------|-------------------------|-------------------------|-----------------------|---------|
| A               | 3.58                    | 0.17                     | 0.12                    | 0.81                    | 1.98                  | 0.15939 |
| B               | 2.94                    | 0.17                     | 0.05                    | 0.88                    | 1.6                   | 0.205903|
| C               | 1.54                    | 0.17                     | 0.11                    | 0.82                    | 0.22                  | 0.63904 |
| D               | 94.01                   | 0.17                     | 0.01                    | 0.92                    | 2.4                   | 0.121335|
| E               | 86.51                   | 0.17                     | 0.24                    | 0.68                    | 104.48                | 1.59e-24 |
| F               | 196.81                  | 0.17                     | 0.89                    | 0.04                    | 4.16                  | 0.041389|
| G               | 999                     | 0.17                     | 0.08                    | 0.84                    | 8.12                  | 0.004378|
| H               | 18.27                   | 0.17                     | 0.09                    | 0.84                    | 11.68                 | 0.000632|
| I               | 1                       | 0.17                     | 0.75                    | 0.38                    | 0                     | 1       |
| J               | 12.04                   | 0.17                     | 0.31                    | 0.62                    | 14.46                 | 0.000143|
| K               | 999                     | 0.17                     | 0.13                    | 0.80                    | 26.36                 | 2.83e-7  |
| L               | 1                       | 0.17                     | 0                        | 0.93                    | 0                     | 1       |
| M               | 1                       | 0.17                     | 0.40                    | 0.52                    | 0                     | 1       |

Bold p values are statistically significant after correcting for multiple hypothesis testing (p = 0.0038). Likelihood ratio test (LRT) statistic applied to a χ2 table with df = 1, critical values 3.84 (5%) and 6.63 (1%). Note that ω1 is constrained to be greater than or equal to 1 (positive to neutral selection). Δ, expression increased 1.5× under heat stress. ΔΔ, expression increased 2× under heat stress. LnL = log likelihood from PAML output.
the best nonself within-family blastn match for 112 Hsp70 loci averaged 86.4% identity at the nucleotide level (Supplementary Fig. 4A) and for 63 AIG1 loci averaged 89.7% identity (Supplementary Fig. 4B). Given the observed genomic heterozygosity is only 1.15%, these data suggest that the Hsp70 and AIG1 genes are diverged paralogs and neither allelic copies nor redundant misassemblies. As a control we extracted 65 collagen genes and performed the same analysis. As might be predicted for an ancient and highly divergent gene family, 74% of collagen genes lacked within-family nonself blast matches entirely, though those that matched averaged 86.6%, similar to Hsp70 (Supplementary Fig. 4C). These sequence divergences are greater than a control analysis performed by blast of the assembly to itself (Supplementary Fig. 4D). We conclude that the expanded Hsp70 and AIG1 families represent paralogy rather than assembly redundancy artifact.

Fig. 3 Phylogenetic analysis of AIG1. a Nematode-only RAxML tree of 17 species showing two clusters: MTG1-related and AIG1-like groups. b Nematode, human, Arabidopsis, and fungal RAxML tree illustrating potential HGT event. IAN/GIMAP are synonyms of AIG1, the original names given to plant and vertebrate sequences. Two fungal (Rhizophagus irregularis) AIG1-like sequences highlighted in light red boxes. For both trees, branch numbers indicate bootstrap support from 200 replicates, and scale bar represents substitutions per site. The sequences used in both trees are indicated with Wormbase (nematode), UniProtKB (Rhizophagus irregularis) or Genbank Accessions (other species).
Often repetitive elements are collapsed by assemblers, so we checked that Hsp70 and AIG1 are not collapsed repeats, which would cause under-representation of their family diversity. Mapping raw reads onto the assembled genome indicates the collapsed repeats as regions of elevated coverage. By mapping the raw reads back to the *H. mephisto* genome we found that the coverage of both Hsp70 and AIG1 families are not elevated relative to the entire genome (Supplementary Fig. 3E). Overall, we

