LETTER

OPEN

doi:10.1038/nature12326

Genomic evidence for ameiotic evolution in the bdelloid rotifer *Adineta vaga*

Jean-François Flot[1,2,3,4,5,6], Boris Hespeels[1,2], Xiang Li[1,2], Benjamin Noel[3], Irina Arkhipova[7], Etienne G. J. Danchin[8,9,10], Andreas Hejnol[11], Bernard Henrissat[12], Romain Koszul[13], Jean-Marc Aury[3], Valérie Barbe[3], Roxane-Marie Barthélémy[14], Jens Bast[15], Georgii A. Bazykin[16,17], Olivier Chabrol[14], Arnaud Couloux[3], Martine Da Rocha[8,9,10], Corinne Da Silva[3], Eugene Gladyshev[7], Philippe Gouret[14], Oskar Hallatschek[6,18], Bette Hecox-Lea[7,19], Karine Labadie[3], Benjamin Lejeune[1,2], Oliver Piskurek[20], Julie Poulain[3], Fernando Rodriguez[7], Joseph F. Ryan[11], Olga A. Vakhrusheva[16,17], Eric Wajnberg[8,9,10], Bénédicte Wirth[14], Irina Yushenova[7], Manolis Kellis[21], Alexey S. Kondrashov[16,22], David B. Mark Welch[7], Pierre Pontarotti[14], Jean Weissenbach[3,4,5], Patrick Wincker[3,4,5], Olivier Jaillon[3,4,5,21]* & Karine Van Doninck[1,2]*

Loss of sexual reproduction is considered an evolutionary dead end for metazoans, but bdelloid rotifers challenge this view as they appear to have persisted asexually for millions of years[1]. Neither male sex organs nor meiosis have ever been observed in these microscopic animals: oocytes are formed through mitotic divisions, with no reduction of chromosome number and no indication of chromosome pairing[2]. However, current evidence does not exclude that they may engage in sex on rare, cryptic occasions. Here we report the genome of a bdelloid rotifer, *Adineta vaga* (Davis, 1873)[3], and show that its structure is incompatible with conventional meiosis. At gene scale, the genome of *A. vaga* is tetraploid and comprises both anciently duplicated segments and less divergent allelic regions. However, in contrast to sexual species, the allelic regions are rearranged and sometimes even found on the same chromosome. Such structure does not allow meiotic pairing; instead, we find abundant evidence of gene conversion, which may limit the accumulation of deleterious mutations in the absence of meiosis. Gene families involved in resistance to oxidation, carbohydrate metabolism and defence against transposons are significantly expanded, which may explain why transposable elements cover only 3% of the assembled sequence. Furthermore, 8% of the genes are likely to be of non-metazoan origin and were probably acquired horizontally. This apparent convergence between bdelloids and prokaryotes sheds new light on the evolutionary significance of sex.

With more than 460 described species[4], bdelloid rotifers (Fig. 1) represent the highest metazoan taxonomic rank in which males, hermaphrodites and meiosis are unknown. Such persistence and diversification of an ameiotic clade of animals are in contradiction with the supposed long-term disadvantages of asexuality, making bdelloids an 'evolutionary scandal'[5]. Another unusual feature of bdelloid rotifers is their extreme resistance to desiccation at any stage of their life cycle[6], enabling these microscopic animals to dwell in ephemeral freshwater habitats such as mosses, lichens and forest litter; this ability is presumably the source of their extreme resistance to ionizing radiation[7].

We assembled the genome of a clonal *A. vaga* lineage into separate haplotypes with a N_{50} of 260 kilobases (kb) (that is, half of the assembly was composed of fragments longer than 260 kb). Assembly size was 218 megabases (Mb) but 26 Mb of the sequence had twice the average sequencing coverage, suggesting that some nearly identical regions were not resolved during assembly (Supplementary Fig. 3); hence, the total genome size is likely to be 244 Mb, which corresponds to the estimate obtained independently using fluorometry (Supplementary Note C2). Annotation of the complete assembly (including all haplotypes) yielded 49,300 genes. Intragenomic sequence comparisons revealed numerous homologous blocks with conserved gene order (colinear regions). For each such block we computed the per-site synonymous divergence (Ks) and a colinearity metric defined as the fraction of colinear genes. Colinear blocks fell into two groups (Fig. 2a): a group characterized by high colinearity and low average synonymous divergence, and a group characterized by lower colinearity and higher synonymous divergence. The presence of two classes of colinear blocks is consistent with a tetraploid structure comprised of alleles (recent homologues) and ohnologues (ancient homologues formed by genome duplication). Allelic pairs of coding sequences are on average 96.2%

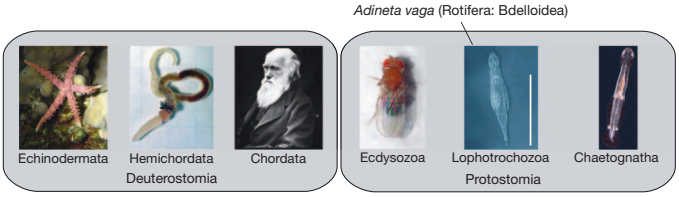


Figure 1 | Position of bdelloid rotifers among metazoans. Bdelloid rotifers ('leech-like wheel-bearers') are a clade of microscopic animals (scale bar, 100 μm) within the phylum Rotifera. Photographs of Hemichordata (*Saccoglossus*), Chordata (*Homo*) and Ecdysozoa (*Drosophila*) courtesy of David Remsen (MBL), John van Wyhe (http://darwin-online.org.uk) and André Karwath, respectively.

[1]University of Namur, Department of Biology, URBE, Laboratory of Evolutionary Genetics and Ecology, 5000 Namur, Belgium. [2]Namur Research Institute for Life Sciences (NARILIS), 5000 Namur, Belgium. [3]CEA-Institut de Génomique, GENOSCOPE, Centre National de Séquençage, 2 rue Gaston Crémieux, CP5706, 91057 Evry Cedex, France. [4]Université d'Evry, UMR 8030, CP5706, 91057 Evry Cedex, France. [5]Centre National de la Recherche Scientifique (CNRS), UMR 8030, CP5706, 91057 Evry Cedex, France. [6]Max Planck Institute for Dynamics and Self-Organization, Biological Physics and Evolutionary Dynamics, Bunsenstraße 10, 37073 Göttingen, Germany. [7]Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, Massachusetts 02543, USA. [8]INRA, UMR 1355 ISA, Institut Sophia Agrobiotech, 400 route des Chappes, 06903 Sophia-Antipolis, France. [9]CNRS, UMR 7254 ISA, Institut Sophia Agrobiotech, 400 route des Chappes, 06903 Sophia-Antipolis, France. [10]Université de Nice Sophia-Antipolis, UMR ISA, Institut Sophia Agrobiotech, 400 route des Chappes, 06903, Sophia-Antipolis, France. [11]Sars International Centre for Marine Molecular Biology, University of Bergen, 5008 Bergen, Norway. [12]Architecture et Fonction des Macromolécules Biologiques, Aix-Marseille University, CNRS UMR 7257, 13288 Marseille, France. [13]Groupe Spatial regulation of genomes, CNRS UMR 3525, Department of Genomes and Genetics, Institut Pasteur, 75724 Paris, France. [14]LATP UMR-CNRS 7353, Evolution Biologique et Modélisation, Aix-Marseille University, 13331 Marseille cedex 3, France. [15]J.F. Blumenbach Institute of Zoology and Anthropology, University of Göttingen, 37073 Göttingen, Germany. [16]Department of Bioengineering and Bioinformatics, M.V. Lomonosov Moscow State University, Leninskye Gory 1-73, Moscow, 119991, Russia. [17]Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Bolshoi Karetny pereulok 19, Moscow, 127994, Russia. [18]Department of Physics, University of California, Berkeley, California 94720, USA. [19]Department of Biology, Northeastern University, Boston, Massachusetts 02115, USA. [20]Courant Research Centre Geobiology, Georg-August-Universität Göttingen, Goldschmidtstraße 3, Göttingen 37077, Germany. [21]MIT Computer Science and Artificial Intelligence Laboratory, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02139, USA. [22]Life Sciences Institute and Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan 48109-2216, USA.
*These authors contributed equally to this work.

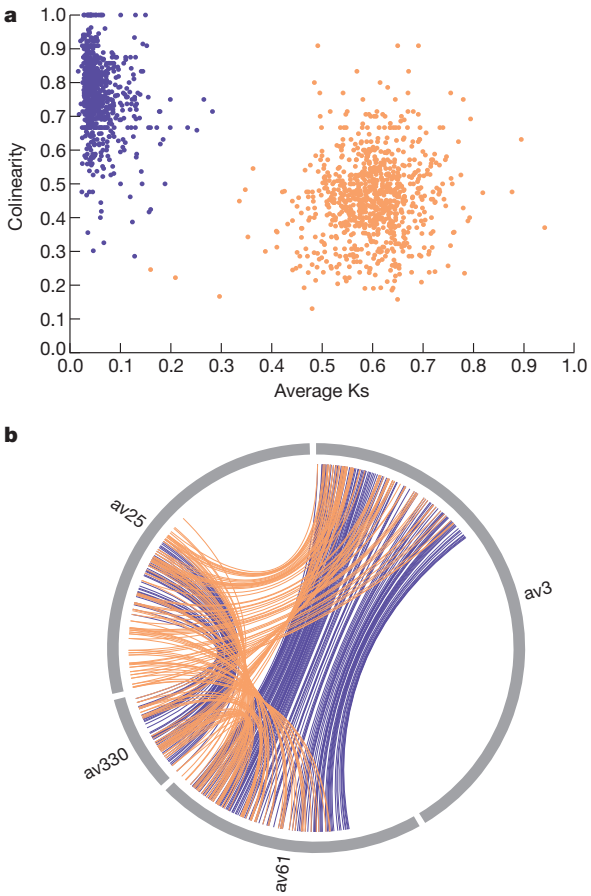


Figure 2 | A locally tetraploid genome. **a**, Analysis of intragenomic synteny reveals two groups of colinear regions: alleles (in violet, regions characterized by a high fraction of colinear genes and low average Ks, that is, synonymous divergence) and ohnologues (in orange, with lower colinearity but higher Ks). **b**, Example of a genomic quartet of four scaffolds: allelic gene pairs are connected with violet curves and ohnologous gene pairs with orange curves.

identical at the nucleotide level (median = 98.6%) versus 73.6% (median = 75.1%) for ohnologous pairs. Nearly 40% (84.5 Mb) of the assembled genome sequence is organized in quartets of four homologous regions A_1, A_2, B_1 and B_2, of which A_1–A_2 and B_1–B_2 are two pairs of alleles and As are ohnologous to Bs[8] (Fig. 2b).

We found evidence of genomic palindromes up to 705 kb in length and involving up to 148 genes. The *A. vaga* genome contains at least 17 such palindromic regions (Fig. 3a) reminiscent of those reported in the Y chromosomes of primates[9]. In all 17 cases, the arms of the palindromes present the colinearity and divergence signatures of allelic regions and do not have other allelic duplicates in the assembly, suggesting that they arose by inter-allelic rearrangements rather than by local duplications. In addition to these 17 inverted repeats, we observed three direct repeats that present the signatures of allelic blocks and involve up to 50 genes (Fig. 3a). The cumulative length of the assembly fragments (scaffolds) bearing these 20 allelic rearrangements is 7.5 Mb or 3.5% of the genome sequence. Allelic regions that are found on the same chromosome clearly cannot segregate during meiosis. Moreover, we found hundreds of colinearity breakpoints between allelic regions, and the total length of the scaffolds that have no full-length homologue in the assembly due to these breakpoints exceeds 109 Mb or 51% of the genome assembly (including 91 of the 100 largest scaffolds, Fig. 3b and Supplementary Fig. 10). As a result, it is impossible to split the assembled genome of *A. vaga* into haploid sets: the apparent ploidy level of *A. vaga* is scale-dependent, with a tetraploid structure at gene scale versus chromosome-scale haploidy. Such relaxation of constraints on genome structure is reminiscent of other mitotic lineages such as cancer cells[10] and somatic tissues[11].

It has been proposed that, in the absence of meiosis, alleles accumulate mutations independently from one another, to the point that ancient asexuals may harbour genome-wide allele sequence divergence (ASD)[12] larger than inter-individual differences (the so-called 'Meselson effect'). However, the average inter-allelic divergence of *A. vaga* is only 4.4% at the nucleotide level (3% when looking at synonymous divergence), which falls in the upper range reported for sexually reproducing species[13]. The absence of genome-wide ASD could be explained by low mutation rates and/or by frequent mitotic recombination (such as gene conversion resulting from DNA repair)[12]. Although there is no evidence of reduced mutation rates in bdelloid rotifers compared with their cyclically sexual sister clade the monogononts[14], we found strong signatures

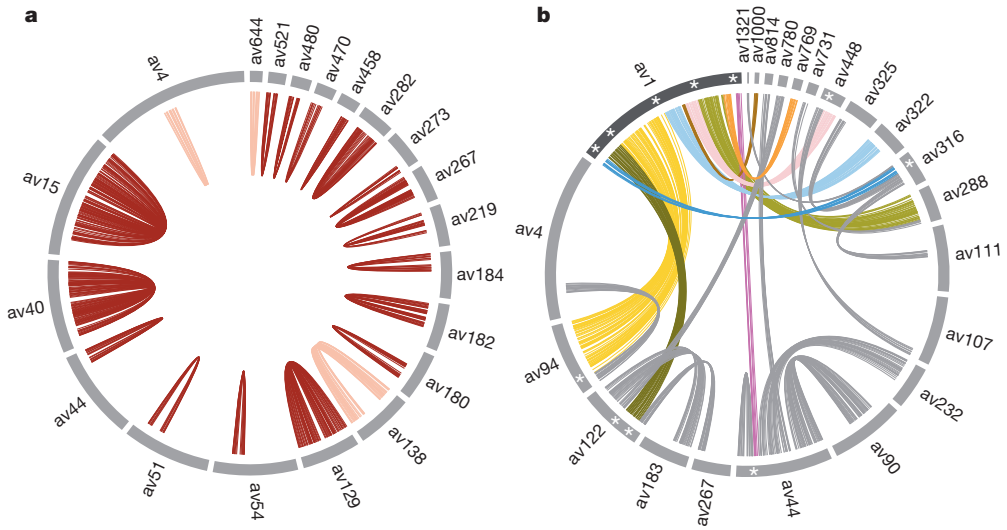


Figure 3 | A genome structure incompatible with conventional meiosis. **a**, In twenty cases, allelic regions are found to occur on the same chromosome. All curves shown connect allelic gene pairs. On three scaffolds both allelic regions have the same orientation (direct repeats, in pink), whereas on the seventeen other scaffolds they are inverted (palindromes, in red). **b**, Local colinearity between alleles does not extend to chromosome scale. Colours are arbitrary and only allelic gene pairs are represented. Asterisks highlight colinearity breakpoints between scaffold av1 and its allelic partners av44, av94, av122, av316 and av448. Further examples for other scaffolds are shown on Supplementary Fig. 10.

of recent gene conversion events in the distribution of identity track lengths, that is, distances between consecutive mismatches (Fig. 4a and Supplementary Note E1). We calculated that the probability that a given base in the genome experiences gene conversion is at least one order of magnitude greater than its probability to mutate (Supplementary Note E1), suggesting that homologous regions in the genome of *A. vaga* undergo concerted evolution[15]. Homogenization through gene conversion may either expose new mutations to selection by making them homozygous or remove them as they get overwritten with the other allelic version (Fig. 4b), thereby slowing Muller's ratchet (that is, the irreversible accumulation of detrimental mutations in asexual populations of finite sizes, Supplementary Note E2 and Supplementary Fig. 11).

Over 8% of the genes of *A. vaga* are much more similar to non-metazoan sequences in GenBank than to metazoan ones (AI log score > 45 (ref. 16), Supplementary Note E4) and were therefore probably acquired through horizontal gene transfer (HGT). This class of genes has significantly fewer introns per kilobase of coding sequence compared with probable core metazoan genes (AI ≤ −45, Supplementary Table 2). More than 20% of genes with AI > 45 are found in quartets (groups of four homologous copies in conserved syntenic regions) and were therefore probably incorporated into the rotifer genome before the establishment of tetraploidy, which itself pre-dates the divergence

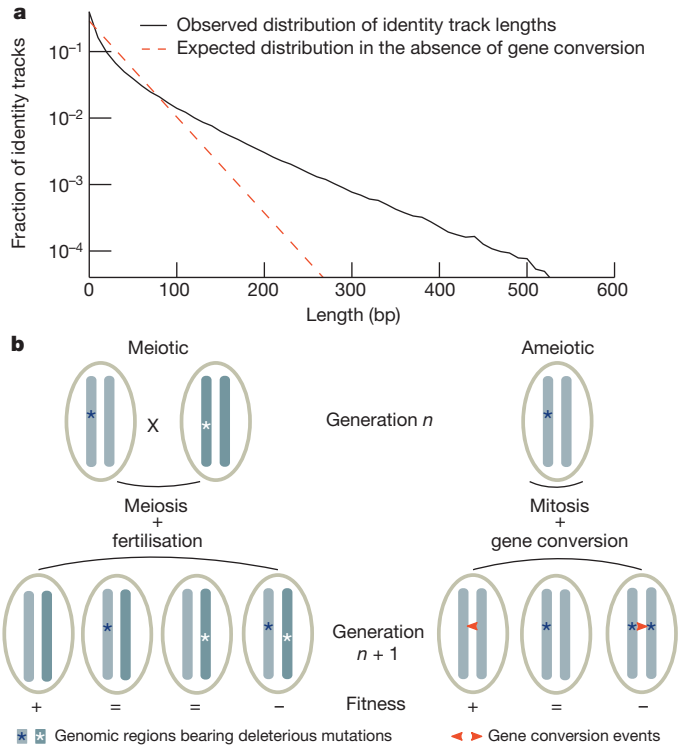


Figure 4 | Gene conversion and its evolutionary consequences in ameiotic organisms. a, Evidence for gene conversion between allelic regions. If we suppose that mutations happen at random in a Poisson process of parameter $1/M$ (where M is the average distance between mutations), then the distance between two consecutive mismatches follows a negative exponential distribution where the proportion of identity tracks of length x equals $e^{-x/M}/M$. Comparison of the observed distribution of identity track lengths with this theoretical distribution reveals a deficit of short tracks and an excess of long tracks, as expected in case of gene conversion. The same pattern was observed when gene-coding regions were excluded from the analysis (data not shown), thereby ruling out a confounding effect of selection. **b**, In sexual organisms, meiotic recombination can generate offspring with fewer or more deleterious mutations (hence increasing or decreasing fitness) than the previous generation. The same outcome is expected in ameiotic organisms that experience gene conversion: a deleterious allele may be overwritten by a beneficial or neutral one, resulting in an increase in fitness, or may overwrite it, resulting in decreased fitness.

of extant bdelloid families[8]. The higher the number of copies of a putative HGT gene, the higher its number of introns and the closer its guanine–cytosine (GC) content to the *A. vaga* genome average (Supplementary Fig. 22), which suggests that these parameters reflect the age of acquisition. We also noticed signatures of possibly very recent HGTs: 60 genes with AI > 45 are present in only one copy (with normal coverage), have no intron and have a GC content that is more than 1% above or below the genome average (the same scaffolds also bear genes of probable metazoan origin with AI < 0). In summary, there seems to be an ancient but still ongoing process of HGT at a level comparable to some bacteria[17].

Some theories predict that transposable elements should be either absent from the genomes of asexuals[18] or undergo unrestrained expansion after the switch to asexuality, potentially leading to species extinction unless transposable element proliferation is prevented[19]. We found that transposable elements cover about 3% of the *A. vaga* genome, which is less than the percentage reported in most other metazoans (including the genome of the obligate parthenogenetic nematode *Meloidogyne incognita*, 36% of which is made up of repetitive elements[20]). Another surprising feature is the high diversity of transposable-element families and the extremely low copy numbers observed for each of them (Supplementary Table 3). Out of 255 families, the overwhelming majority (209) are represented by only one or two full-length copies (for 24 families, no full-length copies could be identified), and for each full-length copy there are, on average, only about ten times as many transposable-element fragments. This relatively low abundance of decayed copies and the fact that long-terminal-repeat (LTR) retrotransposons have identical or nearly identical LTRs (Supplementary Table 4) suggest that most low-copy-number families represent recent arrivals. This is consistent with an ongoing process of acquisition of transposable elements by HGT.

This hypothesis is further supported by the significantly higher density of transposable elements observed around HGTs and vice-versa (Supplementary Note E5). If *A. vaga* has been acquiring transposable elements by HGT, a question that arises is what keeps their number lower than in most other metazoans. Many fragmented copies have apparently been formed through microhomology-mediated deletions. Excision of LTR retrotransposons has also been occurring through LTR–LTR recombination, leaving behind numerous solo LTRs: for example, two *Juno1* insertions, *Juno1.1* and *Juno1.2*, which were present as full-length copies in the 2006 *A. vaga* fosmid library[21], exist in the current assembly only as solo LTRs (in the same genomic environments and with the same target site duplications). Finally, there is evidence for expansion and diversification of the RNA-mediated silencing machinery. In addition to Dicer1 proteins, which are shared by all metazoans, *A. vaga* possesses a deep-branching Dicer-like clade with uncertain taxonomic placement (Supplementary Fig. 20). The Argonaute/Piwi and RNA-directed RNA polymerase (RdRP) families are also expanded (Supplementary Figs 18 and 19). It is plausible that these proteins participate in epigenetic silencing of transposable elements (as was recently observed for single-copy transgenes in *Caenorhabditis elegans*[22]), thereby preventing horizontally transferred transposable elements from multiplying upon arrival.