**Fig. 4** Transcriptome analysis of gene expression in *H. mephisto*. a Boxplot showing that Hsp70 transcripts are induced on heat while AIG1 transcripts are unchanged. Box shows median (center line) and first and third quartiles, while whiskers indicate the 15th to 85th percentiles, and notches represent confidence intervals. *p* values obtained by two-tailed Mann-Whitney test. Source data are provided as a Source Data file. b Volcano plot of gene expression fold change comparing heat (combined replicates: 38 °C, n = 3, and 40 °C, n = 6) versus 25 °C control (replicates, n = 3). Statistically altered transcripts, defined as q value less than 0.05 and upregulated or downregulated at least twofold under heat-stress conditions (38–40 °C) relative to 25 °C controls, are indicated with luminosity of 1 while nonsignificant are luminosity 0.1. The q value was obtained from the `stattest()` function within the Ballgown R package. Genes labeled “Unknown” are novel, as discussed in the text. Proteins encoding no recognizable domain but matching other nematodes by blastp are indicated with the number of nematode species matched (out of the 28 used in constructing the blast database). For genes matching only a single other nematode, the species is given in parenthesis. Color key: red: Hsp70, blue: AIG1, green: ARMET, and magenta: peptidases. **Abbreviations**: Sax-2 *sensory axon guidance 2*, LSM-like SM. BI-1, Bax Inhibitor-1, ARMET arginine-rich, mutated in early-stage tumors. Source data are provided as a Source Data file.
conclude that the Hsp70 and AIG1 genes identified in this study are true paralogs, neither over nor underrepresented in the genome.

*D. pachys* is remarkable for having fused its chromosomes together into a single linkage group and eliminating telomeres. However, over 6,000 telomeric repeat-containing reads (at least four copies of TTAGGGC, the *C. elegans* telomeric repeat (72) were present in the raw Illumina data from *H. mephisto*. By extracting read pairs with telomeric repeats in at least one read, merging them with PEAR (73), and assembling them with MIRA we were able to identify 7 unique subtelomeric regions, suggesting that while the number of *H. mephisto* chromosomes may be reduced relative to *C. elegans*, they are not fused as in *D. pachys*. Consistent with this, we find two homologs of the telomeric gene Protection of Telomeres (POT1) in *H. mephisto* relative to the three in *C. elegans*, all of which are lost in *D. pachys* (69). Also in contrast to *D. pachys*, we were able to identify Telomerase Reverse Transcriptase (TERT) in *H. mephisto*, suggesting that standard telomeres have been retained in the subterrestrial organism and chromosome fusion is not an inevitable consequence of a parthenogenetic lifestyle.

Gene expression provides clues to the adaptation to the warm subterrestrial environment. Among transcripts whose expression changed significantly on exposure to heat, over twice as many are downregulated (675) as upregulated (285) (Fig. 4b). We suggest two potential explanations of this phenomenon: the worms may find lower temperatures more stressful given their native conditions are warmer, and therefore activate more genes at 25 °C relative to 38–40 °C. Or, the downregulation of genes may be itself an adaptation to heat, an idea consistent with the regulated IRE1α dependent decay (RIDD) pathway of the unfolded protein response (UPR) (74,75). When the RIDD pathway is activated, degradation of ribosome-bound transcripts is mediated by the endoribonuclease domain of IRE1α, relieving the immediate protein synthesis demand on the ER, and providing existing proteins time to refold (63,74). Our data cannot distinguish these two theories; however, the elevated expression of protein chaperones like Hsp70 under heat exposure (Fig. 4a) supports the model in which observed changes in transcriptional profile reflect an adaptive response to higher, rather than lower, temperatures. Combined with the observed heat-induced expression of the antiapoptotic factor Bax Inhibitor 1 (BI-1) (Fig. 4b) this response would help *H. mephisto* survive the abiotic heat stress of the subterrestrial environment.

While we report significantly expanded Hsp70 and AIG1 families, only the Hsp70 genes are upregulated under heat stress in the laboratory (Fig. 4a). Interestingly the worms appear to keep per-gene Hsp70 expression low, even under conditions of heat stress. The per-gene expression of Hsp70 at high temperature is slightly lower than all genes (median Hsp70 FPKM = 2.04, and 2.58 for all genes, Fig. 4a), but the dramatic expansion of Hsp70 paralogs effectively elevates the gene dosage for Hsp70 to 51-fold higher than a single-copy gene would be. This is a similar order of induction as the ARMET protein, an UPR-related single-copy gene induced 44-fold (Fig. 4b), and which may synergize with Hsp70. While many studies have shown induction of Hsp70 genes upon heat-shock, the short-term exposure of non-heat-adapted organism to brief extreme heat (31–34), organisms under long-term adaptation to heat tend to minimize overexpression of Hsp70 because it has harmful effects on development, fertility, and growth (39,76–78). Long-term exposure to heat stress led to downregulated Hsp70 expression in both flies (79) and fish (80). The subterrestrial environment of *H. mephisto* is thermostable over time: four readings during a 5-year period showed the water temperature was an average of 36.8 ± 1.2 °C (2–4,81,82). Similarly, we cultured the worms for 2–4 weeks at constant temperatures (25, 38, or 40 °C) in the laboratory for RNA isolation and gene expression analysis. Therefore, *H. mephisto’s* sustained expression of Hsp70 under conditions of constant stable heat stress implies a functional bypass of the genes’ known detrimental effects on growth and development (39). We hypothesize that the divergent Hsp70 genes in *H. mephisto* that respond most strongly to elevated temperatures may have been functionality modified to ameliorate these deleterious effects, marking them as important candidates for future study.

We were surprised to find that the suite of expanded AIG1 genes in *H. mephisto* are not activated by heat (Fig. 4a). These genes respond to abiotic stress, including heat, in Arabidopsis (48) and in mammals the proteins are involved in immune system function, including inhibiting apoptosis during T-cell maturatin (85). Given that *H. mephisto* has dramatically expanded AIG1 copy numbers, it is tempting to speculate that these genes may be involved in responding to hypoxia or other abiotic nonthermal stresses present in the deep terrestrial subsurface, where their survival function should be adaptive. However, this hypothesis remains to be tested in future experiments. Remarkably, however, heat induced another anti-apoptotic factor in *H. mephisto*—the Bax Inhibitor 1, BI-1 (82), suggesting a different method for blocking apoptotic response under subterrestrial heat stress. These data suggest *H. mephisto* has adapted to the subterrestrial environment by managing unfolded protein stresses while upregulating Hsp70 and inhibiting apoptosis.

We note that the pacific oyster *Crassostrea gigas* has convergently expanded Hsp70 and AIG1 gene families (83) and activates the UPR in response to abiotic stress including heat (84), so *H. mephisto* helps define a general evolutionary adaptive response to heat stress (85). While the oyster experiences considerable thermal fluctuation, *H. mephisto* does not, as described above. Therefore, the signature of adaptation we report is not limited to cyclical or temporary temperature fluctuations but extends to adaptation to constant warm environments.

The expansion of Hsp70 is shared, also convergently, by distantly related *Diploscapter* species, soil nematodes which display pronounced thermotolerance (37,39), and we show here that positive selection is detectable in Hsp70 lineages in both *Diploscapter* and *H. mephisto*. These findings may not only relate to environmental heat: *D. coronatus* has been reported as a facultative parasite of humans, so it must survive to 37 °C, human body temperature (38). Consistent with this, the closest relative of *H. mephisto* is a deadly horse parasite, *H. gingivalis*, which has not been fully sequenced, and has also been reported as a facultative and fatal parasite of humans (86). Therefore, the genome signature of adaption to heat

### Table 4 GO terms enriched in *H. mephisto* genes downregulated under heat stress

| GO identifier | Name                                      | Ratio in study | Ratio in population | p_uncorrected | p_bonferroni |
|---------------|-------------------------------------------|----------------|---------------------|---------------|--------------|
| GO:0008234    | Cysteine-type peptidase activity          | 19/649         | 133/28,062          | 4.83E-07      | 0.0014       |
| GO:0042302    | Structural constituent of cuticle          | 13/649         | 89/28,062           | 9.91E-07      | 0.0029       |
| GO:0045887    | Ornithine-oxo-acid transaminase activity  | 3/649          | 3/28,062            | 1.23E-05      | 0.0362       |

* p values from GOATools Python package’s find_enrichment.py script
in *H. mephisto*, *D. coronatus*, and *D. pachys* may serve as a pre-adaptational bridge to parasitic lifestyles at least in some lineages. Therefore, these genomic adaptive strategies are of significant concern to human and animal health, and as our climate warms, it will be increasingly important to understand their evolutionary dynamics.

**Methods**

**H. mephisto culture and isolation of DNA and RNA.** The methods in this work comply with all relevant ethical regulations for research. *H. mephisto* were cultured by standard *C. elegans* methods,

for agar plates seeded with *Escherichia coli* OP50. For DNA extraction the NeuroSpin kit (Cat #740952.250, Macherey-Nagel, Bethlehem, PA, USA) was used. The pellet was resuspended in 540 μl T1 buffer supplemented with 10 μl Proteinase K. Lysis was accomplished by 4 cycles of rapid freeze-thaw using a dry ice-ethanol bath, with thawing on a 56 °C heatblock; cycles were approximately 1 min per freeze or thaw step. After this the sample was treated with an additional 25 μl Proteinase K overnight at 56 °C. After this the manufacturer’s protocol was followed for column purification of high-quality DNA, which was verified by gel electrophoresis prior to library construction.