Overall, the genome of *A. vaga* comprises more genes than usually reported for metazoans (Supplementary Note F2), as its haplotypes were assembled separately. Even taking this into account, the gene repertoire of *A. vaga* features expansion of several gene families. For example, the genome of *A. vaga* comprises 284 homeobox superclass genes, mostly found in four copies (quartets) but not organized in clusters; very few ohnologues have been lost, resulting in more homeobox genes than in any other metazoan genome sequenced (Supplementary Note F5). Genes putatively related to oxido-reduction processes are substantially more abundant in *A. vaga* than in other metazoan species, and most of the corresponding genes appear to be constitutively expressed (Supplementary Table 9). This is consistent with the recent report of an effective antioxidant protection system

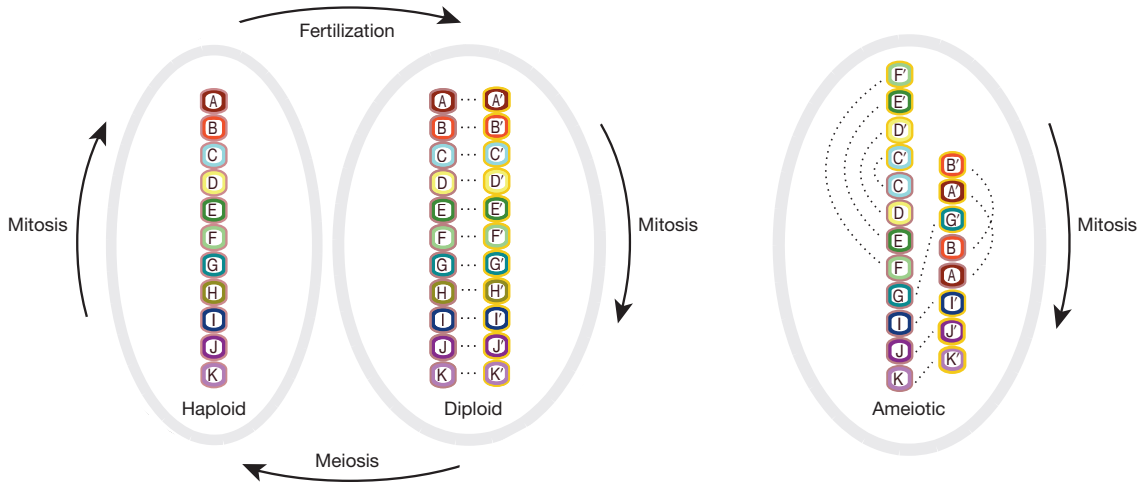


Figure 5 | Meiotic versus ameiotic genome structures. Genes are represented with letters, and dashed lines connect allelic gene pairs. A meiotic genome (left) alternates between a haploid phase (in which a single allele of each gene is present) and a diploid phase (in which the genes are present in two allelic versions arranged colinearly on homologous chromosomes). In the ameiotic genome of *A. vaga* (right), alleles are distributed in blocks that are shuffled across chromosomes, resulting notably in intrachromosomal repeats (direct or inverted). As a consequence, chromosomes have no homologues and cannot be paired.

in bdelloid rotifers[23]. Carbohydrate-active enzymes (CAZymes) in the genome of *A. vaga* are also notably diverse and abundant, with 1,075 genes falling into 202 characterized families. With 623 glycoside hydrolases (involved in the hydrolysis of sugar bonds) and 412 glycosyltransferases (responsible for building sugar bonds), the CAZyme richness of *A. vaga* ranks highest among metazoans and is only comparable to some plants such as poplars[24]. *A. vaga* has the richest repertoire of glycoside hydrolases of any organism sequenced so far, hinting at a diversity of feeding habits; 52% of the CAZymes have an AI > 45 and were therefore probably acquired through horizontal gene transfer.

A. vaga has lost 1,250 genes compared with the inferred last common ancestor of Protostomia, the genome of which comprised at least 7,844 unique protein-coding genes (Supplementary Note E6). A total of 137 PFAM domains typically present in metazoans could not be detected in the assembled genome sequence (Supplementary Data 10). Of particular interest are missing domains involved in reproductive processes (Supplementary Note F1); for example, the *Zona pellucida*-like domain (notably found in sperm-binding proteins[25]) is present in an average of 36 copies in metazoan genomes but is absent in *A. vaga*. In contrast, we found multiple copies of most metazoan genes involved in DNA repair and homologous recombination, including a considerably divergent *Spo11* but no *Rad52* and *Msh3*.

To conclude, our analysis of a lineage of the bdelloid rotifer *Adineta vaga* reveals positive evidence for asexual evolution: its genome structure does not allow pairing of homologous chromosomes and therefore seems incompatible with conventional meiosis (Fig. 5). However, we cannot rule out that other forms of recombination occur in bdelloid populations in ways that do not require homologous pairing, such as parasexuality[26]. The high number of horizontally acquired genes, including some seemingly recent ones, suggests that HGTs may also be occurring from rotifer to rotifer. It is plausible that the repeated cycles of desiccation and rehydration experienced by *A. vaga* in its natural habitats have had a major role in shaping its genome: desiccation presumably causes DNA double-strand breaks, and these breaks that allow integration of horizontally transferred genetic material also promote gene conversion when they are repaired. Hence, the homogenizing and diversifying roles of sex may have been replaced in bdelloids by gene conversion and horizontal gene transfer, in an unexpected convergence of evolutionary strategy with prokaryotes.

METHODS SUMMARY

Genomic DNA was extracted from laboratory cultures of a clonal *A. vaga* lineage and shotgun-sequenced using 454 and Illumina platforms at respective coverage of 25 and 440 times (using both single reads and mate reads from inserts up to 20 kb). The 454 reads were assembled into contigs using MIRA[27]; the contigs obtained were corrected using single Illumina reads and linked into scaffolds using paired Illumina reads[28] (Supplementary Table 1). We annotated protein-coding genes by integrating evidence from RNA sequencing, *ab initio* predictions and comparison with UniProt. Most synteny and Ka/Ks (non-synonymous divergence/synonymous divergence) analyses were performed using the package MCScanX[29] and synteny plots were drawn using Circos[30].

Received 21 November 2012; accepted 30 May 2013.
Published online 21 July 2013.

1. Danchin, E. G. J., Flot, J.-F., Perfus-Barbeoch, L. & Van Doninck, K. In *Evolutionary Biology — Concepts, Biodiversity, Macroevolution and Genome Evolution* (ed. Pontarotti, P.) 223–242 (Springer, 2011).
2. Hsu, W. S. Oogenesis in the Bdelloidea rotifer *Philodina roseola* Ehrenberg. *Cellule* **57,** 283–296 (1956).
3. Davis, H. A new *Callidina*: with the result of experiments on the desiccation of rotifers. *Month. Microscopical J.* **9,** 201–209 (1873).
4. Segers, H. Annotated checklist of the rotifers (Phylum Rotifera), with notes on nomenclature, taxonomy and distribution. *Zootaxa* **1564,** 1–104 (2007).
5. Maynard Smith, J. Contemplating life without sex. *Nature* **324,** 300–301 (1986).
6. Ricci, C. Anhydrobiotic capabilities of bdelloid rotifers. *Hydrobiologia* **387–388,** 321–326 (1998).
7. Gladyshev, E. & Meselson, M. Extreme resistance of bdelloid rotifers to ionizing radiation. *Proc. Natl Acad. Sci. USA* **105,** 5139–5144 (2008).
8. Hur, J. H., Van Doninck, K., Mandigo, M. L. & Meselson, M. Degenerate tetraploidy was established before bdelloid rotifer families diverged. *Mol. Biol. Evol.* **26,** 375–383 (2009).
9. Rozen, S. *et al.* Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* **423,** 873–876 (2003).
10. Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144,** 27–40 (2011).
11. Vijg, J. & Dollé, M. E. T. Large genome rearrangements as a primary cause of aging. *Mech. Ageing Dev.* **123,** 907–915 (2002).
12. Birky, C. W. Jr. Heterozygosity, heteromorphy, and phylogenetic trees in asexual eukaryotes. *Genetics* **144,** 427–437 (1996).
13. Leffler, E. M. *et al.* Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biol.* **10,** e1001388 (2012).
14. Welch, D. B. M. & Meselson, M. S. Rates of nucleotide substitution in sexual and anciently asexual rotifers. *Proc. Natl Acad. Sci. USA* **98,** 6720–6724 (2001).
15. Teshima, K. M. & Innan, H. The effect of gene conversion on the divergence between duplicated genes. *Genetics* **166,** 1553–1560 (2004).
16. Gladyshev, E. A., Meselson, M. & Arkhipova, I. R. Massive horizontal gene transfer in bdelloid rotifers. *Science* **320,** 1210–1213 (2008).
17. Syvanen, M. Evolutionary implications of horizontal gene transfer. *Annu. Rev. Genet.* **46,** 341–358 (2012).
18. Hickey, D. A. Selfish DNA: a sexually-transmitted nuclear parasite. *Genetics* **101,** 519–531 (1982).
19. Arkhipova, I. & Meselson, M. Deleterious transposable elements and the extinction of asexuals. *Bioessays* **27,** 76–85 (2005).
20. Abad, P. *et al.* Genome sequence of the metazoan plant-parasitic nematode *Meloidogyne incognita*. *Nature Biotechnol.* **26,** 909–915 (2008).

21. Gladyshev, E. A., Meselson, M. & Arkhipova, I. R. A deep-branching clade of retrovirus-like retrotransposons in bdelloid rotifers. *Gene* **390,** 136–145 (2007).
22. Shirayama, M. *et al.* piRNAs initiate an epigenetic memory of nonself RNA in the *C. elegans* germline. *Cell* **150,** 65–77 (2012).
23. Krisko, A., Leroy, M., Radman, M. & Meselson, M. Extreme anti-oxidant protection against ionizing radiation in bdelloid rotifers. *Proc. Natl Acad. Sci. USA* **109,** 2354–2357 (2012).
24. Geisler-Lee, J. *et al.* Poplar carbohydrate-active enzymes. Gene identification and expression analyses. *Plant Physiol.* **140,** 946–962 (2006).
25. Bork, P. & Sander, C. A large domain common to sperm receptors (Zp2 and Zp3) and TGF-β type III receptor. *FEBS Lett.* **300,** 237–240 (1992).
26. Forche, A. *et al.* The parasexual cycle in *Candida albicans* provides an alternative pathway to meiosis for the formation of recombinant strains. *PLoS Biol.* **6,** e110 (2008).
27. Chevreux, B., Wetter, T. & Suhai, S. Genome sequence assembly using trace signals and additional sequence information. *Proc. German Conf. Bioinf.* **99,** 45–56 (1999).
28. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27,** 578–579 (2011).
29. Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40,** e49 (2012).
30. Krzywinski, M. *et al.* Circos: An information aesthetic for comparative genomics. *Genome Res.* **19,** 1639–1645 (2009).

Supplementary Information is available in the online version of the paper.

Acknowledgements The authors would like to thank M. Meselson for his support during the initiation phase of this project and for inspiring us with his seminal works on bdelloid genetics. The authors are also grateful to M. Radman for useful discussions, M. Knapen and N. Debortoli for participating in laboratory work, M. Lliros for helping with Fig. 1, S. Henrissat for participating in CAZyme analyses, and S. Oztas, B. Vacherie, P. Lenoble and S. Mangenot for performing PCR validations of the assembly. This work was supported by Genoscope-CES (where most of the sequencing was performed), by US National Science Foundation grants MCB-0821956 and MCB-1121334 to I.A., by German Research Foundation grant HA 5163/2-1 to O.H., by grant 11.G34.31.0008 from the Ministry of Education and Science of the Russian Federation to A.S.K., by grant NSF CAREER number 0644282 to M.K., by US National Science Foundation grant MCB-0923676 to D.B.M.W., by FRFC grant 2.4.655.09.F from the Belgian Fonds National de la Recherche Scientifique (FNRS) and a start-up grant from the University of Namur to K.V.D.; J.F.F. and K.V.D. thank also J.-P. Descy (University of Namur) for funding support.

Author Contributions Bo.H., X.L., and B.N. are joint second authors; O.J. and K.V.D. are joint last authors. Bo.H., X.L., F.R. and B.H.L. maintained the rotifer cultures; Bo.H., X.L., F.R. and B.H.L. prepared the genomic DNA; X.L., D.B.M.W. and B.H.L. carried out gene expression experiments; Bo.H., X.L. and B.H.L. prepared complementary DNAs; K.L., J.P. and B.H.L. carried out the sequencing; J.F.F., A.C., V.B., O.J., B.N., J.M.A. and C.D.S. assembled the genome, validated the assembly and built the gene set; J.F.F., J.M.A., V.B., G.A.B., M.D.R., E.G.J.D., O.A.V., M.K., P.W., O.J. and K.V.D. analysed the genome structure; Bo.H., E.G.J.D., M.D.R., J.F.F., A.H., Be.H., B.H.L., R.K., B.L., J.F.R., F.R., A.S.K., E.W., D.B.M.W. and K.V.D. analysed the gene families; I.A., J.B., O.P. and I.Y. annotated and analysed the transposable elements; O.C., P.G., B.W., R.B., P.P. and K.V.D. carried out orthology analysis; I.A., E.G., E.G.J.D., P.G., B.W., F.R., D.B.M.W., P.P., J.F.F. and O.J. analysed the horizontal gene transfers; O.A.V., J.F.F., G.A.B., A.S.K. and D.B.M.W. analysed the signatures of gene conversion; O.H. modelled the effect of gene conversion on Muller's ratchet; J.F.F., O.J. and K.V.D. wrote the core of the manuscript, with contributions from I.A., E.G.J.D., A.H., B.N., O.H., Be.H., Bo.H., R.K., J.M.A., J.F.R., O.A.V., M.K., A.S.K., D.B.M.W., P.P. and P.W.; and P.W., J.W., R.B., D.B.M.W., P.P., O.J. and K.V.D. designed the project and acquired funding.

Author Information The sequencing reads and assembly are available at the Sequence Read Archive (accessions ERP002115 and SRP020364 for DNA, ERP002474 and SRP020358 for cDNA) and at the European Nucleotide Archive (accessions CAWI010000000-CAWI010044365), respectively. The assembly and annotation can be browsed and downloaded at http://www.genoscope.cns.fr/adineta, whereas the result of the orthology analysis is accessible at http://ioda.univ-provence.fr/. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to O.J. (ojaillon@genoscope.cns.fr or ojaillon@mit.edu), J.F.F. (jean-francois.flot@ds.mpg.de) or K.V.D. (karine.vandoninck@fundp.ac.be).

Supplementary Notes

A. Organism background

A1. The phylum Rotifera

The phylum Rotifera is a part of the Protostomia (basal-branching triploblasts, Figure 1). It is generally divided into three taxa: the Seisonidea, the Monogononta, and the Bdelloidea[1]. Rotifers are typically described as small, free-living metazoans (less than 1 mm) distinguished by a ciliated wheel organ on their head (the corona), a jaw-like grinding organ (the mastax) and bilateral ovaries. They also have ganglia, muscles, digestive and excretory systems, photosensitive and tactile sensory organs, structures for crawling or swimming, and a foot with pedal glands. The taxon Seisonidea consists of four[2] described species in two genera; they reproduce sexually and are exclusively marine, generally living on leptostracan crustaceans[3]. Monogononta are found mostly in freshwater but also in soil and marine environments. The monogononts possess a single gonad and reproduce through cyclical parthenogenesis with diminutive haploid males fertilising sexual females to produce diploid resting eggs. Bdelloidea are found mostly in freshwater and ephemerally aquatic habitats; males are absent and females reproduce by mitotic parthenogenesis[4,5]. In addition, Acanthocephala is a closely allied taxon that should be placed in Rotifera according to most molecular studies[6-9]. Acanthocephalans (thorny-headed worms) are a group of obligate endoparasitic animals that parasitise arthropods or molluscs in their early life stage and vertebrates as adults. Acanthocephalans possess an evertable proboscis in the anterior part of the body; no wheel organ is observed. However, their morphology has become drastically simplified due to their parasitic life style and, as a consequence, it is difficult to establish relationships using traditional morphological characters. All studied acanthocephalans are obligate sexuals.

A2. Choice of rotifer species for genome sequencing

There are more than 460 described bdelloid species classified based on morphology into 4 families and 19 genera[10]. Despite much observation since Van Leeuwenhoek[11], no males, vestigial male structures or hermaphrodites have ever been observed. Hsu[4, 5] studied the oogenesis of two bdelloid rotifer species belonging to two different families and reported in both species the formation of eggs by two mitotic divisions of oocytes with no signs of chromosome pairing and no reduction in chromosome numbers. In addition to their high taxonomic rank as obligate asexuals and their abundance of species, the Bdelloidea were shown to be of ancient origin because bdelloid fossils were found in amber dating from 35 to 40 million years[12]. Another remarkable feature of bdelloid rotifers is their high resistance to desiccation at any stage in their life cycle, unlike monogonont rotifers that can only survive dehydrated conditions as resting eggs. When humidity decreases, bdelloids contract into a compact tun shape and enter a metabolically quiescent state (anhydrobiosis). They can survive prolonged periods of desiccation and come back to life when water is present again[13]. Moreover, bdelloid rotifers are also extraordinarily resistant to ionising radiation, being able to survive and resume reproduction after doses that cause hundreds of DNA double strand breaks (DSB) per genome[14]. By analogy with the desiccation- and radiation resistant bacterium *Deinococcus radiodurans*[15,] it is likely that the extraordinary radiation resistance of bdelloid rotifers reflects an adaptation to survive desiccation in their characteristic ephemerally aquatic habitats and that the damage incurred during desiccation includes DNA breakage. Finally, foreign genes were found at subtelomeric regions in bdelloid genomes, suggesting horizontal gene transfer[16].

A partial genome assembly was previously reported for the monogonont *Brachionus ibericus*[17] but has not yet been deposited in any public database and appears very fragmented. For the reasons outlined above, we chose a species of Bdelloidea to represent the first high-coverage, public rotifer genome. We selected *Adineta vaga* for several reasons. First, a limited amount of genomic data was already available for this species[18,19]. This species had also been screened for the presence of foreign genes and for the presence of

specific transposon families[20,21]. Genome sizes had been determined for several bdelloid species and *A. vaga* had one of the most compact genomes, equivalent to 0.25 pg DNA[22] and the karyotype of *A. vaga* was known to contain 12 chromosomes all similar in size and in morphology[23].

B. Next-generation sequencing

B1. *Adineta vaga* culture and DNA extraction

The laboratories of Karine Van Doninck (University of Namur) and David Mark Welch (MBL) maintain clonal cultures of *Adineta vaga* started from a single individual obtained from the Meselson laboratory at Harvard University. *Adineta vaga* are grown in filter-sterilised commercial spring water and fed *Escherichia coli*.

For genomic DNA extraction at Namur, eggs of *Adineta vaga* were obtained after cleaning dense bdelloid cultures with bleach (5% final). This method is very effective in the destruction and elimination of adult rotifers, bacteria and detritus and yields clean bdelloid eggs (stored at -80°C). Genomic DNA is extracted from frozen bdelloid eggs, grounded in liquid N_2, followed by a standard phenol:chloroform extraction. Aliquots of genomic DNA were sent to Genoscope (Evry, France) for high-throughput sequencing.

For genomic DNA extraction at MBL, animals were starved for 48 hours, then harvested by adding NaCl to a final concentration of ~100mM (to induce the animals to detach from the substrate), transferring rotifers and media to 50-mL sterile centrifuge tubes, and pelleting the animals by centrifugation (2 minutes at 500 rcf, followed by transfer to 1.5ml tubes and centrifugation for 2 minutes at 16,000 rcf to pellet the animals). Supernatant was removed and total DNA immediately extracted with DNeasy Blood and Tissue Kit (Qiagen, Valencia,CA) per manufacturer's protocol. After quality assessment on an agarose gel and quantitation using a Nanodrop 2000 (Thermo Scientific, Wilmington, DE), ~36 µg gDNA were ethanol-precipitated and sent to Genoscope (Evry, France) for sequencing.

B2. cDNA preparation

At the University of Namur, cDNA libraries were constructed from *A. vaga* individuals in various biological states of desiccation: rehydrated control individuals, dried individuals, rehydrated individuals having been subjected to different periods of desiccation (7, 14 and 28 days). In addition, two cDNA libraries were constructed from irradiated *A. vaga* individuals using a ^{137}Cs γ-radiation source. Control and dehydrated *A. vaga* cDNAs were produced with SuperScript II Reverse Transcriptase (Invitrogen) using 1-5 µg RNA template in the first strand synthesis reaction. cDNAs were sent to Genoscope (Evry, France) for high-throughput sequencing.

At MBL, cDNA libraries were constructed from healthy, normally hydrated *A. vaga* cultures containing animals at all life-stages. After removal of supernatant, ~100 mg of animals were immediately resuspended in TRIzol® (Invitrogen, Carlsbad, CA) and extracted following the manufacturer's protocol with the following options: a glass dounce homogeniser was used to quickly disrupt the tissue; glycogen (Ambion/Life Technologies, Grand Island, NY) was used as a co-precipitant; and the RNA wash in 75% ethanol was extended to ~16 hours at -20°C to remove salts. A total of ~96 µg *A. vaga* RNA was extracted and subjected to two rounds of poly-A selection (1 hr incubation for the 1st round, 45 min incubation for the 2nd round) with Ambion's MicroPoly(A) Purist Kit (Life Technologies), which yielded ~2 µg mRNA. cDNA libraries were prepared with a cDNA Synthesis System Kit (Roche-454, Branford, CT) and a cDNA Rapid Library Preparation Kit (Roche-454) per the cDNA Rapid Library Preparation Method Manual, GS FLX Titanium Series, January 2010 revision, including careful quality assessments as recommended. The final library met all quality metrics as defined by Roche, and library quantitation was performed on an Agilent 2100 Bioanalyzer with a High-Sensitivity DNA kit (Agilent, Santa Clara, CA) prior to emPCR titrations.