For RNA, the worms were cultured on 5% agar plates at 25, 30, or 4°C for 2–4 weeks prior to harvest, then pelleted in PBS pH 7.7, and flash-frozen or immersed in DNA/RNA Shield Buffer for storage and extraction of nucleic acids. For RNA extraction Zymo’s Duet DNA/RNA MiniPrep Plus kit (Cat # D7003) was used. The worm pellet was resuspended in 300 μl of DNA/RNA Shield buffer, transferred to a 1.5 ml Run P RNA Shield Reagent (Cat # S6002) and homogenized on a vortexer at maximum speed twice for 5 min with a 1:2 min rest period between (for cooling). After this 30 μl of PK buffer and 15 μl Proteinase K were added, and the solution incubated for 30 min at 55 °C, then supplemented with 345 μl lysis buffer. After pelleting the insoluble material at 16,000g for 1 min, the supernatant was transferred to yellow (for DNA) and green (for RNA) columns as described in manufacturer’s protocol for the Duet DNA/RNA MiniPrep Plus Kit, performing in-column DNase treatment of RNA as recommended.

Based on Agilent BioAnalyzer 2100 output, a total of 12 samples yielded RNA of sufficient quality for sequencing: 3 from 25°C, 3 from 38°C, and 6 from 40°C. For analysis the 38°C and 40°C samples were considered together as a single sample. This data was assembled with Platanus70 as described in the next section. For Illumina, a TruSeq library with insert size of 387 bp was generated with 9 cycles of PCR after gel purification. This library was paired-end sequenced on a HiSeq2500 with 215 bp reads, yielding 58.7 million pairs. This data was assembled with Trinity as described in the next section. For PacBio, 30 lanes were run on the RS II system using libraries generated off the same DNA sample used in Illumina.

**RNA sequencing.** Stranded RNA-seq libraries were generated using the KAPA Total Stranded preparation kit (KAPA Biosystems catalog # KK8484) with an average insert size of 175 bp. Samples were sequenced on a HiSeq2500 in High Output mode using 2 × 50 bp paired-end reads, generating at least 60 M total reads per sample. Preliminary quality control for read mapping was performed using taxMaps version 0.2.11b, on a downsampled subset (1%) with the NCBI BLAST nt and a kmer size of 75 to confirm species identity for the reads generated.

**Genome assembly.** Raw read kmer analysis was performed with the SoapEC v. 2.01 (from the SOAPdenovo2 package), KmerFreq_HA set to 23 and error corrected with Corrector_HA. We found that platanus assembly was optimal at 100× average coverage, so a total of 15,582,039 random paired, error-corrected reads were used in the final assembly. Platanus version 1.2.4 was used to assemble with a stepsize 2, kmer of 21, and q value less than 0.05. The genome of each species was uploaded onto the selected database (thus, for the boxplots of gene expression, the FPKM values at high or low temperature were exported as a text file and imported into Python 2.7.14 where a custom script was used to construct the boxplots with matplotlib 2.2.2.

**Analysis of unknown genes.** *H. mephisto* genes were analyzed by blastp (evalue 1e-4) against a collection of 28 nematodes (Supplementary Table 1). When blasting with *H. mephisto* genes against *C. elegans*, *N. autumnalis*, *P. xylostella*, *M. musculus*, *D. melanogaster*, *D. viridis*, *D. pachys* and *D. pachydermus* with all programs, we found that most of the best hits were removed by using the remaining 95,378 reads, nsp variants were called with bcftools v. 1.9 mpileup and then call command with the –vm flag.
**GO analysis.** The find_enrichment.py script from the GOATOLS v.0.6.10 package\(^{162}\) was used under default settings to examine the 285 upregulated and 675 downregulated genes relative to the entire set of proteins. GO terms were assigned using Interproscan 5.30.69\(^{159}\).

**Hsp70 and AIG1 tree building.** For Hsp70 full-length proteins were aligned, and for AIG1 the proteins were broken down into domains using the envelope coordinates provided by HHMER and a custom Python script. We labeled the domains by order within the original protein. Alignments performed with MAFFT\(^{104}\) v.7.017 and refined manually to minimize indels. Trees were generated with MrBayes\(^{105}\) v.3.2.6, using the bloom state matrix and invgamma rate variation; and RAxML\(^{106}\) v.8.2.12\(^{107}\) with the PROTACBLOSUM62 rate matrix and 200 bootstrap replicates. All non-mehisto sequences in Fig. 2 are identified using their Wormbase ParaSite (https://parasite.wormbase.org/index.html) gene identifier. Accessions are given in Supplementary Table 1. For Fig. 3 we prepended non-mehisto sequences with the shortest possible (normally two letter) genus and species abbreviation (e.g., *B. xylophilus*).