B3. Genome and cDNA sequencing

For Illumina libraries sequenced at Genoscope, DNA was sonicated using the S2 Covaris instrument (Covaris, Inc., USA). Single end libraries were prepared following Illumina's protocol (Illumina DNA sample kit). Briefly, fragments were end-repaired, then 3`-adenylated, and Illumina adapters were added. Ligation products of 350-400 bp were gel-purified, and size-selected DNA fragments were PCR-amplified using Illumina adapter-specific primers. Libraries were purified and then quantified using a Qubit Fluorometer (Life technologies); libraries profiles were evaluated using an Agilent 2100 bioanalyser (Agilent Technologies, USA). Each library was sequenced using 76-bp read chemistry in a single flow cell on an Illumina GA IIx (Illumina, USA).

For 454 libraries sequenced at Genoscope, DNA was fragmented to a range of 5-10 kb or 18-22 kb using a HydroShear instrument. Fragments were end-repaired and extremities were ligated with 454 circularisation adaptors. Fragments were size selected respectively to 8kb or 20 kb through regular gel electrophoresis, and circularised using Cre-Lox recombination. Circular DNA was fragmented again by nebulisation. Fragments were end-repaired and ligated with library adaptors. Mate-pair libraries were amplified and purified. Single-stranded libraries were isolated, then bound to capture beads and amplified in an oil emulsion (emPCR). Libraries were then loaded on a pico-titer plate and pyrosequenced using a GS FLX according to the manufacturer's protocol.

The cDNA library produced at MBL was sequenced on a Roche 454 GS-FLX, yielding 865,534 reads (average trimmed length 383 bp, modal trimmed length 470 bp). Reads were assembled using Newbler v2.5 with the -urt option, resulting in 26,607 isotigs and 21,551 isogroups.

C. Genome assembly and annotation

C1. Genome assembly

Genome assembly was performed in three steps: assembly of 454 data into contigs, correction of the contigs using Illumina data, and assembly of the corrected contigs into scaffolds using Illumina data.

For the first step, we used the multi-pass assembler MIRA[24] version 3.2.1_prod (normal mode, default options except the number of cycles) to generate contigs from the 454 genomic libraries (one single-read and three paired-end 3kb, 8kb and 20kb libraries, for a total coverage of about 25X). Eight cycles were performed to separate a maximum of repeats and polymorphic regions. We subsequently used Illumina data to correct the homopolymer errors of the 454 contigs following a standard procedure[25]. The corrected contigs were linked into scaffolds using the program SSPACE[26] (parameters k = 4 and a = 0.7, no contig extension) with three Illumina libraries of insert sizes ranging from 425 bp to 11 kb. The first library had 525616 (70.0%) satisfied read pairs (in terms of distance and orientation) and 259284 (30.0%) unsatisfied ones; the second library had 243179 (46.3%) satisfied pairs and 282539 (53.7%) unsatisfied one; and the third library, with the largest insert sizes, had 320960 (40.0%) satisfied pairs and 481121 (60%) unsatisfied ones.

C2. Validation of the assembly

We assessed the validity of our assembly by focusing on the following two points: first, we verified whether our contigs were consistent (i.e. colinear) with previously published genomic regions of *A. vaga*. Second, we took a closer look at two peculiar features of this genome (the presence of palindromes and of colinearity breakpoints between allelic regions) to make sure that they were not artefacts resulting from the assembly procedure.

Alignment of the scaffolds with published genomic regions

Previous genomic studies on *A. vaga* based on fosmid sequencing often focused on telomeric regions replete with transposable elements. These types of genomic regions are poorly assembled using whole genome shotgun approaches, ending up fragmented in many very small contigs. We therefore built our reference sequence set according to the following protocol:
- among the 321 *A. vaga* nucleotide sequence entries at NCBI, we retained 28 that were longer than 30 kb;
- 13 sequences that contained annotation of telomeric repeats were removed from the set of sequences ;
- 4 sequences that contained annotated transposable elements we removed from the set of sequences;
- the 11 remaining sequences were used to validate our assembly.

Our validation dataset totalled 500 kb and contained 176 genes. These eleven reference sequences are referenced in EMBL by the following accession numbers: GU373045, GU373047, EU850438, EU831279, EU652315, EU652316, EU643484, EU850438, EU643476, EU643474, EU637018 and EU637017. We aligned these sequences on the genome assembly using BLAT and/or BLAST. Each of our eleven control sequences aligned to a single scaffold of our assembly, with perfect global continuity and no local inversion (Supplementary Fig. 1). The only differences we noticed were the absence in our scaffolds of 3 regions of the reference sequences:
- one 10 kb region of sequence GU373045 was missing in the corresponding scaffold av690 as it fell in an assembly gap between two contigs. This region contains one gene corresponding to a leucine-rich repeat-containing protein that matches numerous other locations in the assembly;
- one 3 kb region of sequence GU373047 was missing from the corresponding scaffold av744. This region does not contain any annotated gene. A match to this short region was detected in scaffold av1973.
- one 8 kb region of sequence EU643474 was missing from the corresponding scaffold av344. This region contains annotations of two putative genes: one similar to a TPR-repeat-containing protein, and a second one similar to the glutamate receptor GLR3.3. Those genes are however present in the corresponding allelic region of scaffold av34.

Overall, the colinearity of the scaffolds with published genome regions was excellent. The average similarity level between the eleven scaffolds and the corresponding reference sequence was 98.3% (average of blast identity weighted according to the lengths of the alignment) (See Supplementary Fig. 1). Our scaffolds contained 97% of genes of the matching reference sequences, and 96% of their nucleotides.

Validation of palindromic regions
Since our analysis of the genome structure revealed the frequent occurrence of palindromic regions (inverted repeats), we verified them by examining the depth of coverage by reads and read inserts along the assembly. Our rationale was that the presence of mate reads in wrong orientation or unexpected distance would be a signature of errors in the assembly. The distribution of correct mate pairs along the palindromes, including their central parts, did not deviate noticeably when compared to the rest of the assembly.

Validation of breakpoints between allelic regions
We identified 869 blocks involved in colinearity breakpoints between alleles. These blocks fell in 301 scaffolds covering 109.7 Mb of the genome. This is certainly an underestimation since breakpoints involving small scaffolds are not detectable. Most of the longest scaffolds contain breakpoints: 18 of the 20 longest scaffolds, 46 of the 50 longest, and 91 of the 100 longest. Since the chromosomes of *A. vaga* are all of similar size[23], any scaffold has the same probability to fall on any of them; therefore most chromosomes, if not all, have no full-length homologue in the genome.

To validate this result, we analysed 72 examples of breakpoints and tested their correctness by examining the continuity of the insert pair coverage along the corresponding genomic

regions. At places where insert coverage was low, we also performed PCR and resequencing to confirm those scaffolds. Using these two complementary approaches we succeeded to get a complete physical validation for 31 breakpoints out of 72, thereby ruling out that this important feature of the genome of *A. vaga* resulted from errors during the assembly process.

Assessment of repeat separation and genomic variation
To assess whether the contigs obtained from the assembly of the 454 data contained a higher-than-expected amount of variation between reads (that would suggest misassemblies of repeats and/or differences in genome sequence among the DNA extracts we sequenced), we used SAMtools[27] to generate a pileup of all bases at each position of the contigs. The distribution of this variation among consensus bases is shown in Supplementary Fig. 2. Assuming an error rate of ~1% for the 454 Titanium platform[28], 96.3% of the consensus bases had less variants than expected and 3.7% of the bases had more variants than expected. The average percentage of variants per consensus base was 0.31%, in line with other published studies[29]. Indeed, the program we used to assemble our 454 reads into contigs looked in detail at the variation at each site and returned IUPAC ambiguity characters at position where the variability exceeded expectations: there were less than one ambiguous base per 15,000 positions in the assembly (one per 30,000 when considering only contigs larger than 500 bp). These figures demonstrate that the separation of repeats in our assembly was excellent and that variation was very low in and among the DNA extracts we sequenced.

We also checked the percentage of repeats (including duplicated genome regions) that may have been fused in the assembly because they were identical or nearly identical (and would therefore not have been detected by the analysis above). To do so, we remapped part of the genomic data on the scaffolds using bowtie2[30], turned the resulting SAM alignment into BAM and sorted it using SAMtools[27], then calculated the per-base coverage using BEDtools[31].

As can be seen in Fig. S3 (which presents the distribution of coverage of the genome), most repeats assembled separately. However, some scaffolds had two-fold coverage (e.g. scaffolds 6 and 9). Overall, 26 Mb of the assembly (i.e. 11.2%) has a coverage higher than 1.5 times the normal coverage: if we assume that all of these are genome regions that became fused during the assembly process because they were identical or nearly identical, then the total genome size is 26 +218 = 244 Mb, which corresponds to the independent size estimate obtained using fluorometry[22].

C3. Gene prediction procedure
To construct gene models we used a rationale applied and published previously for other metazoan genomes[32], consisting of combining evidences provided by diverse resources.

Repeat masking
Most of the genome comparisons were performed using repeat-masked sequences. For this purpose, we searched and masked sequentially several kinds of repeats: known rotifer repeats and transposons available in Repbase with the Repeat masker program[33], tandem repeats with the TRF program[34], and *ab initio* repeat detection with RepeatScout[35]. This pipeline masked 2.8% of the assembly.

Protein alignments
We extracted a pool of 2,187,378 metazoan protein sequences from UniProt[36] to detect conserved genes between *A. vaga* and other species. As Genewise[37] is computationally expensive, we first aligned the protein database with the *A. vaga* genome assembly using BLAT[38]. Subsequently, we extracted the genomic regions where no protein hit had been found by BLAT and realigned Uniprot proteins with more permissive parameters. We then refined each significant match using Genewise in order to identify exon/intron boundaries.

RNA-Seq

We generated 76 bp Illumina reads from the cDNA libraries constructed at Namur, and mapped 233,235,652 usable reads (after quality and adaptor trimming) to the *A. vaga* genome using SOAP2[39] with default parameters. Reads that aligned on exon-exon junctions could not be mapped to the genomic sequence. To improve read mapping, we split each 76 bp unmapped read into two halves and relaunched the mapping using SOAP2, mapping 79% of usable reads. Running Gmorse[32] on the SOAP2 mapping and unmapped read resulted in 121,154 transcript models.

Isotigs from the MBL cDNA assembly were aligned against genomic scaffolds by Blat, using default parameters. To refine the Blat alignment, we used Est2Genome[40]. We required 90% identity and only the best match for each contig was retained. This method aligned 99% of the contigs on the genomic assembly, with an average identity of 99%.

Rotifera ESTs
We used BLAT[38] to align a collection of 53,322 public Rotifera mRNAs (downloaded from the EMBL database) with the *A. vaga* genome assembly, using default parameters between translated genomic and translated ESTs. To refine BLAT alignment, we used Est2Genome[40] with an identity threshold of 50%. For each mRNA sequence we selected the best match and all matches with a score higher than 90% of the score of the best match.

Ab initio gene predictions
We trained the SNAP[41] *ab inito* gene prediction programme on open reading frames derived from the Gmorse transcript models. SNAP subsequently predicted 94,395 gene models in the genome assembly.

Integration of resources using GAZE
We broke down the individual predictions from each of the programs described above (SNAP, Genewise, Est2genome, Gmorse) into segments (coding, intron, intergenic) and signals (start codon, stop codon, splice acceptor, splice donor, transcript start, transcript stop) that we used to automatically build gene models using GAZE[42]. Exons predicted by SNAP, Genewise, Est2genome, Gmorse were used as coding segments. Introns predicted by Genewise, Est2genome and Gmorse were used as intron segments. Intergenic segments were created from the span of each mRNA, with a negative score to coerce GAZE into not splitting genes. We defined predicted repeats as intron and intergenic segments, to avoid prediction of genes coding proteins in such regions. The whole genome was scanned to find signals (splice sites, start and stop codons).

GAZE used exon boundaries from Genewise, Est2genome or SNAP only if it chose the same boundaries. Each segment or signal from a given program was given a value reflecting our confidence in the data, and these values were used as scores for the arcs of the GAZE automaton. All signals received a fixed score, but segment scores were context-sensitive: coding segment scores were linked to the percentage identity (%ID) of the alignment; intronic segment scores were linked to the %ID of the flanking exons. We assigned a weight to each resource to further reflect its reliability and accuracy in predicting gene models, and multiplied the score of each information source by this weight before processing by GAZE. Finally, gene predictions created by GAZE were filtered according to their scores and their lengths.

C4. Automatic functional annotation
We use HMMER 3.0 to compare the proteome of *A. vaga* as well as those of 12 other species (*Lactobacillus acidophilus* 30SC uid63605, *Deinococcus radiodurans* R1 uid57665, *Saccharomyces cerevisiae, Neurospora crassa, Meloidogyne incognita, Caenorhabditis elegans, Strongylocentrotus purpuratus, Nematostella vectensis, Lottia gigantea*, *Drosophila melanogaster, Oikopleura dioica, Homo sapiens*) against the PFAM-A database of manually curated Hidden Markov Models (HMMs). To avoid introducing biases in case of multiple splice variants we kept only one model per locus (the longest one).

In the whole proteome of *A. vaga*, a total of 74,431 PFAM domains were assigned to 32,721 different genes (Supplementary Data 10). Hence, 66.37 % of the 49,300 predicted proteins of this species were assigned at least one PFAM domain. This proportion is comparable to that typical of other proteomes annotated (54-84% of proteins with at least one PFAM domain assigned).

Mapping of PFAM to GO term was performed using pfam2go (www.geneontology.org).

C5. Annotation of repeats and transposable elements

TE identification

For initial screening, we mostly followed the scheme of Kapitonov and Jurka which was used to characterise TE complements in *Nematostella vectensis*[43] and *Xenopus tropicalis*[44]. Briefly, DNA fragments with homology to known types of reverse transcriptases, transposases, and DNA polymerases from Repbase[45] (http://www.girinst.org/repbase) were detected using WU-BLAST/CENSOR and clustered with BLASTclust, after which element boundaries were extended where possible, using the knowledge of overall element structure, and verified as discontinuities in a multiple query-anchored alignment in a BLASTN search against *A. vaga* scaffolds (-m 1). Strikingly, in many cases TE boundaries were visible in a reverse pattern to what is normally observed, i.e. a single-copy TE would be embedded into a more repetitive genomic environment. Low copy number often resulted in unreliable consensus sequences, in which case preference was given to an apparently intact genomic copy. The initial TE set was then supplemented with additional sequences using outputs from the TEdenovo pipeline in the REPET package[46] and from ReAS[47], which were processed in the same manner to extend each TE to its boundaries wherever possible. Consensus sequences were kept if reliable; otherwise, preference was given to intact copies. Neither REPET nor ReAS could identify 44 single-copy TE families, as these programs rely on identification of over-represented sequences. In addition to known TEs, which represented about one-half of the REPET output, it also contained 140 tandem repeats, 226 host genes, and 232 unclassified repeats of uncertain origin, with redundancy between clusters. The ReAS output added 8 extra sequences to the TE set. On balance, these findings underscore the need to apply several independent complementary approaches for *ab initio* TE identification in non-model organisms.

Using the compiled TE set, the overall TE content in the assembly was estimated with RepeatMasker (rmblat engine), which yielded a total coverage of 3.03% without further validation, and with BLAT after removal of overlapping hits, hits shorter than 50 bp, and correction for secondary insertions (2.24%). We chose to present only validated hits from the second estimate in the summary table (Supplementary Table 3), as RepeatMasker yielded numerous short fragments that in many cases could not be confirmed as TEs (*e.g.* were located within known CDS).

TE content

Despite its low overall TE content, the *A. vaga* genome harbors a large diversity of TE families. We identified 23 families of LTR retrotransposons (0.58 Mb), 24 families of non-LTR retrotransposons (0.71 Mb), 23 families of *Penelope*-like retroelements (0.72 Mb) and 184 families of DNA transposons (2.7 Mb). Although DNA TEs exhibit much higher family diversity (184 vs 70 families), they occupy only a slightly larger fraction of the assembly than retrotransposons (2.7 vs 2.0 Mb), due to their generally shorter length (Supplementary Table 3 and 4). Characteristically, the most abundant TE families are those which have already been described: *Juno*, *Vesta*, *Hebe*, *R9* and *Athena* retroelements, and *mariner/Tc*, *hAT*, *piggyBac*, and *Helitron* DNA TEs[20, 21, 48-50]. The high diversity of TE families, accompanied by extremely low copy numbers per family, apparently prevented their detection in earlier PCR screens with degenerate primers aimed at detecting TEs present in high copy number[51].

While the longest gene-rich scaffolds carry very few TE insertions per Mb, the overall density of TEs per Mb is generally higher in smaller scaffolds, indicating that TEs are concentrated towards genome periphery and in poorly assembled regions (Fig. S5). However, the overall

increase per bin is non-uniform with regard to individual scaffolds: TE distribution is highly uneven among scaffolds, and within some of the larger scaffolds there is a clear compartmentalisation between gene-rich and TE-rich regions (e.g. scaffolds av24, av37, av88, av136, av270).

Paradoxically, the copy number for members of the most prominent multigene families, such as LRR, TPR, PPR, Kelch, NHL, and FG-GAP repeat-containing proteins, exceeds that for most TE families, and repeats of genic origin make a substantial contribution to the overall *A. vaga* repeat content (over 1Mb). In addition, the assembly contains unclassified short repeats with no obvious relationship to known TEs.

D. Synteny computation and analysis

D1. Detection of colinear blocks

Preliminary Oxford grids[52] revealed numerous duplicated regions with conserved gene order (Supplementary Fig. 6). More detailed analyses were conducted using the MCScanX package[53]. Out of the 49,300 annotated genes, 5,915 (12%) were detected as singletons, 11,111 (22.5%) as dispersed duplicates, 244 (0.5%) as proximal duplicates, 298 (0.6%) as tandem duplicates, and 31,732 (64.4%) as segmental duplicates putatively resulting from whole genome duplication.

D2. Identification of ohnologues and alleles

For each block of two colinear regions we computed various statistics including the Ks between homologues, i.e. their divergence at synonymous positions, as well as a colinearity metrics defined as the number of colinear genes in the block divided by the total number of genes present in the colinear regions. The distributions of Ks and colinearity were both bimodal and strongly correlated (Fig. 2A). Moreover, the points corresponding to the larger blocks were tightly grouped whereas the points corresponding to the smaller blocks displayed a larger variance in average Ks and colinearity (Supplementary Fig. 7A). We used the ratio (average Ks)/colinearity to divide the set of colinear blocks into two groups: allelic blocks with a ratio (average Ks)/colinearity <0.5, and ohnologous blocks with a ratio (average Ks)/colinearity>0.5 (Fig. S7B). Lists of relevant information for allelic and ohnologous blocks, as well as allelic and ohnologous gene pairs, are available in Supplementary Data 1, 2, 3 and 4, respectively.

D3. Inverted segmental repeats (palindromes)

The genome of *A. vaga* comprises 17 large palindromes (colinear regions found in opposite directions on the same scaffold). To shed light on the origin of these features, we compared the distribution of Ks of the genes in the arms of the palindromes with those of all pairs of colinear genes in the genome (Supplementary Fig. 8). The Ks signature of the gene pairs in the palindromes was found to be identical to the one of alleles, suggesting that these large inverted repeats did not arise by recent reduplication of genomic regions but rather by shuffling of alleles caused by rearrangements of the genome. Dotplots showing the internal structure of some representative palindromes are displayed in Supplementary Fig. 9.

E. Genome dynamics

E1. Evidence for gene conversion between alleles

The fact that the observed nucleotide divergence between the alleles of *A. vaga* is much lower than predicted in the long-term absence of meiotic recombination[54] suggests that gene conversion (resulting for instance from break-induced repair or synthesis-dependant strand annealing) may be playing an important role in preventing the accumulation of sequence divergence in the genome of *A. vaga*. To test this hypothesis, we computed alignments of allelic regions of the genome, then looked at the distribution of distances between interallelic mismatches and compared it with the one expected in the absence of gene conversion.

For the first step (computation of allelic alignments), we selected colinear blocks of more than ten genes with over 70% synteny and aligned them locally using BLASTZ[55] with stringent parameter (Y=1000). We then used the program single_cov2 from the TBA/MULTIZ package[56] to select for each segment of reference sequence only the best hit among all query sequences. We further got rid of redundant hits and kept only best bi-directional alignments: if a given alignment segment between contigs A and B was found in alignment of scaffold A vs its colinear scaffolds as well as in alignment of scaffold B vs its colinear scaffolds, we kept only the alignment with the longer reference scaffold. However, if a given alignment segment between contigs A and B was found in an alignment of scaffold A vs its colinear scaffolds, but not in an alignment of scaffold B vs its colinear scaffolds, we discarded this alignment. The final set of alignments was made of 415 blocks of total length 47 Mb and included about 43% of the assembly.

We then looked at the distribution of the distances between successive mismatches in the alignments: the average identity track length across all blocks was 30.09bp (SD=54.20) and the longest was 1064 bp long. In the absence of gene conversion, this distribution would be exponential (provided that the per nucleotide mutation rate is uniform[57]) since in such case the differences between the genomes would be generated by a homogeneous Poisson process. Gene conversion, however, would change this distribution (unless all conversion tracks are just one nucleotide long). Indeed, the observed distribution of distances between successive intergenomic differences did not match the exponential distribution but was instead characterised by an excess of very short and very long distances (Figure 4A).

Let us assume that mutation, which creates differences between allelic sequences in the genome, and gene conversion, which eliminates such differences, are at equilibrium. Then $\mu = k*L*x$, where μ is the mutation rate per nucleotide per generation, k is the gene conversion rate per nucleotide per generation (understood as the probability that a gene conversion event begins at a nucleotide site in the course of a generation), L is the average length of a gene conversion track, and x is the probability that a given position carries a difference (i.e. the divergence between the two sequences). The probability of a nucleotide to be converted ($k*L$) can be estimated directly: $k*L = \mu/x$ hence, if $x = 0.04$ (4% nucleotide divergence between alleles) then $k*L = 25\,\mu$. The probability of a nucleotide to be converted is therefore a least one order of magnitude greater than its probability to mutate.