**PAML with branch-sites model.** Using CodeML (PAML v.4.9)\(^{108}\) branch site \(\omega\) was estimated using runmode = 0, settype = 1, model = 2, Nsites = 2, in the Codeml control file. All other parameters were used on the default setting. Each branch was estimated twice: once with a neutral model (fix omega = 1 and omega = 1) and once using a purifying selection model (fix omega = 0, omega = 1). The \(p\) values were determined using the LRT statistic 2\(\chi_l\) compared against \(\chi_2\) with critical values of 3.84, 5% significance level, and 6.63, 1% significance\(^{109}\). In our analysis there were (after gap removal) 891 sites across 55 different aligned nucleotide (coding) sequences.

**Analysis of subtelomeric sequences.** Extraction of telomere repeat-containing read pairs (at least one of which contains at least four copies of TTAGGG) from raw Illumina data was performed using a custom Python script. These reads were merged at their overlaps using PEAR\(^{110}\) v.0.9.8, and the resultant fused reads were assembled with MIRA\(^{111}\) v4.0. The resulting contigs were manually inspected and redundant sequences were collapsed to obtain an estimate of subtelomeric region number.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Data availability**

This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession SWTD00000000. The version described in this paper is version SWTD01000000. The raw Illumina DNA and RNA data, and PacBio DNA data are available on the Sequence Reads Archive (SRA) at accession PRJNA528747. Transcriptome expression data as FPKM values genome-wide is available in the Gene Expression Omnibus (GEO) at accession GSE133178. The genome annotation file and transcript and protein sequences are available from WormBase Parasite at accession Halicephalobus_mehisto:prjna528747. The source data underlying Fig. 4A, B are provided as a Source Data File.

Received: 6 November 2018; Accepted: 24 October 2019; Published online: 21 November 2019