To try to infer the distribution of conversion track lengths from the observed distribution of identity track lengths, we conducted simulations of the joint action of mutation and gene conversion on a diploid genome and looked at the resulting distribution of the lengths of conversion tracks. The simulations were carried out as follows. We started from a sequence of length 50,000,000 containing only As (matches). This sequence was subject to 10,000,000 generations of mutation and gene conversion. In each generation, in the course of mutation a A became a B (mismatch) with probability $\mu = 10^{-6}$. After this, gene conversion converted λ consecutive letters to As with probability $k = \mu / \lambda .x$, where x is the average divergence between alleles.

We calculated the distribution of distances between successive B's in the simulated sequences. Simulations in which the length of gene conversion track λ was constant gave us a rough idea of the characteristic track length in *A. vaga*, which was likely to be below 200 nucleotides. Still, the observed distribution could not be reproduced precisely under any fixed λ. In contrast, simulations in which the lengths of gene conversion tracks were drawn from distributions with two possible values produced very good fits. For example, if $\lambda = 30$ with probability 0.95, and $\lambda = 220$ with probability 0.05, simulations produced a distribution of the distances between successive mismatches that was very close to the observed one. Thus, it appears that the real distribution of λ contains a long right-hand tail, corresponding to rare long gene conversion tracks, although the average track length is not high.

In addition to tracks of longer-than-expected identity between alleles, the assembly contains also long tracks (up to 1 Mb) of two-fold coverage that can only be explained by the fusion during the assembly of identical or near-identical regions, i.e. long tracks of sequence

identity. Such long tracks of identity were reported previously from the genome of *A. vaga*[18]. Whereas the size of the smaller tracks are consistent with gene conversion from repair processes such as synthesis-dependent strand annealing[58], the longer tracks may have resulted from break-induced replication[59] (that has been reported to result in gene conversion over 100kb in size[60]), from reciprocal (mitotic) crossover[61], or simply from the accumulation of many DNA repair events in a region of the genome particularly prone to such damage.

E2. Effect of gene conversion on Muller's ratchet

In the following, we devise a simple population genetic model similar to the one of Connallon and Clark[62] to see whether gene conversion can efficiently slow down Muller's ratchet.
Consider a population of N individuals with a genome of L diploid loci. We assume that each locus is made up of a combination of two possible alleles, a wild-type allele A and a less fit mutant allele B. The three possible states of the locus, their relative fitness and the transitions between them due to mutations and gene conversion are summarised in Supplementary Fig. 11A. Gene conversion is modelled such that if gene conversion occurs at one locus it replaces at random one allele by the other. This can lead to the removal of a mutant allele B if the second allele is wild type (A). Note that it is the absence of a BB → AB transition that allows Muller's ratchet to click even in the presence of gene conversion: if all individuals carry at least one locus homozygous for the mutated type, the fittest genotype cannot be restored, as illustrated in Supplementary Fig. 11B. Our goal is to estimate how much slower the ratchet mechanism deteriorates fitness in our model with gene conversion compared to the one without.

Muller's ratchet is by now well understood theoretically for asexual, haploid models without gene conversion under the assumptions of multiplicative epistasis and identical mutational effect sizes[63-66]. In these models, the click rate depends crucially on the deterministic equilibrium number n_0 of individuals in the fittest class, and the mean fitness difference s_0 between the two fittest classes[64]. If $n_0 s_0 << 1$, the ratchet clicks frequently at a rate which is on the order of the genomic deleterious mutation rate. In this regime, selection is inefficient and cannot substantially reduce the speed of the ratchet below the neutral expectation. In the opposite regime $n_0 s_0 >> 1$, the time between clicks is exponentially small in the parameter $n_0 s_0$: $\ln(\text{click time}) \propto n_0 s_0$. In this "slow-ratchet" regime, the rate of clicks depends on the (long) time it takes until a rare fluctuation due to genetic drift is able to drive the fittest class to extinction against the stabilising force of selection. A system protected against Muller's ratchet should be in this regime and hence characterised by $n_0 s_0 >> 1$. In the following, we show that the *effective* $n_0 s_0$ in our model is much larger with gene conversion than without, and that it is much larger than 1 for a broad range of parameter values.

In order to map our problem to the established haploid models, we need to determine an effective size n_0 of the fittest class and the effective selective difference s_0 between the two fittest classes of our model. Crucially, the relevant population n_0 is not the number of fittest genotypes but the number of all genotypes from which the fittest genotype can be restored by gene conversion. We call this class the "restorable class", and it comprises all genomes that have no locus in the homozygous deleterious state, $n_{BB} = 0$ (Supplementary Fig. 11B).

Thus, we need to calculate the deterministic equilibrium number n_0 of all individuals that have no loci in the BB state. In the deterministic limit, the allele frequencies at different loci evolve independently. Therefore, we may first determine the equilibrium at a single locus and then infer the multilocus frequency distribution as a product of single locus allele frequency distributions.

In equilibrium, the per generation change due to natural selection and mutations balance each other. Mathematically, this can be written as

$$0 = (1/W - 1)p_{AA} - 2\mu p_{AA} + \gamma p_{AB} \qquad (1)$$

$$0 = ((1 - hs)/W - 1)p_{AB} - (\mu + 2\gamma)p_{AB} + 2\mu p_{AA} \quad (2)$$

$$0 = ((1 - s)/W - 1)p_{BB} + (\mu + \gamma)p_{AB} \quad (3)$$

In each of the above equations, the first term on the right hand side is the change in genotype frequencies due to selection and the others represent the change due to mutations and gene conversion. Deleterious mutations A→B occur at rate μ, and gene conversion at a rate γ. The parameter *s* measures the fitness detriment of the homozygous mutated state (BB); *h* parameterises dominance effects. The mean relative fitness W appearing in these equations equals $p_{AA} + (1 - hs)p_{AB} + (1 - s)p_{BB} = 1 - hsp_{AB} - sp_{BB}$.

We assume in the following that the per-locus rates of mutation and gene conversion are much smaller than the involved selective differences ({μ, γ} << {s, hs}), which themselves are assumed to be small compared to the overall fitness (s << 1). Furthermore, we assume that the frequency of the homozygous deleterious state (BB) is small (p_{BB} << 1). Then, the equilibrium frequencies of the three states are given by

$$p_{AB} \approx 2\mu/(hs + 2\gamma + \mu) \approx 2\mu/(hs) \quad (4)$$

(This equation holds only if it yields a value smaller than 1, otherwise $p_{AB} = 1$.)

$$p_{BB} \approx p_{AB}(\mu + \gamma)/s \; . \quad (5)$$

To proceed to the genomic level, we need to calculate the number of individuals with none of their L loci in the homozygous deleterious state (BB). In the deterministic equilibrium, the number of BB loci follows a Poisson distribution with mean Lp_{BB} because different loci evolve independently for multiplicative epistasis. Hence, the population fraction $P_{r.c.}$ of the restorable class is given by

$$P_{r.c.} = \exp(-Lp_{BB}) = \exp(-Lp_{AB}(\mu + \gamma)/s) \quad (6)$$

The mean fitness differential s between the restorable class and the first non-restorable one (n_{BB} =1), is given by

$$s_0 = s - hsp_{AB} \approx s \; . \quad (7)$$

The control parameter $n_0 s_0 = P_{r.c.} N s$ of the ratchet is therefore estimated to be

$$n_0 s_0 \approx N s \exp(-Lp_{AB}(\mu + \gamma)/s) \quad \text{(with gene conversion)} \; . \quad (8)$$

This is our main result for the gene conversion model. If the parameters are such that $n_0 s_0 >> 1$, the ratchet will be very slow and Muller's ratchet a weak effect. We argue that this will typically be the case if the mutations are not effectively neutral. Recall that μ and γ are mutation and gene conversion rates per locus. For the size of the locus, one should take the typical length of gene conversion tracks, say $10^2 - 10^3$ base pairs. With a gene conversion rate of 10^{-8} per base pair similar to typical (total) mutation rates, we may estimate $\mu + \gamma \approx 10^{-6} - 10^{-5}$. Hence, the absolute value of the exponent in (8) will typically be small unless we consider tiny selection coefficients. As a consequence, $n_0 s_0$ will be on the order of N s, the product of the total population size and the selection coefficient. Hence, the ratchet will be slow if N s is larger than 1, i.e. if the mutations are not effectively neutral.

In order to appreciate the effect of gene conversion, we compare our result to the case of no gene conversion. Then, the clicks are determined by the size of the fittest class, $n_{AB} = n_{BB} = 0$, which is generally smaller than the class restorable through gene conversion, $n_{BB} = 0$. The population fraction $P_{m.f.}$ of the fittest class is given by

$$P_{m.f.} \approx \exp(-L(p_{BB} + p_{AB})) = N \exp(-(\mu/s + 1)Lp_{AB}) \; . \quad (9)$$

The fitness differential s_0 between the unloaded and first loaded class will be hs. Therefore, the parameter $n_0 s_0$ controlling the ratchet speed is given by

$$n_0 s_0 \approx N\, hs \exp(-Lp_{AB}) \qquad \text{(without gene conversion)} \qquad (10)$$

where we used $\mu/s << 1$. The main difference to the case with gene conversion is that the factor $(\mu + \gamma)/s$ is missing in the exponent of (10). As a consequence, a wide range of parameters lead to a fast ratchet, $n_0 s_0 < 1$, even if the relevant mutations are non-neutral, $N s >> 1$. All that is required is that $Lp_{AB} = 2\mu L/(hs) >> 1$, i.e. that the total genomic mutation rate $L\mu$ is significantly larger than the selection coefficient s.

To summarise, the mechanism of Muller's ratchet relies on the stochastic extinction of a subpopulation of genotypes. Without gene conversion, Muller's ratchet clicks as soon as the class of (currently) fittest genotypes goes extinct. With gene conversion, it is the population of "restorable" genotypes that has to go extinct for Muller's ratchet to click. Because this population is often much larger than the class of only the fittest genotypes, stochastic extinction is less likely. As a consequence, Muller's ratchet clicks slower in the presence of gene conversion unless the relevant deleterious mutations are effectively neutral. These results are consistent with the stochastic simulations of Connallon and Clark[62].

Remarks

• The factor of $\mu + \gamma$ of mutation and gene conversion rates per locus in Eq. (8) is proportional to the size of the loci in number of base pairs. The typical size of gene conversion tracks sets the effective size of the loci considered in our model. Hence, the smaller the typical size of gene conversion tracks the smaller the factor $\mu + \gamma$ and the more efficient is the beneficial effect of gene conversion on the speed of the ratchet. This can be understood intuitively by considering two neighbouring loci 1 and 2. The state $A_1B_1 - B_2A_2$ belongs to the restorable class. If the gene conversion tracks were however twice as long then this state could not be restored, because gene conversion leads to either $B_1B_1 - A_2A_2$ or $A_1A_1 - B_2B_2$, which both carry deleterious alleles.

• One might argue that one click in the gene conversion case leads to a (potentially) stronger fitness detriment s rather than hs in the no-gene-conversion case. Hence, one should compare 1/h clicks in the no-gene conversion case with the gene conversion case. This however will not change the strong discrepancy of time scales, due to the exponential dependence of the click time on the size of the restorable class.

• The population fraction $P_{r.c.}$ of the restorable class is given by $P_{r.c.} \approx P_{m.f.} \exp((\mu+\gamma)/s)$ where $P_{m.f.}$ is the population fraction of the fittest genotypes. From this relation, it seems that the smaller the gene conversion rate the larger the restorable class. On the other hand, it is clear that one needs at least some gene conversion to decelerate Muller's ratchet. The analysis of the threshold gene conversion is left to a more detailed analysis of Muller's ratchet with gene conversion that will be published elsewhere (this analysis may also deal with the case of a fast clicking ratchet, $n_0 s_0 << 1$).

E3. Analysis of gene divergence

Because selection can act very locally within proteins, leading to a diluted signal at the entire gene scale, we conducted sliding window analyses of Ka/Ks along alignments of allelic gene pairs. Sliding window analyses were performed on 104 allelic pairs where the two coding sequences were of equal length (to avoid ambiguity in alignment) and Ka > 0.03 (to provide sufficient signal for the analysis), using DNAsp v5 (http://www.ub.edu/dnasp/) with a window length of 50 and step size of 10 (longer window lengths were used in some pairwise comparisons when a length of 50 encompassed regions with nonsynonymous differences but no synonymous differences). Because of the small amount of data in the window, Ka and Ks were calculated using the method of Nei and Gojobori[67]. In Supplementary Data 6 a pair with at least one region where Ka/Ks > 1.2 is indicated as "Positive" and a pair with no region

where Ka/Ks > 1.2 but with a region with Ka/Ks between 0.9 and 1.2 is indicated by "Neutral". Pairs with Ka/Ks < 0.9 along their entire length are indicated as "Purifying". Among the 104 alignments, 48 showed localised regions of Ka/Ks > 1 and 4 had Ks of 0 (Supplementary Fig. 12).
We also examined Ka and Ks in 5360 allelic pairs that had coding sequences of equal length and lacked ambiguous bases. More than three quarters of pairs had a Ka/Ks < 0.3, indicating strong purifying selection, and 85 pairs had Ks > 0.03 but no nonsynonymous differences. A total of 87 pairs had a Ka/Ks > 1, potentially indicating positive selection related to neo- or sub-functionalisation, but Fisher tests yielded significant P values < 0.05 for only two of these 87 pairs. In a detailed analysis of quartets, ~36% displayed a much greater difference in Ka than Ks between ohnologues (Supplementary Fig. 13) and ~8% displayed evidence of asymmetrical evolution within one of their allelic pairs (Supplementary Fig. 14).

E4. Detection of horizontal transfers

Alien Index (AI) analyses of the gene set were performed as described previously[16] (Supplementary Data 5). The AI is defined as log((Best E-value for metazoans) + e-200) - log((Best E-value for non-metazoans) + e-200) and can take values between -460 and +460. This approach only identifies putative horizontal gene transfers (HGTs) from non-metazoans (e.g. bacteria, plants or fungi). We used the conservative thresholds of -45 and +45 in our downstream analyses, i.e. genes with AI≤-45 were considered of probable metazoan origin, genes with -45<AI≤45 of uncertain origin, and genes with 45<AI as probable non-metazoan origin.

E5. Co-occurrence of TEs and horizontal gene transfers (HGTs)

It had previously been suggested that both TEs and HGTs accumulate in the subtelomeric regions of *A. vaga*. A consequence of this should be that TEs and HGTs co-occur in the genome, a prediction that we tested statistically using the annotated TEs and AI>45 genes.

First, we counted the number of TEs in windows of 500 bp, 1,000 bp, 2,000bp and 5,000 bp, around HGTs and around the rest of protein-coding genes. Using a t-test we compared the distribution of TEs density around HGTs genes and the rest of protein-coding genes for the four genomic window sizes.

We found that the density of TEs was significantly higher around HGT genes for all genomic windows of size ≥ 2000 bp (significant p-values are presented in bold).

	Size windows	500 bp	1,000 bp	2,000 bp	5,000 bp
Non-HGT genes	Number	491	668	943	1802
	Mean	$1.10\ 10^{-2}$	$1.49\ 10^{-2}$	$2.11\ 10^{-2}$	$4.04\ 10^{-2}$
	Standard deviation	$1.84\ 10^{-1}$	$2.21\ 10^{-1}$	$2.67\ 10^{-1}$	$3.82\ 10^{-1}$
HGT genes	Number	44	68	145	357
	Mean	$9.24\ 10^{-3}$	$1.42\ 10^{-2}$	$3.04\ 10^{-2}$	$7.50\ 10^{-2}$
	Standard deviation	$1.46\ 10^{-1}$	$2.05\ 10^{-1}$	$3.14\ 10^{-1}$	$5.21\ 10^{-1}$
	T-test pvalue	0.43	0.82	**0.049**	**$8.91\ 10^{-6}$**

Second, we counted the number of HT genes and other protein-coding genes in windows of 500 bp, 1,000 bp, 2,000 bp and 5,000 bp, around TEs. Using Fisher's exact test we compared the distribution of TE density around HT genes and the other protein-coding genes for the four genomic window sizes.

We found that the density of HT genes was significantly higher around TEs for windows of size ≥ 2000 bp (significant p-values are presented in bold).

	Size windows	500 bp	1,000 bp	2,000 bp	5,000 bp
No HT genes	Number	44	68	145	357
	Frequency	$0,9\ 10^{-2}$	$1,4\ 10^{-2}$	$3\ 10^{-2}$	$7.5\ 10^{-2}$
HT genes	Number	491	668	943	1802
	Frequency	$1,1\ 10^{-2}$	$1,4\ 10^{-2}$	$1.9\ 10^{-2}$	$4\ 10^{-2}$
	pvalue	$3\ 10^{-1}$	$6.1\ 10^{-1}$	**$1.01\ 10^{-4}$**	**$2.77\ 10^{-22}$**

E6. Analysis of gene losses

A search for completely sequenced genomes available among metazoans led to the selection of genome sequences from 50 bilaterian and 5 non-bilaterian species. The 50 bilaterian species comprised 21 deuterostomes (*Homo sapiens*, *Pan troglodytes*, *Pongo abelii*, *Macaca mulatta*, *Canis lupus familiaris*, *Mus musculus*, *Rattus norvegicus*, *Equus caballus*, *Bos taurus*, *Monodelphis domestica*, *Ornithorhynchus anatinus*, *Gallus gallus*, *Meleagris gallopavo*, *Taeniopygia guttata*, *Xenopus tropicalis*, *Tetraodon nigroviridis*, *Oryzias latipes*, *Danio rerio*, *Branchiostoma floridae*, *Ciona intestinalis*, *Strongylocentrotus purpuratus*); 12 arthropods (*Camponotus floridanus*, *Atta cephalotes*, *Solenopsis invicta*, *Daphnia pulex*, *Acyrthosiphon pisum*, *Tribolium castaneum*, *Bombyx mori*, *Aedes aegypti*, *Anopheles gambiae*, *Drosophila melanogaster*, *Nasonia vitripennis*, *Apis mellifera*); 12 nematodes (*Caenorhabditis elegans, Caenorhabditis brenneri*, *Caenorhabditis japonica*, *Caenorhabditis briggsae* AF16, *Caenorhabditis angaria*, *Caenorhabditis remanei*, *Pristionchus pacificus*, *Brugia malayi*, *Meloidogyne hapla*, *Meloidogyne incognita acrita*, *Haemonchus contortus*, *Trichinella spiralis*); the platyhelminthes *Schistosoma mansoni*; 2 annelids (*Helobdella robusta* and *Capitella teleta*); the mollusc *Lottia gigantea*, and of course, our newly sequenced rotifer *A. vaga*. The 5 non-bilaterian species comprised 2 cnidarians (*Nematostella vectensis* and *Hydra magnipapillata*), the poriferan *Amphimedon queenslandica*, the placozoan *Trichoplax adhaerens*, and the choanoflagellate *Monosiga brevicollis* (which is not a metazoan species, but an opistokont protozoan, closely related to metazoans).

The complete protein set of several of these species was downloaded from the GenPept section of the National Center of Biotechnology Information (NCBI): *Amphimedon queenslandica* (whole genome shotgun sequencing project ACUQ00000000, 9814 predicted proteins), *Brugia malayi* (whole genome shotgun sequencing project AAQA00000000, 11472 predicted proteins), *Camponotus floridanus* (whole genome shotgun sequencing project AEAB00000000, 14864 predicted proteins), *Nasonia vitripennis* (whole genome shotgun sequencing project AAZX00000000, 12988 predicted proteins), and *Solenopsis invicta* (whole genome shotgun sequencing project AEAQ00000000, 14180 predicted proteins). The protein sets of the nematodes *Caenorhabditis angaria* and *Haemonchus contortus* were downloaded from WormBase (ftp://ftp.wormbase.org). The complete proteomes of *Meloidogyne hapla* and *Meloidogyne incognita* were downloaded from the Plant Nematode Genomics Group website (http://www.pngg.org/cbnp/index.php) and from the INRA website (http://www.inra.fr/meloidogyne_incognita), respectively. The *Schistosoma mansoni* protein set was downloaded from the Wellcome Trust Sanger Institute (ftp://ftp.sanger.ac.uk/pub/pathogens/Schistosoma/mansoni/genome/gene_predictions/). Finally, the complete proteome of several species were downloaded from the DOE Joint Genome Institute (JGI) website: *Branchiostoma floridae* (http://genome.jgi-psf.org/Brafl1/Brafl1.download.ftp.html), *Capitella teleta* (http://genome.jgi-psf.org/Capca1/Capca1.download.ftp.html), *Helobdella robusta* (http://genome.jgi-psf.org/Helro1/Helro1.download.ftp.html), *Hydra magnipapillata* (ftp://ftp.jgi-psf.org/pub/JGI_data/Hydra_magnipapillata/hydra_Hma2.pep.fa.gz), *Lottia gigantea* (http://genome.jgi-psf.org/Lotgi1/Lotgi1.home.html), and *Monosiga brevicollis* (http://

genome.jgi.doe.gov/Monbr1/Monbr1.download.ftp.html). The complete proteomes of all other species were downloaded from the Ensembl Database (http://www.ensembl.org/), or from the Metazoan section from the Ensembl Database (http://metazoa.ensembl.org/info/data/ftp/index.html).

To determine the genes that would have been specifically lost and gained in the *A. vaga* genome during the course of evolution, we focused on orthologues from the complete proteomes of these 55 metazoan species. Orthologous associations were obtained by clustering the protein sequences using OrthoMCL[68], where the smallest OrthoMCL clusters are composed of at least two putative orthologous proteins. The OrthoMCL clusters were examined to determine the repertoire of ancestral protein-coding genes present in the genome of the last common ancestor of Protostomia. We then carried out a study of gene gain and loss in the *A. vaga* genome using this ur-protostome geneset and the programs PARS[69] and PhyloPattern[70]. The PARS algorithm infers gains and losses along a phylogenetic tree assuming that because gene transfer is rare in metazoan species, a gene gain occurs only once, and that a gain must be preceded by a loss. PhyloPattern then inventories the different patterns of gene gain and loss along the phylogenetic species tree. Finally, we counted the total number of orthologous groups that may have been present at each node of our species tree, as well as the number of orthologous clusters specifically gained and specifically lost at each node.

The reconstruction of ur-protostome geneset suggests that 7844 OrthoMCL groups of putative orthologous proteins were present in the last common ancestor of Protostomia. Our analysis of the *A. vaga* geneset revealed that 1257 OrthoMCL orthologous clusters and thus at least 1257 protein-coding genes, have been specifically lost in the lineage leading to the rotifer *A. vaga*.

The OrthoMCL groups and and their phyletic distribution are available online through the IODA browser[71] (http://ioda.univ-provence.fr, link on the left of the page).

F. Analyses of specific gene families

F1. Meiosis genes

We created a data set of genes involved in meiosis, homologous recombination, DNA repair, and mitotic and meiotic cohesion using the NCBI Homologene resource and collected searches of the NCBI refseq database (see Supplementary Data 7). This dataset was used in reciprocal best BLAST searches with the 49300 annotated *A. vaga* gene set (using blastp) and the 44635 contigs of the assembly (using blastx and tblastn) to identify potential *A. vaga* homologs. Each potential homolog was then compared to the NCBI refseq database for final confirmation of orthology. The only case in which a gene was not found in the gene set but was found in the contigs was SPO11; for REC114 one gene copy was found in the gene set and the other was found in a contig. Nearly all genes were found in 2-4 copies. Among genes for which we did not identify a homolog in *A. vaga*, most were vertebrate- or yeast-specific. Only four genes were found in transcriptomic and partial genome coverage of monogonont rotifers but not in the genome of *A. vaga* (data not shown) and may therefore be considered bdelloid-specific losses: HOP2, MND1, NBS1, and REC8.

F2. Gene families expansion

We examined the distribution of PFAM domain abundances in *A. vaga* and in the average of 8 other metazoan species (Supplementary Fig. 4). The slope of the linear regression equation was 1.88, indicating that the *A. vaga* genome contained almost twice as many domains as the average for a metazoan genome (since most haplotypes were assembled separately and retained in the annotation process). Some domains are clearly more than two times more abundant in *A. vaga* than in an average metazoan species. Excluding *A. vaga*-specific domains (that were probably acquired via HGT), there are 1,581 PFAM domains that

are at least 3 times more abundant in *A. vaga* than in the average metazoan. Of note is the tetratricopeptide repeat domain (TPR), with 14 different families. Overall, the genome of *A. vaga* comprises 5,647 TPR domains compared to an average of 703 for the 8 other metazoan species analysed. We also observe that domains related to oxidoreduction activities, Glycosyl Hydrolase domains (23 in total), and proteins linked to chitin degradation or binding are frequent in the list of 1,581 over-represented PFAM domains (see sections below).Overall, there are 57 matches to GH19 chitinase domains and chitin-binding domains in *A. vaga*, a number 12 times higher than the average for a metazoan species (4.75).

F3. Carbohydrate active enzymes (CAZymes)

CAZymes form the major catalytic machinery for the assembly and breakdown of complex carbohydrates that play essential roles as structural components and carbon source, but also in intra- and intercellular recognition events. CAZymes are characterised by specific catalytic modules, organised in the following classes: GH for Glycoside Hydrolases, GT for Glycosyl Transferases, PL for Polysaccharide Lyases, CE for Carbohydrate Esterases. These latter catalytic modules are occasionally appended to carbohydrate-binding modules (CBM).

To identify candidate CAZymes, we first compared the protein sequences of *A. vaga* to the full length sequences of the Carbohydrate-Active enZymes (CAZy) database[72] (http://www.cazy.org) using BLAST[73] with an e-value cutoff of 0.1. We subjected these hits in parallel to (i) a BLAST search against a library built with partial sequences corresponding to individual GH, GT, PL, CE and CBM modules and (ii) a HMMer search[74] using hidden Markov models custom built for each CAZy module family. A sequence was considered reliably assigned when it was placed in the same family with the two methods. Borderline cases were manually inspected for the presence of conserved features such as catalytic residues. When necessary a functional annotation was performed by pairwise comparison of *A. vaga* putative CAZymes against a library containing solely the sequences of experimentally characterised CAZymes, derived from the CAZy database. The abundance and distribution (in classes and families) of the whole set of CAZymes encoded by *A. vaga* was compared to those of six other metazoan species, *Danaus plexippus, Caenorhabditis elegans, Drosophila melanogaster, Meloidogyne incognita, Oikopleura dioica* and *Homo sapiens.*

CAZymes typically correspond to 1-5% of the gene products of a living organism; in *A. vaga* 1075 CAZymes were detected constituting around 2% of its total geneset. More specifically, 623 GHs, 412 GTs, 13 PLs, 27 CEs were detected, along with 154 CBMs always associated to GHs. Compared to the other metazoans, *A. vaga* features by far the most diverse and abundant CAZyme repertoire, especially for the classes GH, GT, PL and CE (Supplementary Fig. 15). In total, 37 CAZyme families (totalling 202 genes) appear as unique to *A. vaga*. Interestingly, a high number of those CAZymes may have been acquired via horizontal gene transfer (Supplementary Table 6), especially for the enzymes involved in the breakdown of complex carbohydrates (GHs, PLs and CEs). The highest AI values are for the classes PLs that are typically not found in metazoans and the CEs. Finally, CAZymes involved in the degradation of chitin (GH18, GH19 and GH20) are also overabundant in *A. vaga* (Supplementary Table 7).

F4. Antioxidant enzymes

Starting from the PFAM annotation, we observed a total of 82 PFAM domains related to oxido-reduction processes. Overall, these 82 PFAM domains represent a total of 2,397 hits in *A. vaga*, almost 5 times the average for a metazoan species (484 hits). Some of these domains are 16 times (NADPH-dependent FMN reductase) to 24 times (e.g. Dyp-type peroxidase family) more abundant in *A. vaga* than in an average metazoan.

An automated annotation was conducted on a selected list of primary antioxidant genes by using blast2GO at NCBI and by blasting reference candidate genes against the *A. vaga* geneset. A total of 278 antioxidant candidate genes were retrieved for *A. vaga*, some of

which apparently acquired through HGT (see Supplementary Table 8). In comparison to *C. elegans* (Supplementary Table 8) and six other metazoans (Supplementary Table 9), *A. vaga* contains more genes within each category of selected antioxidants except for glutathione peroxidase (see also Supplementary Fig. S16 and S17). Interestingly, 213 of the 278 candidate genes were found in the cDNA library of hydrated *A. vaga* individuals (control conditions), indicating constitutive expression of the antioxidant genes.

F5. Homeobox genes

We used the HMMER software package[75] and the homeobox model from PFAM[76] to identify homeodomains in the *Adineta vaga* predicted proteome and translated genome, recovering 307 homeodomains from 293 protein models. We used the alignment generated by hmmsearch, which included human and *Drosophila* homeodomains, as input to a maximum likelihood analysis. We used the Perl script ProteinModelSelection.pl (available from the RAxML[77] web site) to determine that the LG model of evolution was the most appropriate for this superclass tree alignment. We ran RAxML version 7.28[77] on this superclass alignment to generate a maximum likelihood tree with 100 bootstraps, and used the tree (Supplementary Data 8) to divide the *A. vaga* homeodomains into classes. We generated individual alignments for the POU, LIM, and PRD classes, as well as the HOXL and NKL subclasses of the ANTP class, and finally a combined alignment of TALE and SINE classes. We used the HomeoDB[78] to add human and *Drosophila* homeodomains from each class. We used the ProteinModelSelection.pl script to determine the most appropriate model for each alignment. We found that the best model was LG for all classes except the POU class for which RTREV was best. We generated maximum likelihood trees with 100 bootstraps for each class.

We compared the family counts provided for ten species in HomeoDB[78] (downloaded in August 2012) to our manual counts of *A. vaga* family-specific clades from our HOXL, NKL, POU, LIM, and PRD trees. Each clade of *Adineta vaga* homeoboxes was counted as having one, two, three, four, or more than four homeoboxes. Families with multiple homeoboxes (i.e. Dux, Rhox, Hdx) were not considered and pseudogenes were not counted.

We found that *A. vaga* includes many more homeobox gene families with four copies than are seen in other animals, including vertebrates that have undergone two rounds of genome duplication and even the zebrafish that has undergone an additional round (Supplementary Fig. 21). This inordinate maintenance of four gene copies likely points to constraints on gene balance during embryogenesis counterselecting the independent loss of gene copies is counterselected.

F6. TE defense genes

Gene datasets

To search for RNA-mediated silencing proteins in the *A. vaga* genome, we performed blastp and tblastn searches with the available gene models and scaffolds. The reference datasets were assembled using published protein sequences for Piwi-Argonaute[79], for RNA-dependent RNA polymerases[80] and for Dicers[81]. The initial homology output was then verified by reciprocal blast (tblastn) against the NCBI nr database. Finally, the selected protein ID sequences confirmed by reciprocal blast were compared with the *A. vaga* genomic sequence to perform manual adjustment of boundaries in a few cases of incomplete protein annotation, and phylogenetic analysis of each dataset was performed to assess their evolutionary relationships.

Phylogenetic analysis

Multiple sequence alignments were performed using Muscle v3.6 with default settings[82]. Phylogenetic trees were generated under the maximum likelihood criterion using PhyML 3.0 (LG model, NNI topological moves, optimizing branch lengths and likelihood branch supports)[83]. All manipulations of phylogenetic trees were performed using Figtree[84].

While the *A. vaga* genome contains numerous proteins with homology to different components of the RNA-mediated silencing machinery, we limited our phylogenetic analysis to its major components, such as RdRP, Piwi/Argonaute, and Dicer, for which published reference datasets and phylogenetic trees were available[79-81]. For RdRPs, previous studies revealed three major clades with relatively strong support, designated RDRα, RDRβ and RDRγ[80]. While most of the *A. vaga* inferred RDR homologs are assigned to the RDRα clade, one of them can be assigned to the mostly fungal RDRβ clade with good support (Supplementary Fig. 18). For Piwi/Ago homologs, a clear dichotomy between the Piwi and Argonaute subfamilies can be observed (Supplementary Fig. 19). A diverse group of *C. elegans* Argonaute proteins represents worm-specific Argonautes (WAGOs)[85] but the *A. vaga* Argonautes are not grouped with WAGOs, indicating that their diversification has occurred independently. Bilaterian Dicer genes form two distinct groups: Dicer1 genes present as single copy in most organisms, and insect Dicer2 genes which arose via a lineage-specific duplication event[81]. The *A. vaga* Dicer homologs also show an ancestral duplication event, which is distinct from the insect-specific duplication (Supplementary Fig. 20). Such highly diversified TE silencing machinery can be expected to slow down TE proliferation within the genome, thereby increasing further the chances for their inactivation via deletion and point mutation.

Supplementary Figures

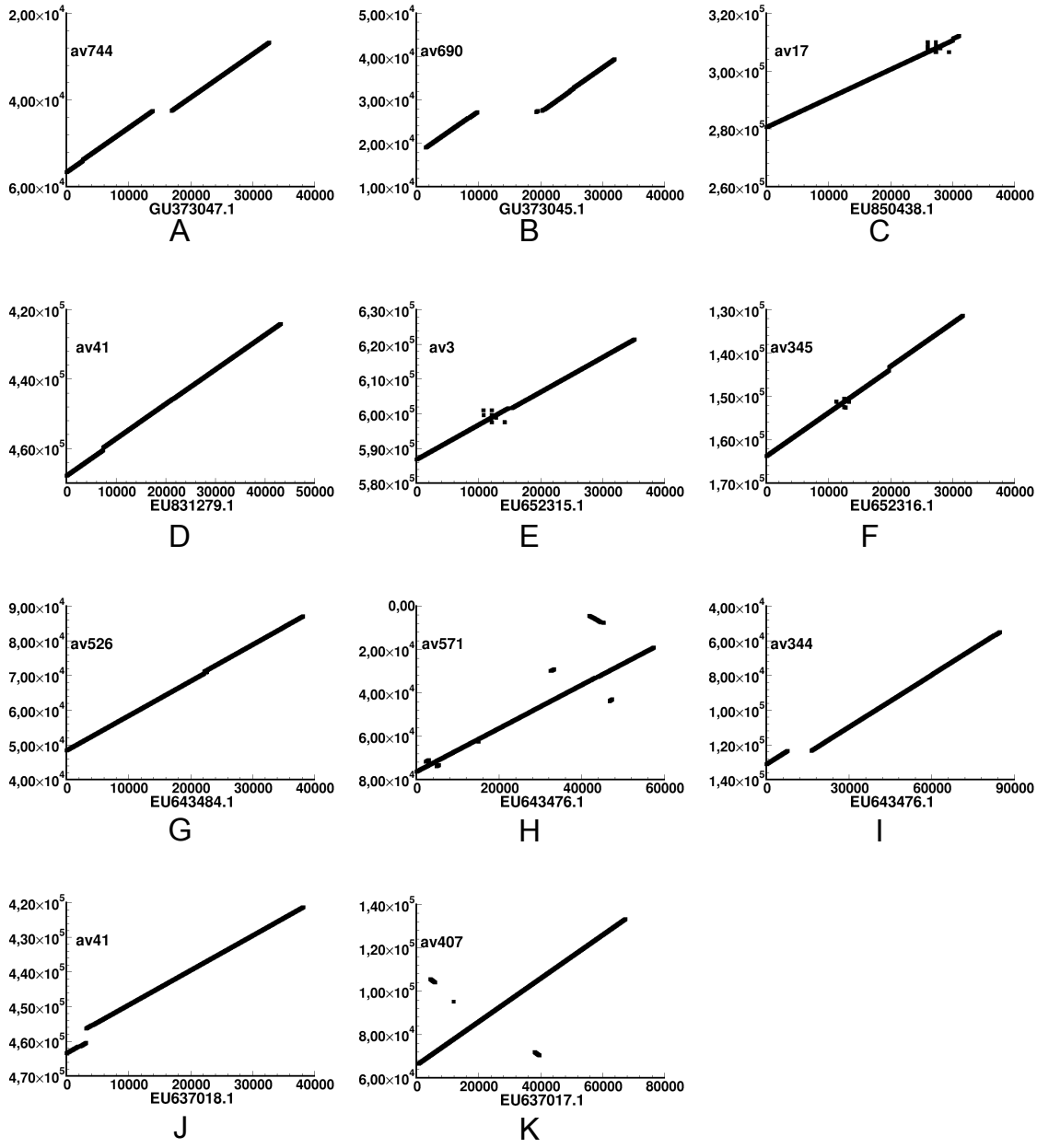


Fig. 1. Validation of the genome assembly by comparison with eleven published sequences. Each fosmid aligned along a single scaffold, notably without inversion. However, some regions were missing from two scaffolds compared with the reference fosmids (see A and B). Levels of similarity (computed for local BLAST alignments longer than 500 bp) are 97.7% for A, 98.25% for B, 98.4% for C, 99.7% for D, 98.4% for E, 98.8% for F, 97.5% for G, 98.3% for H, 96.8% for I, 99.4% for J and 99.9% for K.

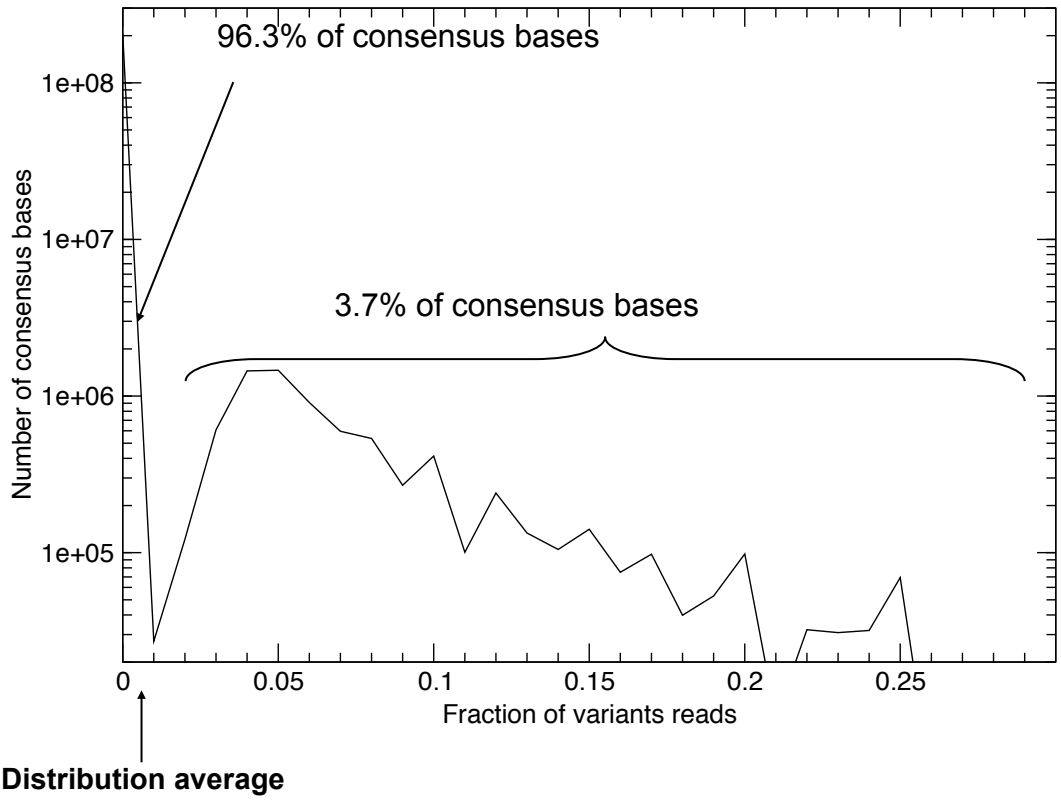


Fig. 2. Distribution of the fraction of variants reads at each position in the assembly.
Most consensus bases (96.3%) had less variants than expected from the 1% average error rate of 454 Titanium, whereas 3.7% of the bases had more variants than expected. Overall the average proportion of variants per consensus base was 0.31%, about one third of the reported error rate. (The drop in the distribution between 1% and 4% is an artefact due to the ~19X average coverage by 454 reads in the assembly: only genomic positions with more than twice this average coverage can yield a percentage of variant reads comprised between 1 and 2%).

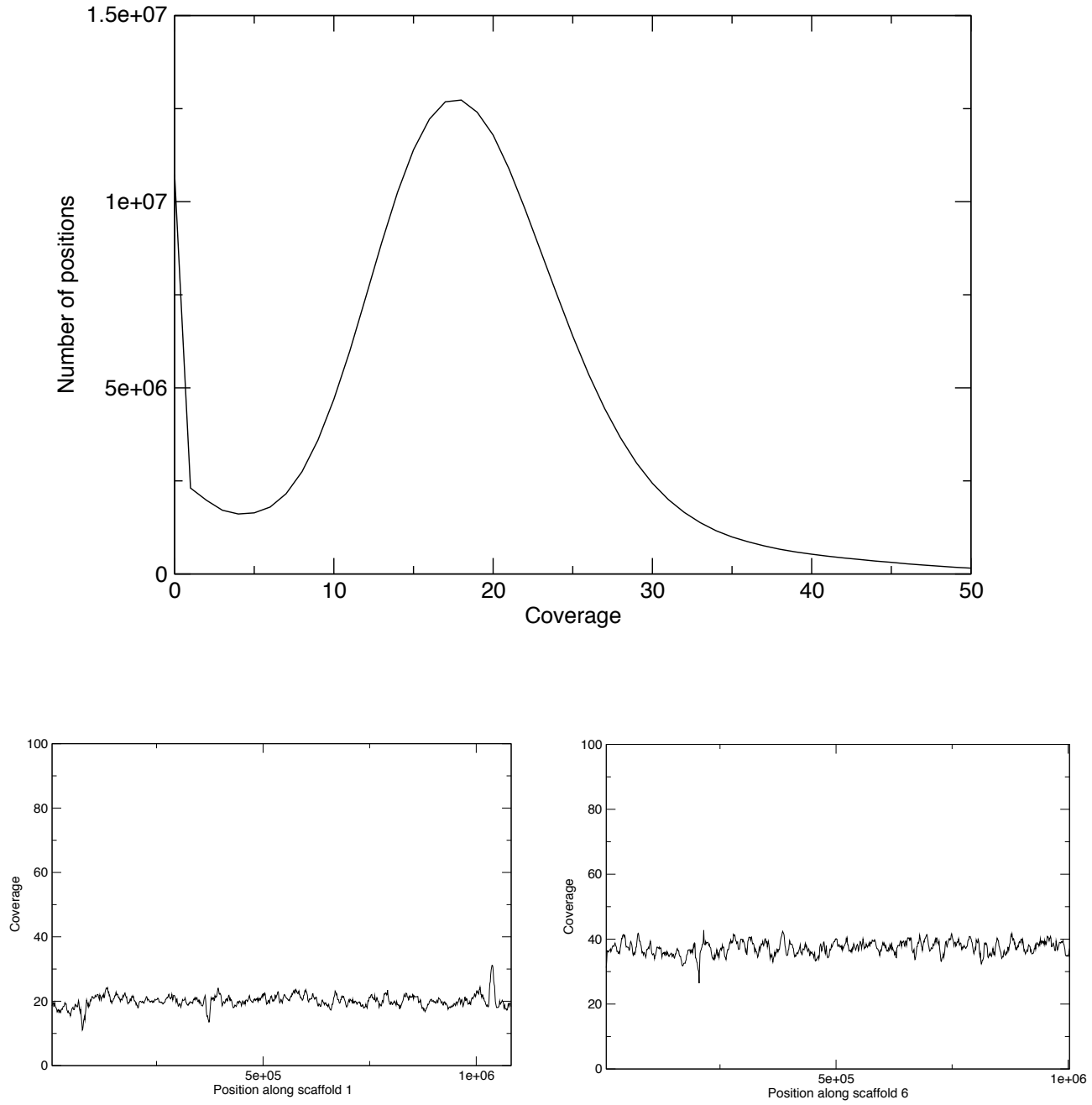


Fig. 3. Distribution of 454 sequencing coverage across the genome and along two representative scaffolds.
Although the distribution of sequencing coverage across the genome appears unimodal (top), detailed sliding window analyses (with a window size of 1000 bp and a step of 100 bp) show that some of the scaffolds such as av6 have double coverage across their whole length and likely result from the fusion of two identical or near-identical copies during the assembly.