**References**

1. Borgonie, G. et al. Nematomata from the terrestrial deep subsurface of South Africa. *Nature* **474**, 79–82 (2011).
2. Lau, M. C. Y. et al. An oligotrophic deep-subsurface community dependent on syntrophy is dominated by sulfur-driven autotrophic denitrifiers. *Proc Natl Acad. Sci. USA* **113**, E7927–E7936 (2016).
3. Magnabosco, C. et al. Fluctuations in populations of subsurface methane oxidizers in coordination with changes in electron acceptor availability. *FEMS Microbiol. Ecol.* **94**, https://doi.org/10.1093/femsec/fiy089 (2018).
4. Lau, M. C. Y. et al. Phylogeny and phylogeography of functional genes shared among seven terrestrial subsurface metagenomes reveal N-cycling and microbial evolutionary relationships. *Front. Microbiol.* **5**, ARTN 53110.3389/lmich.2014.00531 (2014).
5. Borgonie, G. et al. Eukaryotic opportunists dominate the deep-subsurface biosphere in South Africa. *Nat. Commun.* **6**, 8952 (2015).
6. Rödelsperger, C., Streit, A. & Sommer, R. J. In *LS* (John Wiley & Sons, Ltd., 2013).
7. Wasmuth, J., Schmid, R., Hedley, A. & Blaxter, M. On the extent and origins of genic novelty in the phylum Nematoda. *PLoS Negl. Trop. Dis.* **2**, e258 (2008).
8. Coghlan, A. Nematode genome evolution. *WormBook*, 1–15, https://doi.org/10.1895/wormbook.1.151.05 (2005).
9. Borgonie, G. et al. Refuge from predation, the benefit of living in an extreme acidic environment? *Bioul. Biol.* **219**, 268–276 (2010).
10. Ribeiro, V. L. et al. The genomic basis of parasitism in the Strongyloides clade of nematodes. *Nat. Genet.* **48**, 299–307 (2016).
11. Yin, D. et al. Rapid genome shrinkage in a self-fertile nematode reveals sperm competition proteins. *Science* **359**, 55+–, (2018).
12. Leipnitz, O., Wolf, Y. I., Koonin, E. V. & Aravind, L. The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res.* **12**, 1048–1059 (2002).
13. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).
14. Parra, G., Bradnam, K., Ning, Z., Keane, T. & Korf, I. Assessing the gene space of living eukaryotes. *Nat. Genetics* **39**, 289–297 (2007).
15. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–965 (1997).
16. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
17. Lowe, T. M. & Eddy, S. R. TRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–965 (1997).
18. Wheeler, T. J. & Eddy, S. R. nhmmer: DNA homology search with profile HMMs. *Bioinformatics* **29**, 2487–2489 (2013).
19. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinforma.* **12**, 491 (2011).
21. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. 7, 562–578 (2012).
22. P Pertea, M., Kim, D., Chang, M. C., Lee, I. T. & Sal c, S. I. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballg o. Nat. Protoc. 11, 1650–1667 (2016).
23. Haas, B. J. et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 9, 1512–1530 (2014).
24. Chervitz, S. A. et al. Comparison of the complete protein sets of worms and yeast: orthology and divergence. Science 282, 2022–2028 (1998).
25. Thomas, J. H. Analysis of homologous gene clusters in Caenorhabditis elegans reveals striking regional cluster domains. Genetics 172, 124–143 (2006).
26. Wang, Y., Coleman-Derr, D., Chen, G. & Gu, Y. Q. OrthoVenin: a web server for genome-wide comparison and annotation of orthologous clusters across multiple species. Nucleic Acids Res. 43, W78–W84 (2015).
27. Assimon, V. A., Gillies, A. T., Rauch, J. N. & Gestwicki, J. E. Hsp70 proteins as complex drug targets. Curr. Pharm. Des. 19, 404–417 (2013).
28. Sherman, M. Y. & G_absi, V. L. Hsp70 in cancer: back to the future. Oncogene 34, 4153–4161 (2015).
29. Sorensen, J. G., Kristensen, T. N. & Loschekc, V. The evolutionary and ecological role of heat shock proteins. Ecol. Lett. 6, 1025–1037 (2003).
30. Usman, M. G. et al. Molecular analysis of Hsp70 mechanisms in plants and their function in response to stress. Biotechnol. Genet. Eng. 33, 26–39 (2017).
31. Brunquell, J., Morris, S., Lu, Y., Cheng, F. & Westerheide, S. D. The genome-wide (HGS) influence of the regulation of gene expression in Caenorhabditis elegans. BMC Genomics 17, ARTN 55910.1186/s12864-016-2837-5 (2016).
32. Lakhotia, S. C., Srivastava, P. & Prasanth, K. V. Regulation of heat shock proteins, Hsp70 and Hsp64, in heat shocked Malpighian tubules of Drosophila melanogaster larvae. Cell Stress Chaperones 7, 347–356 (2002).
33. Prahlad, V., Cornelius, T. & Mormito, R. I. Regulation of the cellular heat shock response in Caenorhabditis elegans by thermosensory neurons. Science 320, 811–814 (2008).
34. Rodriguez, M., Snek, L. B., De Bon, M. & Kammenga, J. E. Worms under stress: C. elegans stress response and its relevance to complex human disease and aging. Trends Genet. 29, 367–374 (2013).
35. Yu, A. et al. Roles of Hsp70s in stress responses of microorganisms, plants, and animals. Biomed. Res. Int. 2015, 510319 (2015).
36. Bocchieri, L., de Macario, E. C. & Macario, A. J. L. hsp70 genes in complex drugs as targets. Drug. Des. Devel. Ther. 7, 4153–4161 (2013).
37. Yang, Z. Inference of selection from multiple species alignments. NATURE COMMUNICATIONS|           (2019) 10:5268 | https://doi.org/10.1038/s41467-019-13245-8 | www.nature.com/naturecommunications
38. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. Nat. Protoc. 4, 510–513 (2009).
39. Yang, Z., Wong, W. S. & Nielsen, R. Bayes empirical bayes inference of amino acid changes in multiple species. Nucleic Acids Res. 34, W78–W84 (2015).
40. Yang, Z. Inference of selection from multiple species alignments. NATURE COMMUNICATIONS|           (2019) 10:5268 | https://doi.org/10.1038/s41467-019-13245-8 | www.nature.com/naturecommunications
41. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. Nat. Protoc. 4, 510–513 (2009).
42. Yang, Z., Wong, W. S. & Nielsen, R. Bayes empirical bayes inference of amino acid changes in multiple species. Nucleic Acids Res. 34, W78–W84 (2015).
43. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. Nat. Protoc. 4, 510–513 (2009).
44. Reuber, T. L. & Ausubel, F. M. Isolation of Arabidopsis genes that differentiate fi from soil, as a vector of food-borne pathogenic bacteria to preharvest fruits and animals. Arch. Virol. 152, 639–643 (2012).
45. Lin, K. et al. Single nucleotide genome sequencing reveals high similarity among nuclei of an endomycorrhizal fungal pion. Plos Genet. 6, ARTN e100047810.1371/ journal.pgen.1004078 (2014).