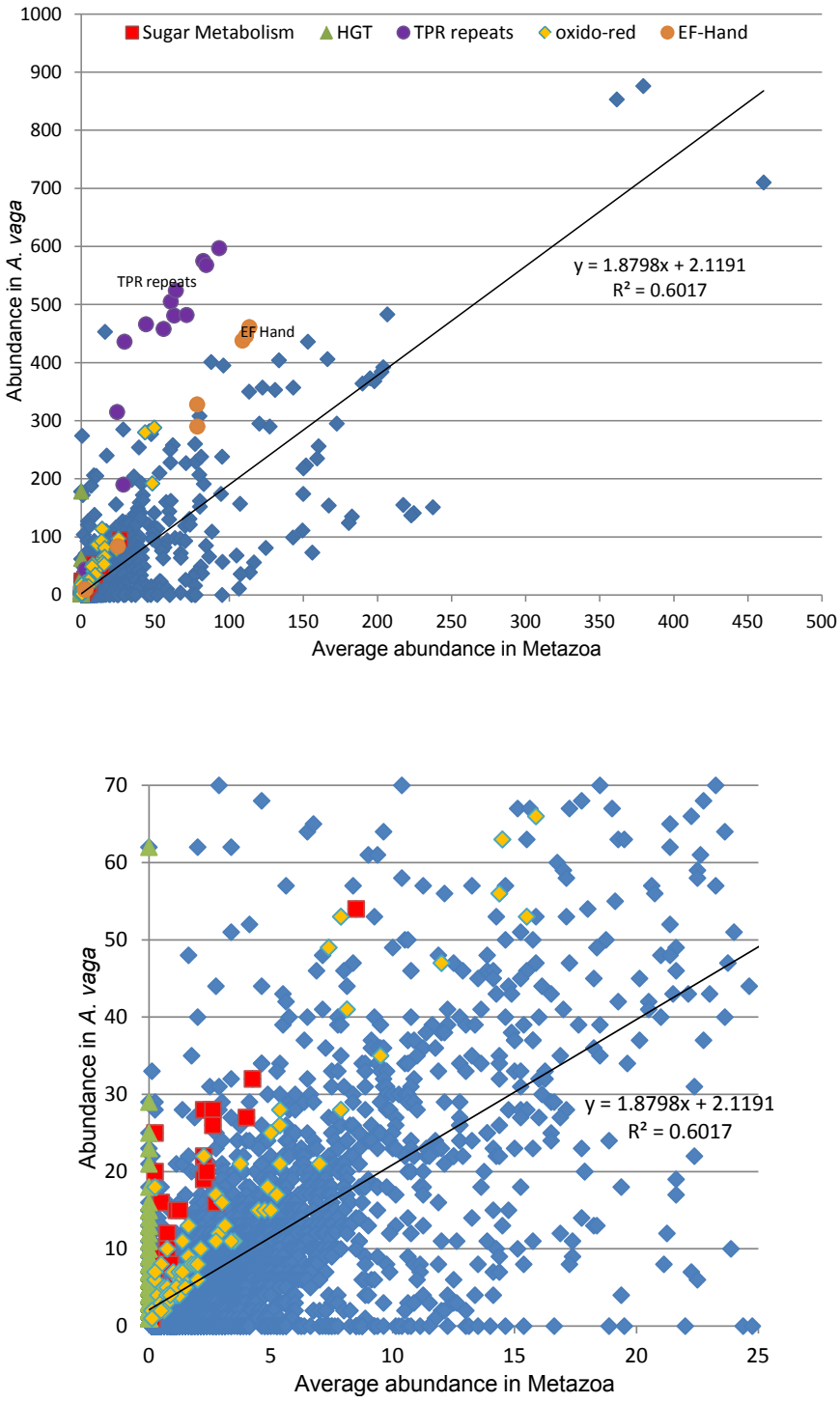


Fig. 4. Comparison of the distribution of abundance of PFAM domains in *A. vaga* to the average abundance in metazoans.
Above: abundance of the different PFAM domains in *A. vaga* vs the average abundance measured in 8 other metazoan species. PFAM domains present along the y axis represent *A. vaga*-specific domains absent from the 8 other metazoan genomes. Proteins probably acquired via HGT are represented with green diamonds. Among the over-abundant domains, we indicate Tetratricopeptide repeats (TPR) in purple, CAZyme domains (red), domains involved in oxido-reductive functions (yellow) and EF-Hand domains (orange). Under: close-up on the bottom-left part of the previous graph.

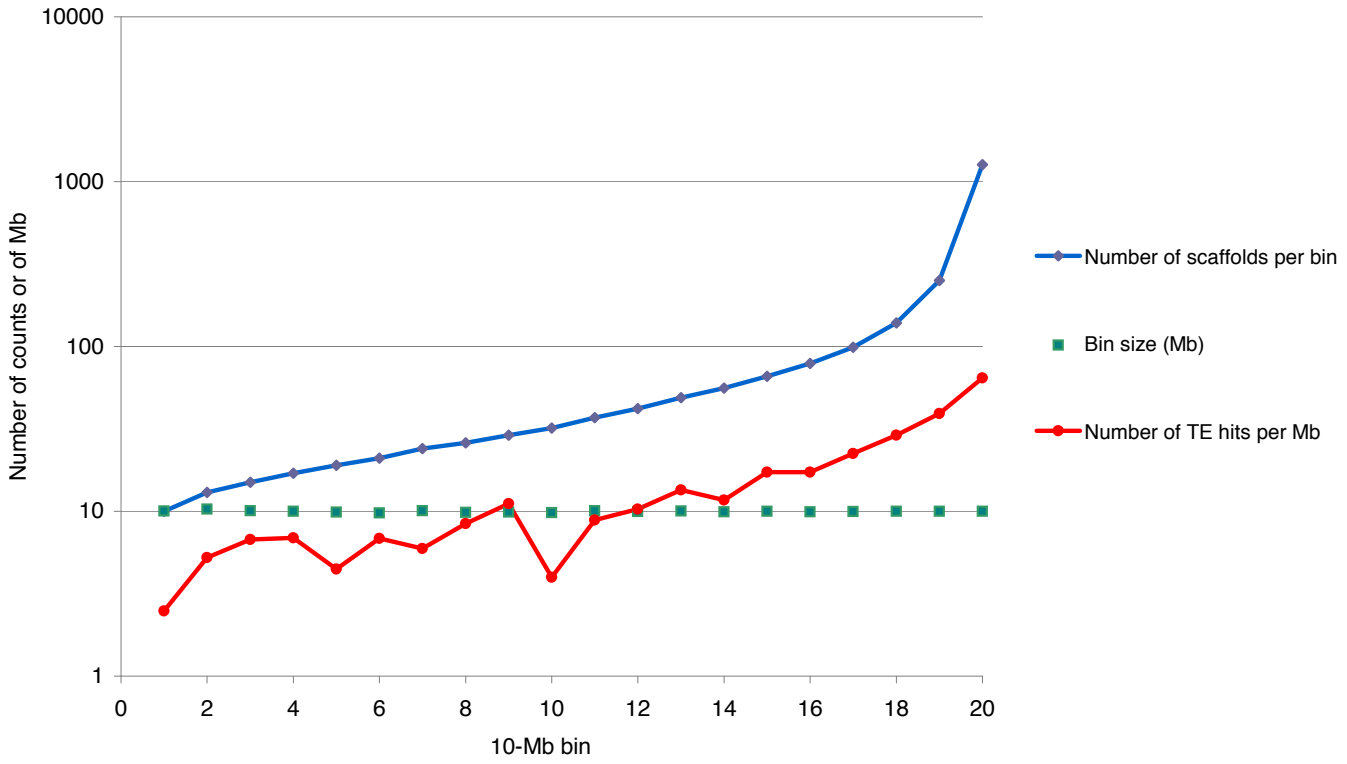


Fig. 5. Distribution of TE insertions among *A. vaga* scaffolds.
The *A. vaga* genomic scaffolds were consecutively grouped into 20 *ca*. 10-Mb bins that were plotted on the X-axis, beginning with scaffold av1. The Y axis (log scale) shows that the number of TE hits per Mb in each bin (red) increases with the number of scaffolds per bin (blue) (i.e. the density of TEs is generally higher in small scaffolds), whereas the number of Mb in each bin (green) remains constant.

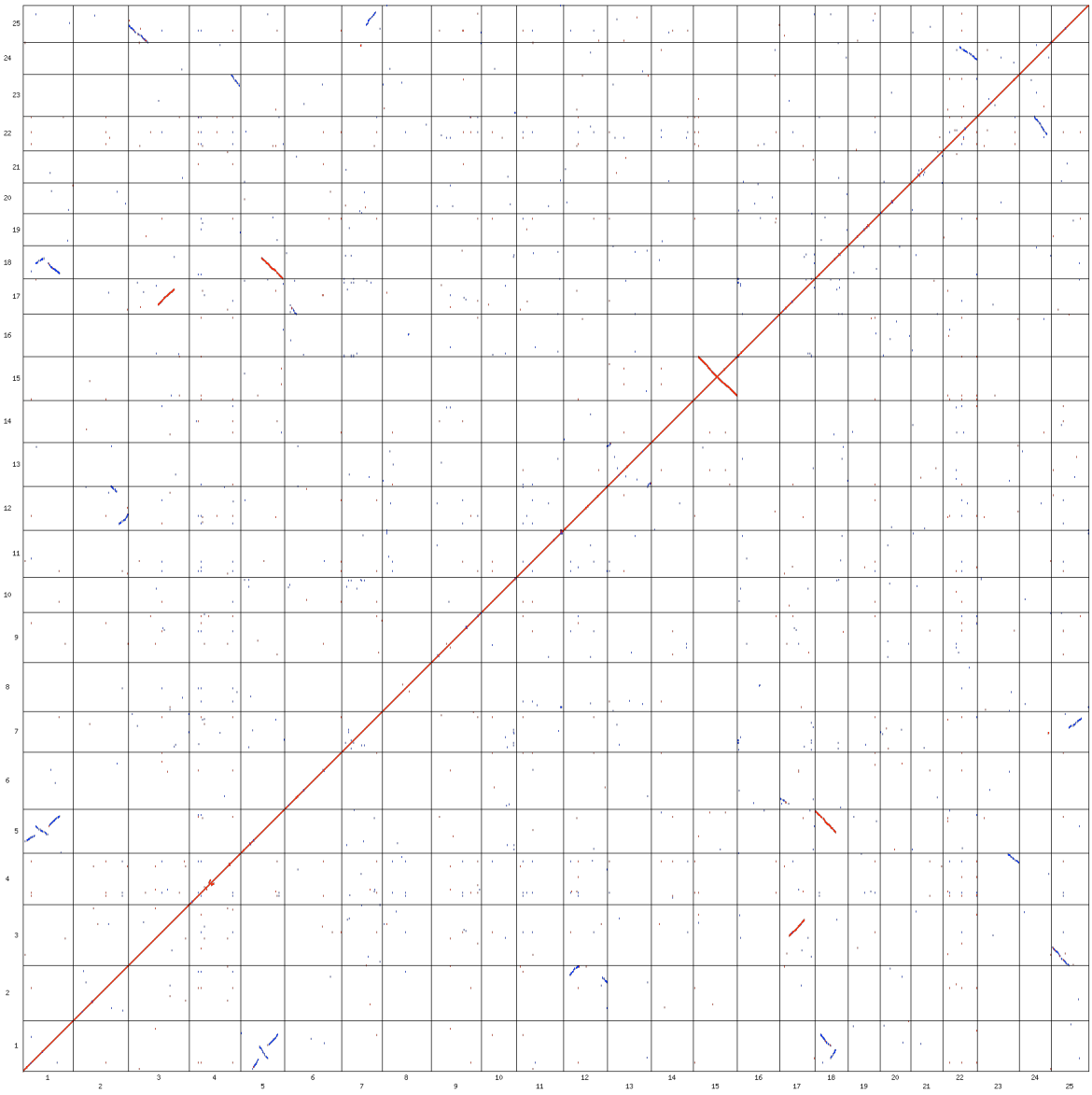
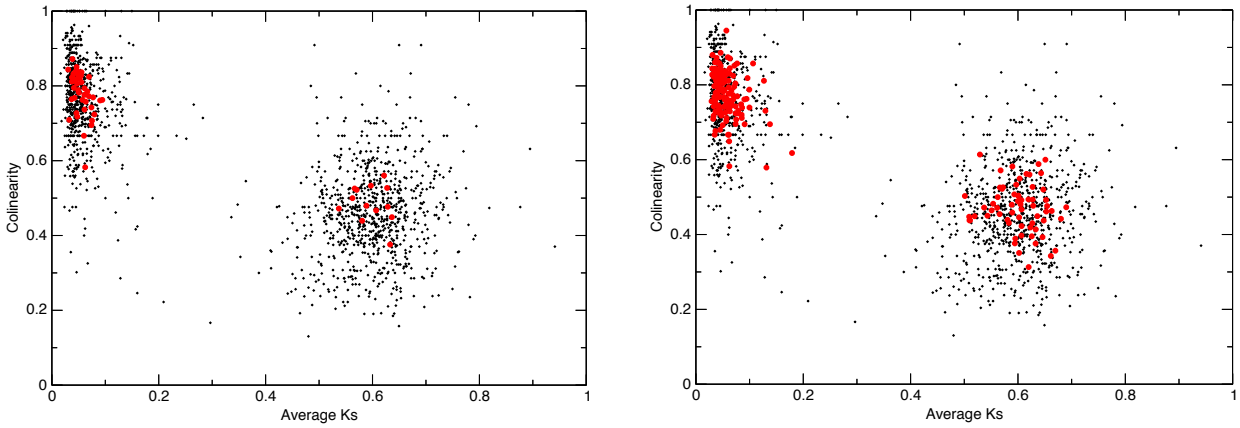


Fig. 6. Oxford grid of synteny conservation between the 25 largest scaffolds.
In red: gene pairs that have at least 95% identity at the nucleotide sequence level; in blue: gene pairs that have less than 95% identity at the nucleotide sequence level

A



B

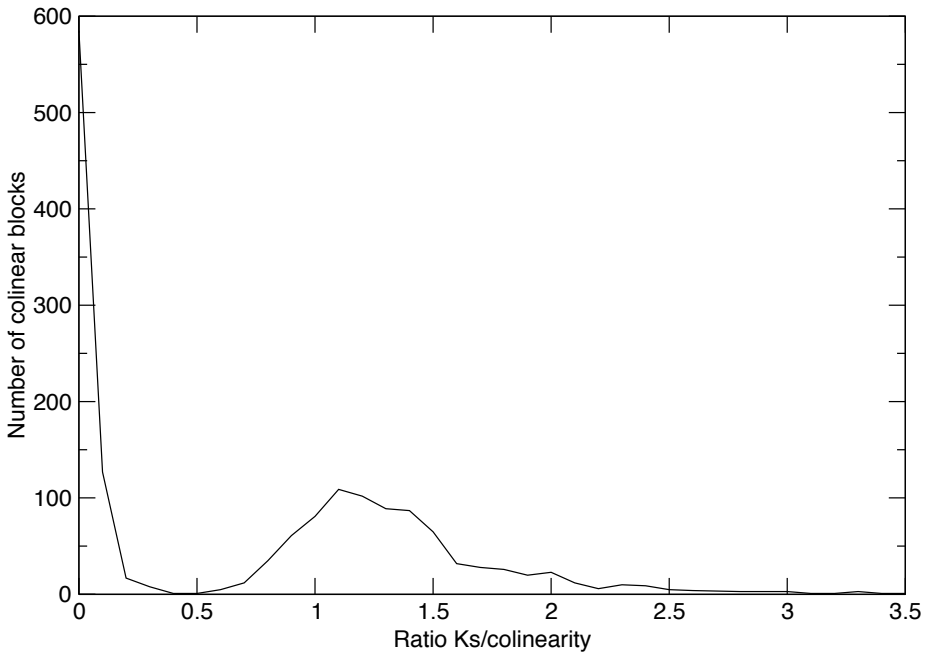


Fig. 7. Distribution of the average Ks and colinearity among colinear blocks.
a, The variance in the average Ks and colinearity metrics of colinear blocks is strongly dependent on the size of the blocks: on the left, blocks comprising more than 50 colinear genes are highlighted in red; on the right, the threshold is set to 30 genes.
b, The distribution of the ratio (average Ks)/colinearity is bimodal with a clear gap between allelic blocks (ratio below 0.5) and ohnologous blocks (ratio above 0.5).

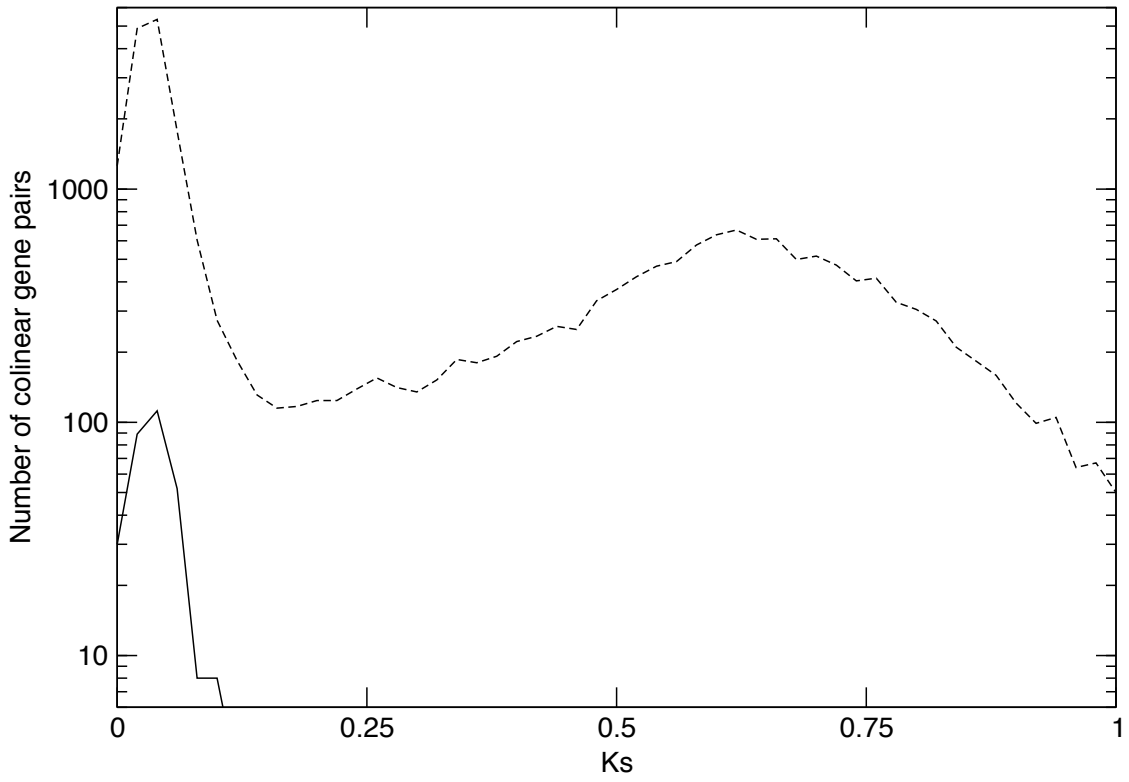


Fig. 8. Distribution of Ks for the genes in the arms of the palindromes (lower curve) vs the complete geneset (upper curve).
The Ks signature of the gene pairs in the palindromes is identical to the one of alleles, suggesting that palindromes did not arise by recent reduplication of genome regions but rather by shuffling of alleles caused by genome rearrangements.