46. Nitta, T. & Takahama, Y. The lymphocyte guard-lINCs: regulation of lymphocyte survival by IAN/GIMAP family proteins. Trends Immunol. 28, 58–65 (2007).
47. Steiner, G. Opuscula miscellanea nematologica. IX. Parasitology 28, 58–65 (2007).
48. Opperman, C. H. et al. Sequence and genetic map of Meloidogyne hapla: a compact nematode genome for plant parasitism. Proc. Natl Acad. Sci. USA 105, 14802–14807 (2008).
49. Abad, P. et al. Genome sequence of the metazoan plant-parasitic nematode Meloidogyne incognita. Nat. Biotechnol. 26, 909–915 (2008).
50. Fradin, H. et al. Genome architecture and evolution of a unicellular nematode. Curr. Biol. 27, 2928–2939 e2926 (2017).
51. Kajitani, R. et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res. 24, 1384–1395 (2014).
52. Treangen, T. J. & Salzberg, S. L. Repeatative DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet. 13, 243–252 (2012).
53. Wicky, C. et al. Tethered repeats (TTAGGC)n are sufficient for chromosome capping function in Caenorhabditis elegans. Proc. Natl Acad. Sci. USA 93, 8983–8988 (1996).
54. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-end read merger. Bioinformatics 30, 614–620 (2014).
55. Hollen, J. & Weissman, J. S. Decay of endoplasmic reticulum-localized mRNAs during the unfolded protein response. Science 313, 104–107 (2006).
56. Maurel, M., Chevet, E., Tavernier, J. & Gerlo, S. Getting RIDD of RNA: IRE1 in cell fate regulation. Trends Biochem. Sci. 39, 245–254 (2014).
57. Krebs, R. A. & Feder, M. E. Deleterious consequences of Hsp70 overexpression in Drosophila melanogaster larvae. Cell Stress Chaperones 2, 60–71 (1997).
58. Krebs, R. A. & Loschek, V. Effects of exposure to short-term heat-stress on fitness components in Drosophila melanogaster. J. Evolut. Biol. 7, 39–49 (1994).
59. Williams, K. D., Hedin, A. B., Poslusny, J., Roberts, S. P. & Feder, M. E. Effect of heat shock, pretreatment and hsp70 copy number on wing development in Drosophila melanogaster. Mol. Biol. Evol. 12, 1165–1177 (2005).
60. Sorensen, J. G., Dahlgaard, J. & Loschekc, V. Genetic variation in thermal tolerance among natural populations of Drosophila buzzatii: down regulation of Hsp70 expression and variation in heat stress resistance traits. Fungl. Ecol. 15, 289–296 (2008).
61. Narum, S. R., Campbell, N. R., Meyer, K. A., Miller, M. R. & Hardy, R. W. Thermal adaptation and acclimation of ectotherms from differing austral climates. Mol. Ecol. 22, 3090–3097 (2013).
81. Magnabosco, C. et al. Comparisons of the composition and biogeographic distribution of the bacterial communities occupying South African thermal springs with those inhabiting deep subsurface fracture water. Front. Microbiol. 5, ARTN 67910.3389/fmicb.2014.00679 (2014).
82. Simkus, D. N. et al. Variations in microbial carbon sources and cycling in the deep continental subsurface. Geochim. Cosmochim Acta 173, 264–283 (2016).
83. Powell, D. et al. The genome of the oyster Saccostrea offers insight into the Eddie’s puzzle. Nature 529, 655–664 (2018).
84. Zhang, G. et al. The oyster genome reveals stress adaptation and complexity of shell formation. Nature 490, 49–54 (2012).
85. Guerin, M. N., Weinstein, D. J. & Bracht, J. R. Stress adapted mollusca and nematoda exhibit convergently expanded Hp70 and AIG1 gene families. J. Mol. Evol. https://doi.org/10.1007/s00239-019-09900-9 (2019).
86. Lim, C. K. et al. First human case of fatal Halicephalobus gingivalis meningonecrophilis in Australia. J. Clin. Microbiol. 53, 1768–1774 (2015).
87. Brenner, S. The genetics of Caenorhabditis elegans. Genetics 77, 71–94 (1974).
88. Corvelo, A., Clarke, W. E., Robine, N. & Zody, M. C. taxMaps: comprehensive and highly accurate taxonomic classification of short-read data in reasonable time. Genome Res. 28, 751–758 (2018).
89. Li, R. et al. De novo assembly of human genomes with massively parallel short reads. Genome Res. 20, 265–272 (2010).
90. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4. http://www.repeatmasker.org.
91. Eddy, S. R. Accelerated profile HMM searches. PLoS Comput. Biol. 7, e1002195 (2011).
92. Wheeler, T. J. et al. Dfam: a database of repetitive DNA based on profile hidden Markov models. Nucleic Acids Res. 41, D70–D82 (2013).
93. Fraze, A. C. et al. Ballgown bridges the gap between transcriptome assembly and expression analysis. Nat. Biotechnol. 33, 243–246 (2015).
94. Eddy, S. R. Profile models for hidden Markov models. Bioinformatics 14, 755–763 (1998).
95. Fischer, S. et al. Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. Curr. Protoc. Bioinforma. 12(Chapter 6), 11–19 (2011). Unit 6.
96. Li, L., Stoeckert, C. J. Jr. & Roos, D. S. OrthoMCL: identification of orthology groups for eukaryotic genomes. Genome Res. 13, 2178–2189 (2003).
97. Camacho, C. et al. BLAST+: architecture and applications. BMC Bioinforma. 10, 421 (2009).
98. Ekstedt, O. K., Kuiper, M. & Mironov, V. orthAgogue: an agile tool for the rapid multiple sequence alignment based on fast Fourier transform. Bioinformatics 28, 3066–3074 (2012).
99. Edgar, R. C. MUSCLE: a program package for multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 33, 74–78 (2005).
100. Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1915–1917 (2009).
101. Darrriba, D., Taboada, G. L., Doallo, R. & Posada, D. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27, 1164–1165 (2011).
102. Klopfenstein, D. V. et al. GOATOOLS: a Python library for Gene Ontology analysis. Sci. Rep. 8, 10872 (2018).
103. Jones, P. et al. InterProScan 5: genome-scale protein function classification. Bioinformatics 30, 1236–1240 (2014).
104. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30, 3059–3066 (2002).
105. Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17, 754–755 (2001).
106. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313 (2014).
107. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13, 555–556 (1997).
108. Yang, Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. 19, 568–573 (1998).
109. Chevreux, B. et al. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res. 14, 1147–1159 (2004).
110. Foth, B. J. et al. Whipworm genome and dual-species transcriptome analyses provide molecular insights into an intimate host-parasite interaction. Nat. Genet. 46, 693–700 (2014).