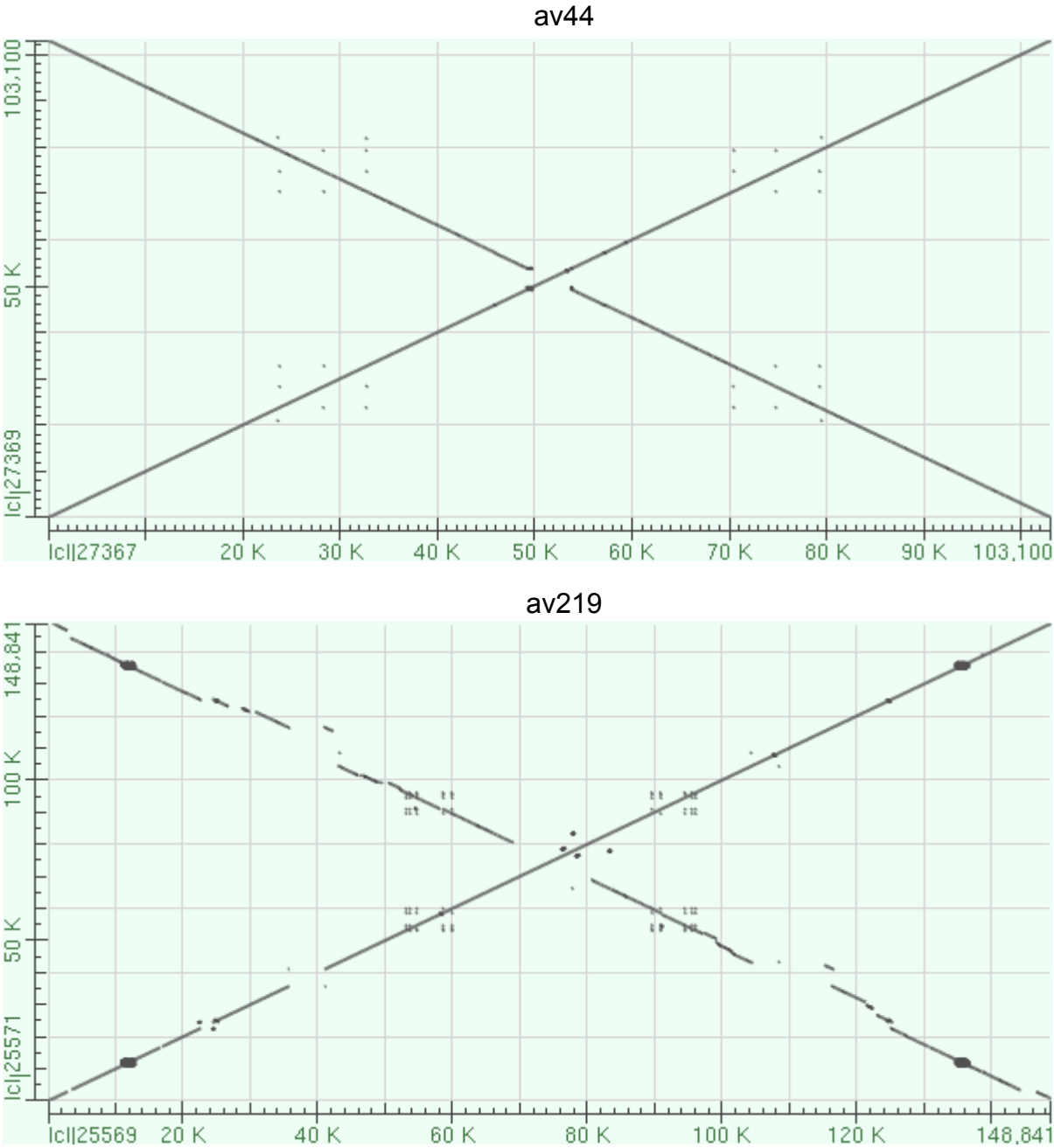


Fig. 9. Internal structure of two representative palindromes.
These dotplots were obtained by blasting the palindromic regions of two scaffolds on themselves.

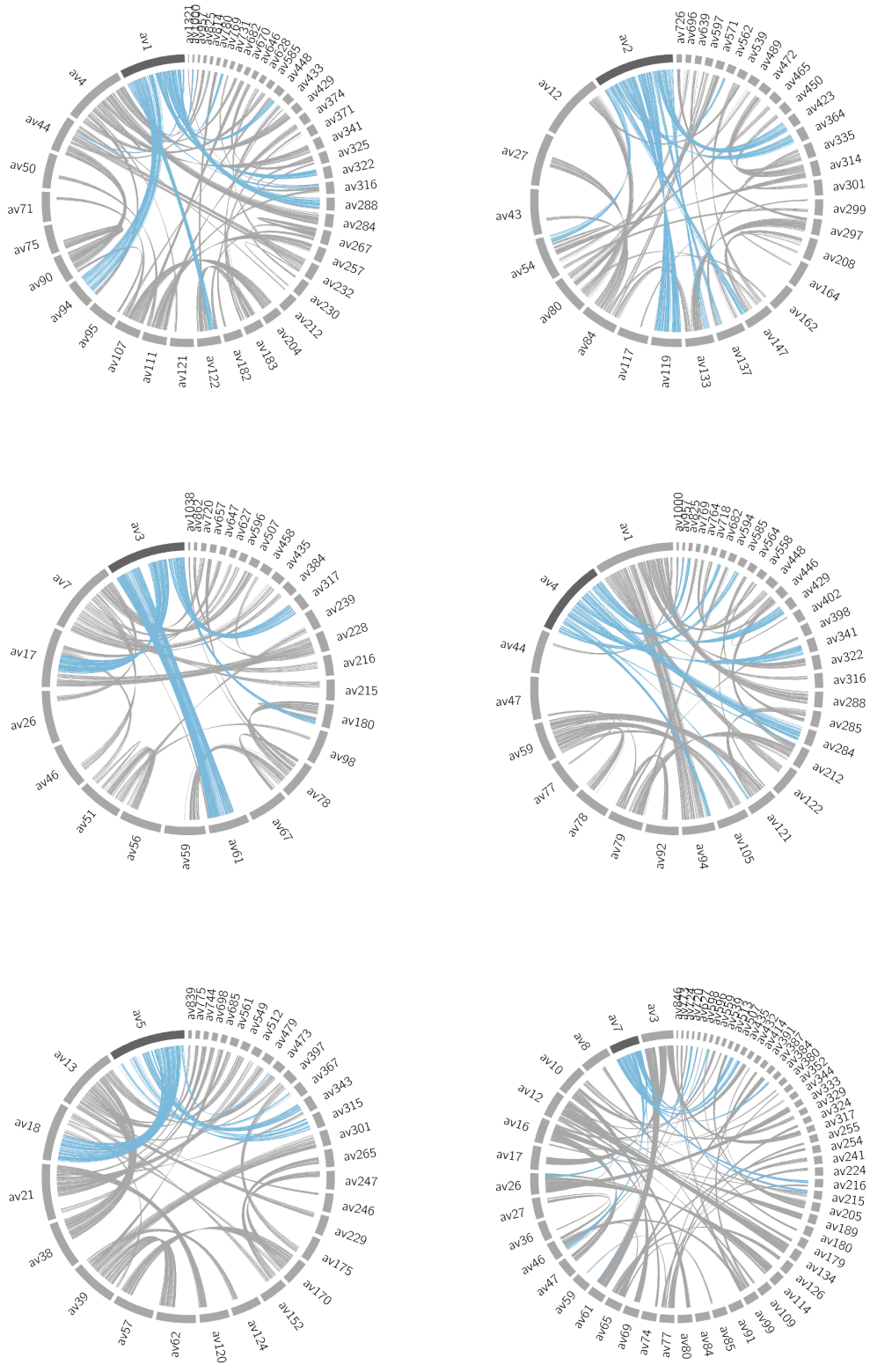


Fig. 10. Evidence for genome-scale haploidy.
18 of the 20 largest scaffolds do not have any possible homologue in the assembly due to the presence of colinearity breakpoints between alleles (all connections drawn are allelic relationships, and the one starting from the scaffold of interest are highlighted in blue).

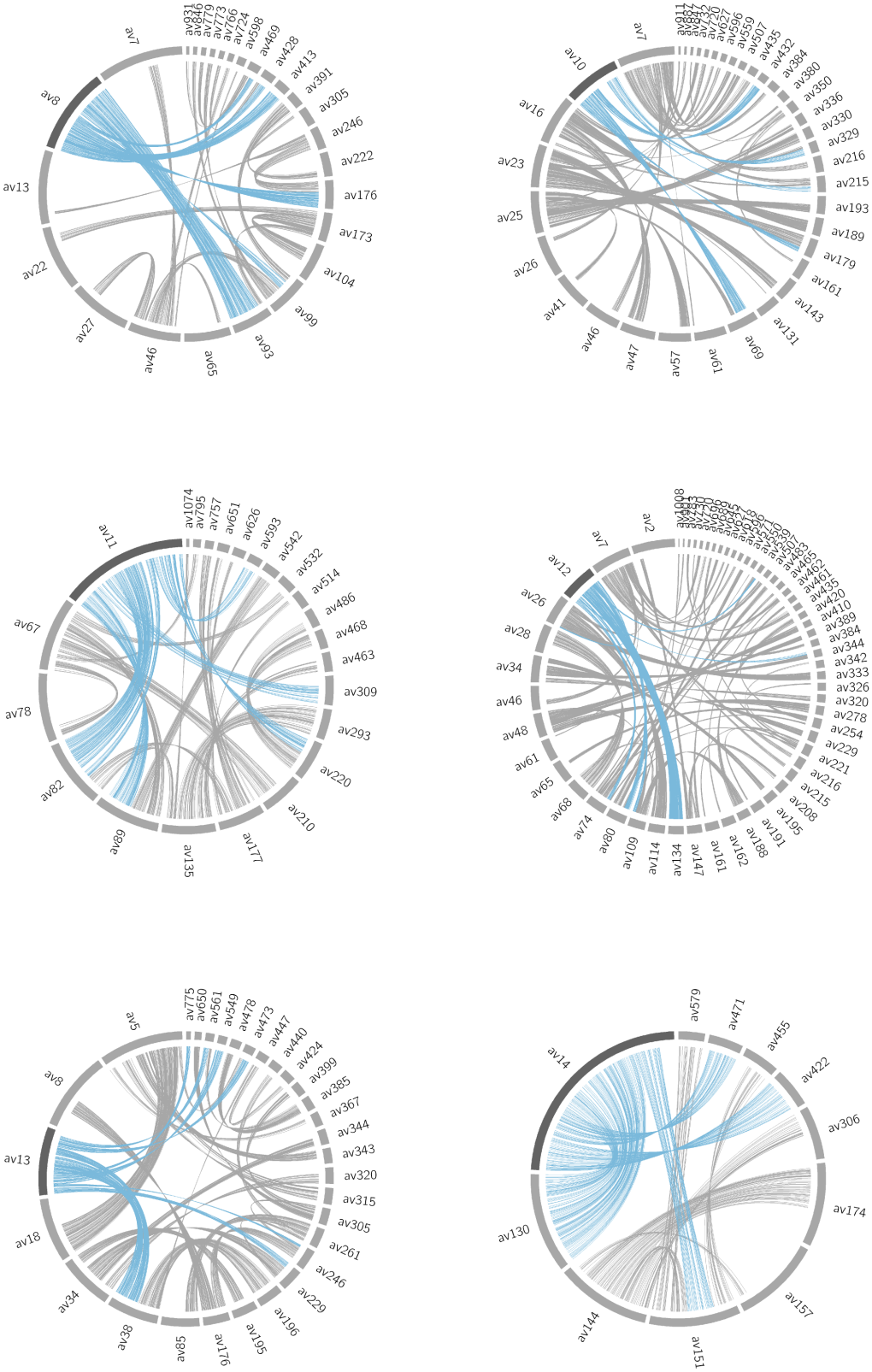


Fig. 10. Evidence for genome-scale haploidy (continued)

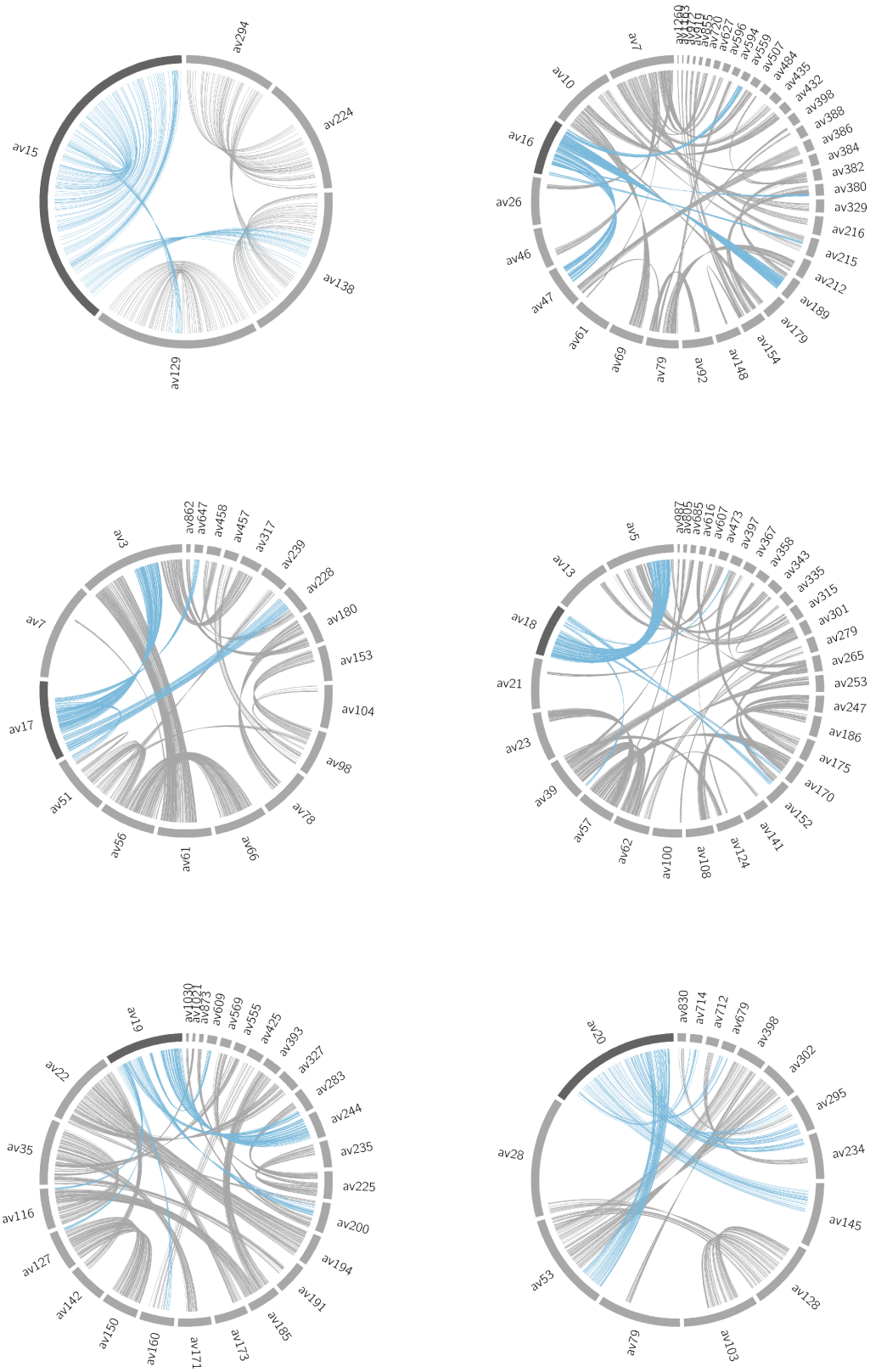


Fig. 10. Evidence for genome-scale haploidy (continued)

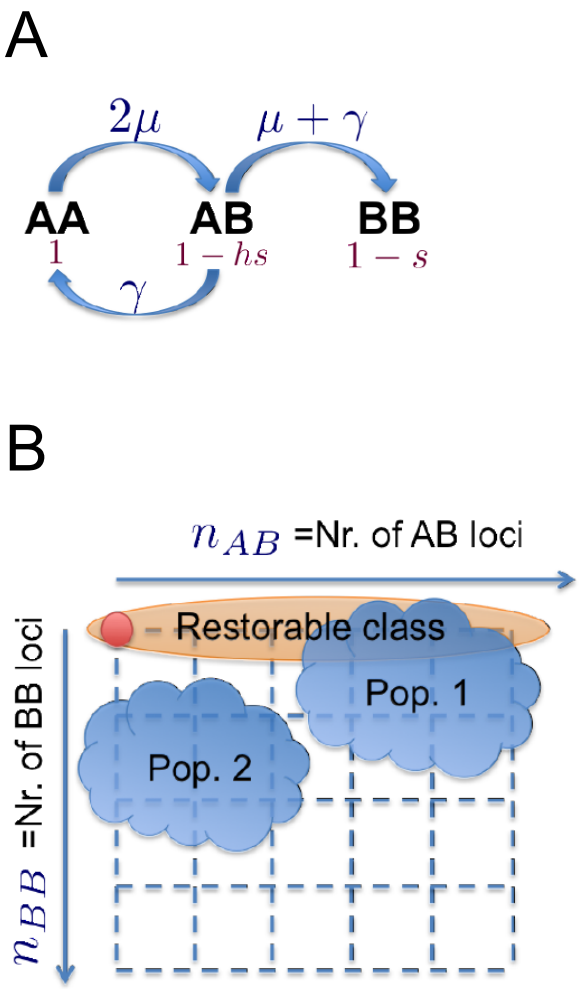


Fig. 11. Model of Muller's ratchet with gene conversion

A. Evolutionary dynamics of a diploid locus experiencing gene conversion. With a wild-type allele A and a mutant allele B, three states are possible: AA, AB and BB. The arrows indicate transitions due to deleterious mutations A→B, which occur at rate μ, and due to gene conversion occurring at rate γ. The terms underneath each state indicate the fitness of the state relative to the homozygous wild-type state. The parameter s measures the fitness detriment of the homozygous mutated state (BB); h parameterises dominance effects. B. Consequence of gene conversion for Muller's ratchet. The state of a genome with L diploid loci can be described by the number n_{AB} of AB loci and the number n_{BB} of BB loci (=L). In the absence of gene conversion, the ratchet clicks if there is no individual left at the optimum $n_{AB} = n_{BB} = 0$ (red circle), which applies to both population examples Pop.1 and Pop. 2. In the presence of gene conversion, however, Muller's ratchet does not click until there is no individual left with $n_{AB} = 0$ (the "restorable class"), which is only the case of Pop.2.

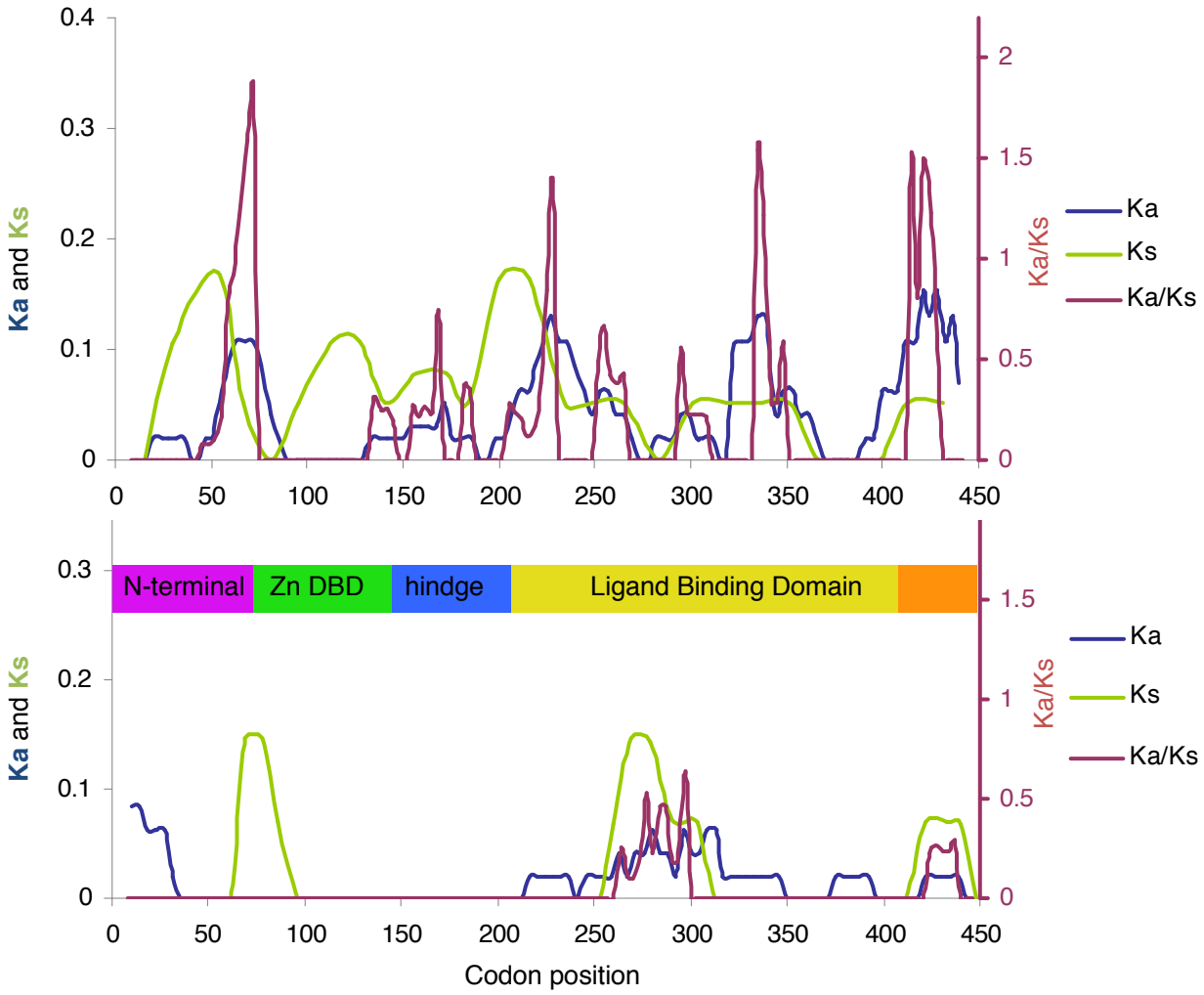


Fig. 12. Sliding windows of Ka, Ks (left axes), and Ka/Ks (right axes) of two allelic pairs of a quartet.
Total Ka, Ks, and Ka/Ks for pair A are 4.2%, 5.7%, and 0.73, respectively; and for pair B, 1.5%, 2.1%, and 0.71. Pair A has accumulated multiple local regions where Ka/Ks > 1 as well as regions where Ks > 10%. The gene encodes a Zn-finger nuclear receptor similar to the Vitamin D3 receptor B of *Crassostrea gigas*; the NCBI CDD identifies a NR_DBD_CAR C4-type Zn finger domain at codon positions 75-135 and weak similarity to a NR_LBD generalised ligand-binding domain at positions 260-409. Other domains of a generalised nuclear receptor are shown in the cartoon. The DNA binding domain is highly conserved in Pair A while areas of potential positive selection are localised to the N-terminal, Hinge, Ligand Binding, and C-terminal domains. Pair A is GSADVT00047851001 and GSADVT00053277001; Pair B is GSADVT00019567001 and GSADVT00031973001. These and other sliding window analyses were performed using DNAsp v5[86] (http://www.ub.edu/dnasp/) with a window length of 50 and step size of 10 (longer window lengths were used in some pairwise comparisons when a length of 50 encompassed regions with nonsynonymous differences but no synonymous differences).