Acknowledgements The authors acknowledge Laura Landweber, in whose laboratory J.R.B. prepared the H. mephisto Illumina DNA library. We thank Wei Wang, Jessica Wiggins, and Donna Storten of the Princeton Sequencing Core Facility for assistance with Illumina library preparation and sequencing. We also acknowledge Aaron Goldman for early work done on the H. mephisto project, and Elizabeth Ginsburg for comments on the paper. We thank Evgeny Bisk for assistance with the installation and troubleshooting of software on the Zorro Computer Cluster at American University. Mariana Erasmus acknowledges support from the Technology Innovation Agency (TIA), an agency of the Department of Science and Technology (DST) and is grateful to Izelle de Beer and Bernice Jordaan for their assistance in culturing of the nematodes. We are indebted to the logistical support of Sibanye Gold Limited, the management and staff of Beatrix Gold Mine that enabled the capture of H. mephisto and follow-up samples for metagenomic studies. The work of M.C.Y.L. was supported by NSF Grant DDEB-1441646 to T.C.O., and J.R.B. was supported by NIH 1K22CA184297 and a NASA DC Space Grant Consortium grant which funded the RNA and PacBio sequencing.

Author contributions D.J.W., S.E.A., and J.R.B. performed gene prediction and phylogenetic tree construction; D.J.W. and J.R.B. performed transcriptome analysis and wrote the paper, and J.R.B. performed genome assembly; M.C.Y.L. constructed the multilocus phylogenetic tree; M.E. and E.V.H. grew the nematodes for RNA; K.C.A. performed analysis of positive selection and K.W.C. performed repetitive element analysis; G.D. and R.S. performed Illumina RNA and PacBio DNA sequencing. G.B. cultured the nematodes for DNA extractions and with T.C.O. provided critical feedback on the manuscript drafts and input into experimental design.

Competing interests The authors declare no competing interests.

Additional information Supplementary information is available for this paper at https://doi.org/10.1038/s41467-019-13245-8.

Correspondence and requests for materials should be addressed to J.R.B.

Peer review information Nature Communications thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019