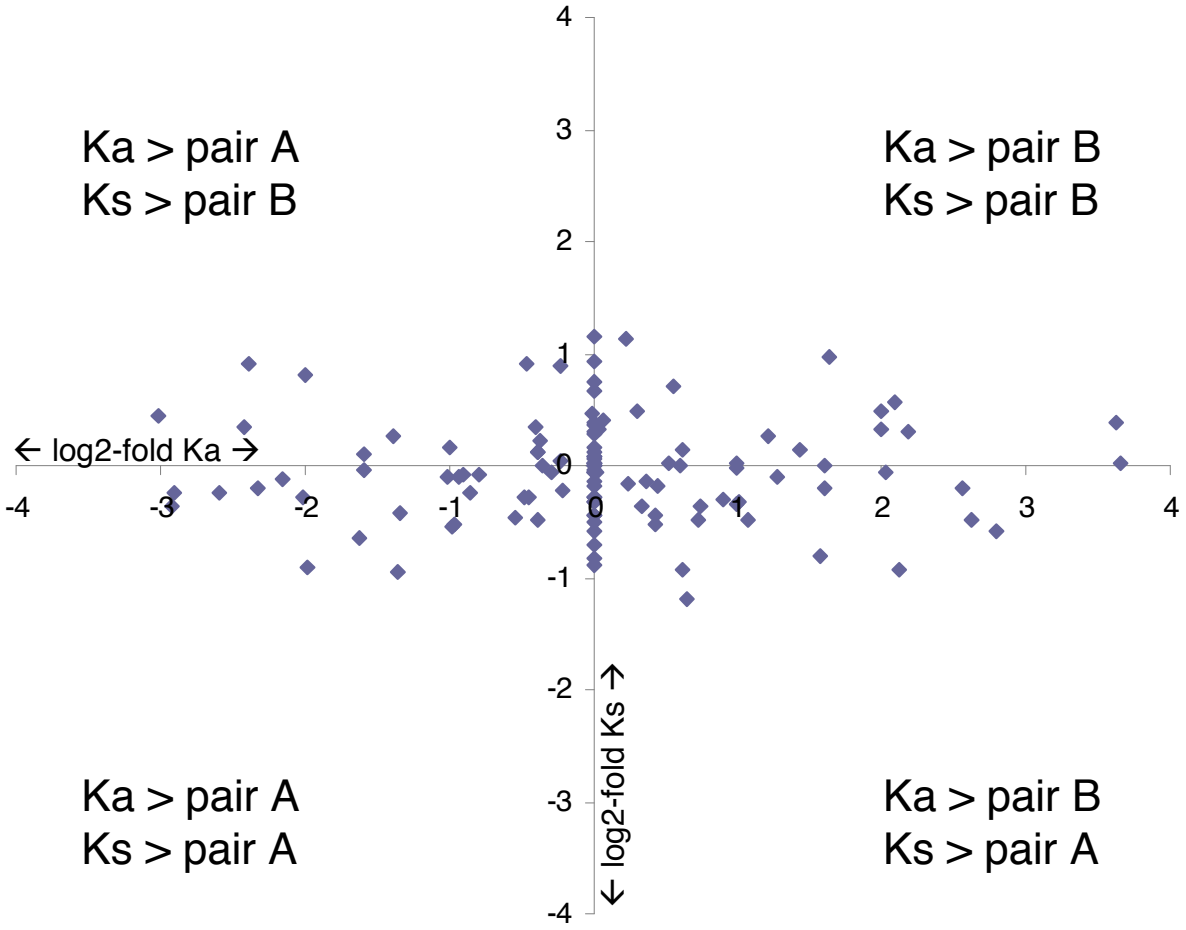


Fig. 13. Comparison of Ka and Ks between allelic pairs in quartets.

Each point represents the log2 difference in Ka (x axis) by the log2 difference in Ks (y axis) for a quartet (i.e. quadrant I shows genes where Ka and Ks are both greater in pair A than pair B; and in quadrant II Ka is greater in pair A and Ks is greater in pair B). There are 120 quartets of genes where the A pair is of equal length and the B pair is of equal length (chosen to avoid ambiguity in the alignment) and where Ks > 3% (chosen to avoid disproportionate effects of small numbers as well as recent gene conversion). In 43 quartets the difference in Ka was > 2-fold (log2 difference > 1) while the difference in Ks was much less.

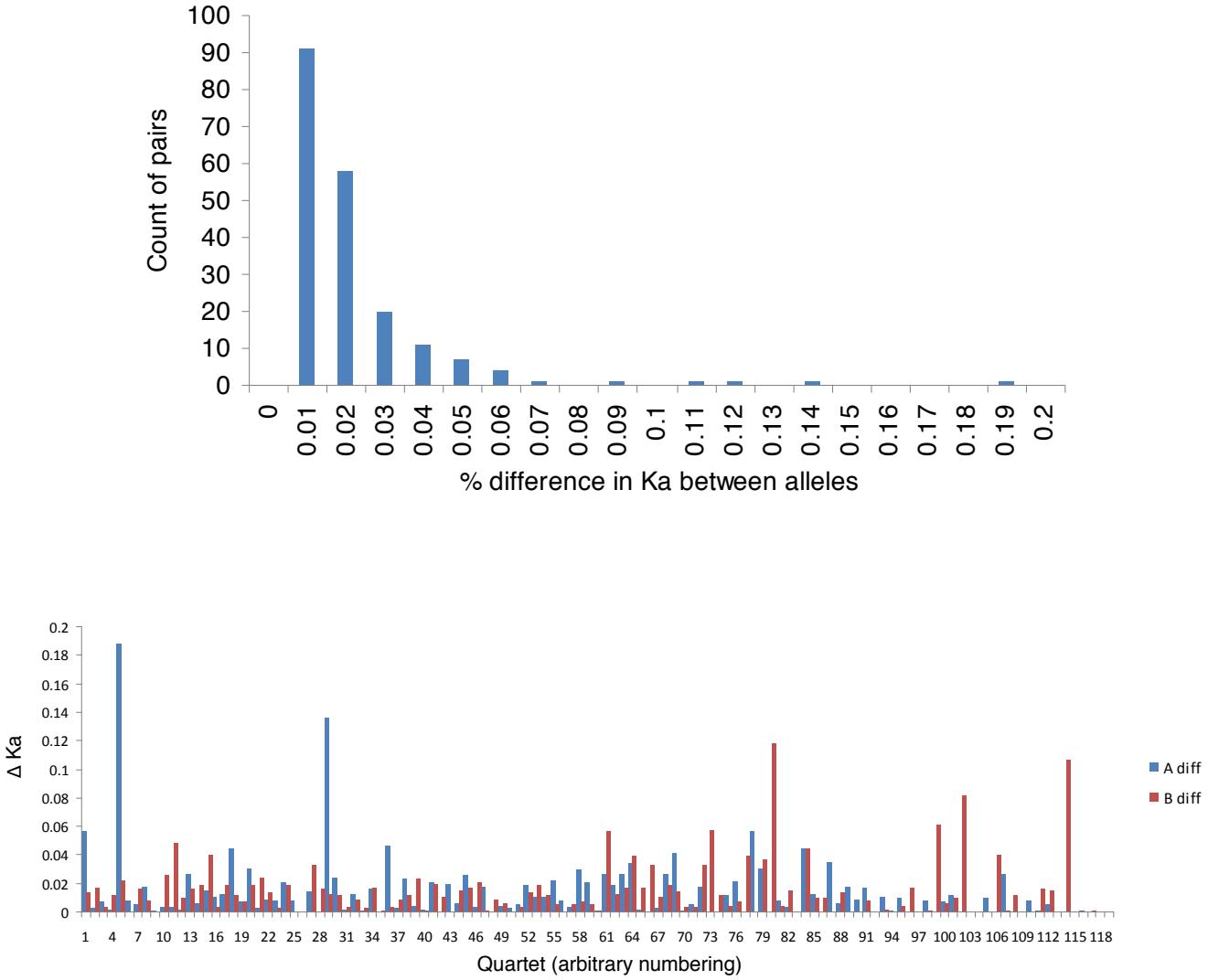


Fig. 14. Evidence for asymmetrical evolution within allelic pairs.

To assess whether some pairs of alleles may be experiencing asymmetrical evolution, we computed the % difference in Ka between each allele of a pair to the ohnologous pair, and divided it by the average distance of the pair to the ohnologous pair.

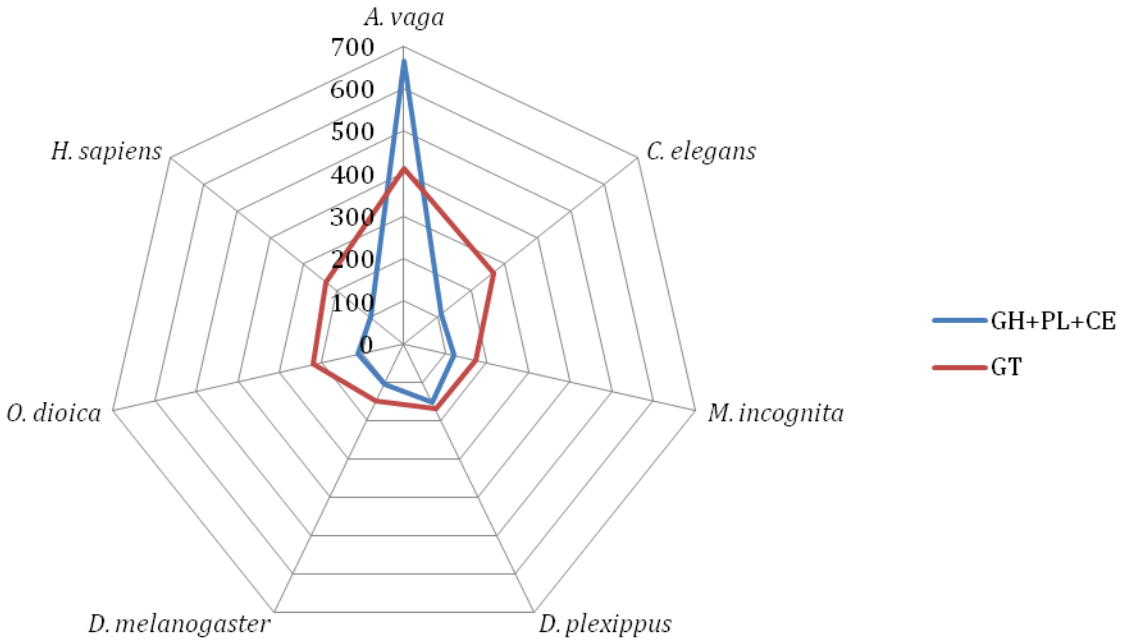


Fig. 15. Abundance of carbohydrate-degrading enzymes (GH+PL+CE) compared to enzymes involved in carbohydrate assembly (GT) in *A. vaga* and 6 other metazoans.

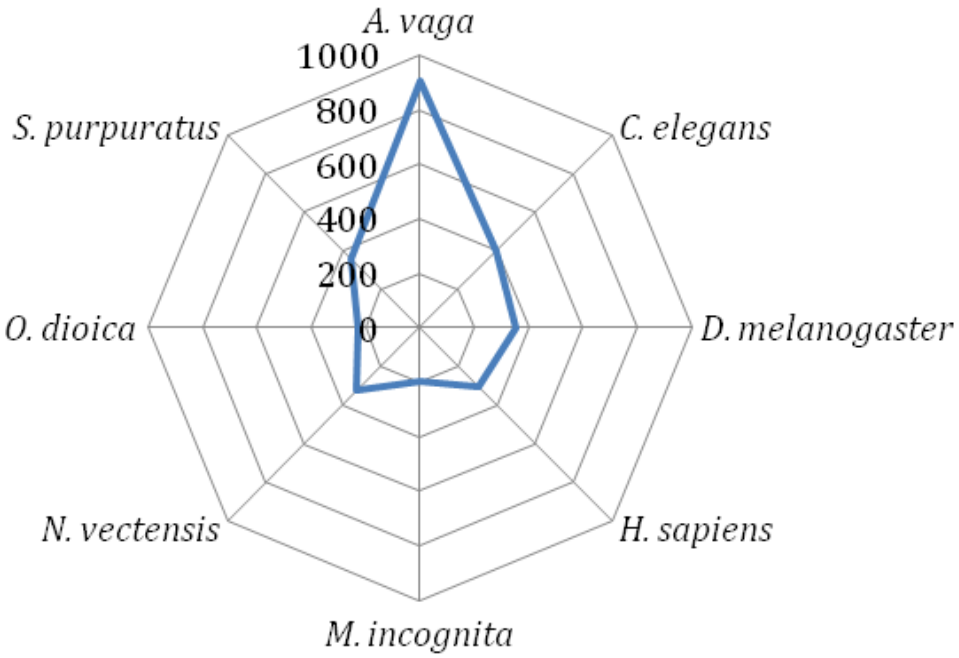


Fig. 16. Distribution of PFAM domains associated with antioxidant processes in *A. vaga* and 7 other metazoans.

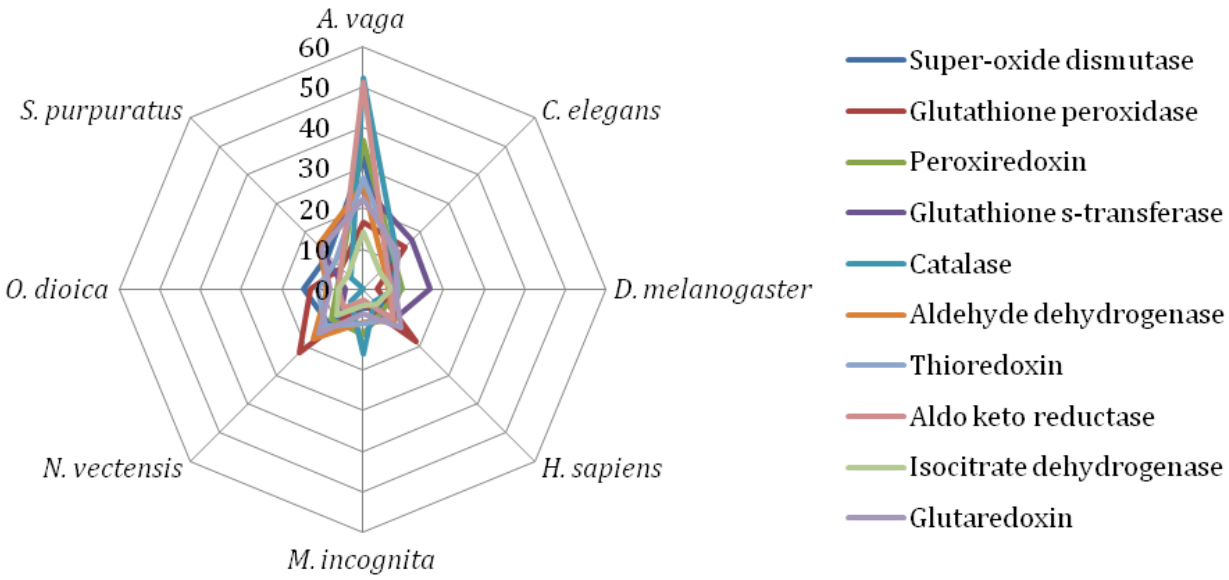


Fig. 17. Distribution of PFAM domains associated with 10 candidate antioxidant genes in *A. vaga* and 7 other metazoans.

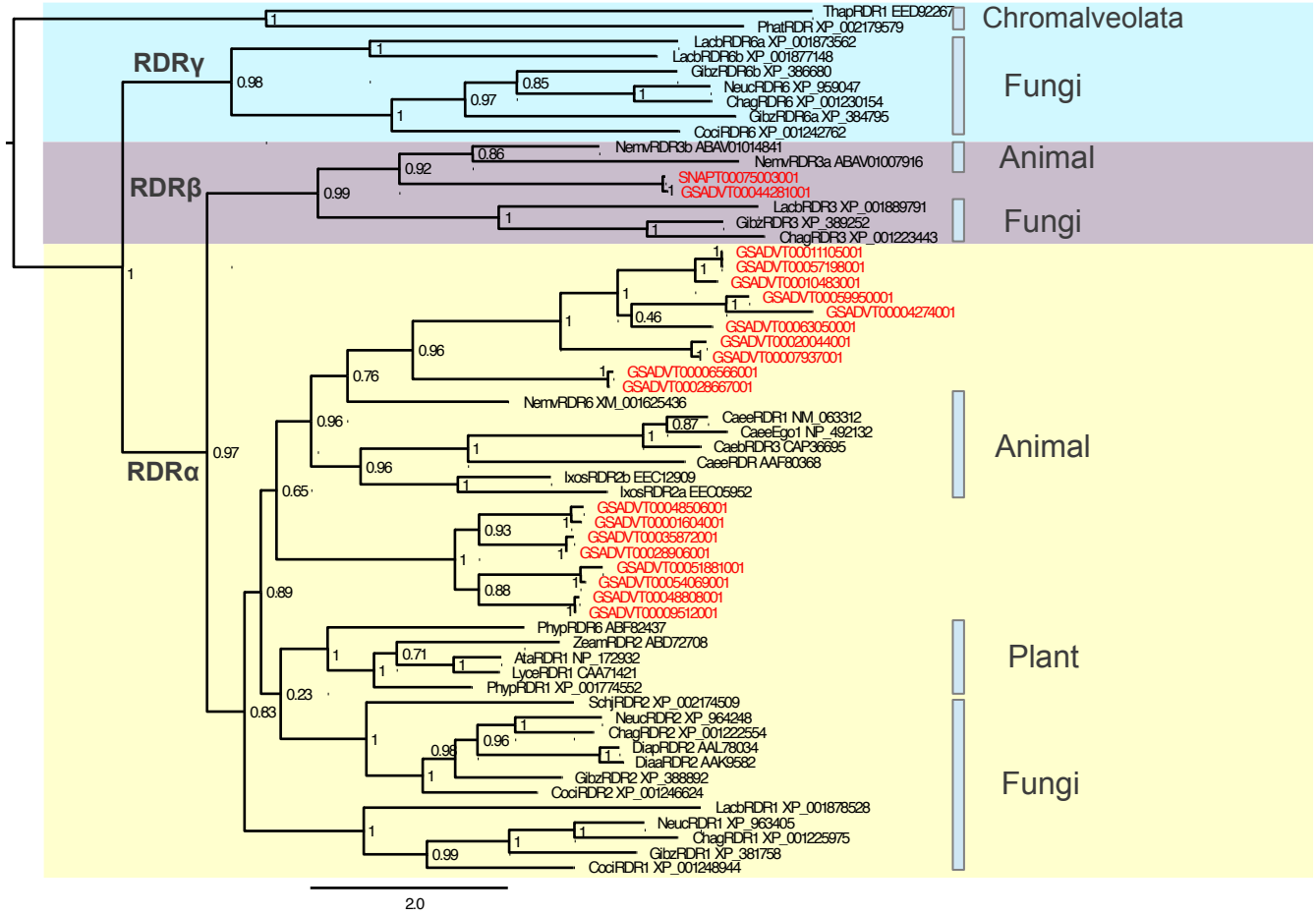


Fig. 18. Phylogenetic analysis of eukaryotic RNA-dependent RNA polymerases (RDR) and the corresponding *A. vaga* candidate proteins.
RDR genes from a subset of plants, fungi and animals[80,87] were aligned and used to construct a ML tree. The RDRα, RDRβ and RDRγ-like proteins are highlighted by different background colours. The *A. vaga* inferred gene products are in red colour. Gene names in the analysis are the same as published[80], and protein accession numbers used were: IxosRDR2a EEC05952, IxosRDR2b EEC12909, CaebRDR3 CAP36695, LacbRDR1 XP_001878528, LacbRDR3 XP_001889791, LacbRDR6a XP_001873562, LacbRDR6b XP_001877148, CociRDR1 XP_001248944, CociRDR2 XP_001246624, CociRDR6 XP_001242762, ChagRDR1 XP_001225975, ChagRDR2 XP_001222554, ChagRDR3 XP_001223443, ChagRDR6 XP_001230154, NeucRDR1 XP_963405, NeucRDR2 XP_964248, NeucRDR6 XP_959047, DiaaRDR2 AAK95829, DiapRDR2 AAL78034, GibzRDR1 XP_381758, GibzRDR2 XP_388892, GibzRDR3 XP_389252, GibzRDR6a XP_384795, GibzRDR6b XP_386680, SchjRDR2 XP_002174509, ThapRDR1 EED92267, PhatRDR XP_002179579, PhypRDR6 ABF82437, ZeamRDR2 ABD72708, LyceRDR1 CAA71421, NemvRDR6 XM_001625436, NemvRDR3a ABAV01007916, NemvRDR3b ABAV01014841, CaeeRDR1 NM_063312, PhypRDR1 XP_001774552, CaeeEgo1 NP_492132, CaeeRDR AAF80368, AtaRDR1 NP_172932.

Fig. 19. Maximum-likelihood analysis of phylogenetic relationships among Argonaute/Piwi proteins.

The Piwi and Argonaute subfamilies are highlighted by different background color areas. Numbers on each node represent branch support. The candidate *A. vaga* genes are displayed in red color. Taxa are designated by the species name, the ID of each protein, and the GenBank accession number: C.elegans_Prg-1 NM_059720, C.elegans_Prg-2 NM_068593, Drosophila_Piwi NM_057527, Drosophila_Aubergine NM_001103674, Drosophila_Ago3 NM_001043162, Planaria_Smedwi-3 EU586258, Zebrafish_Zili NM001080199, Mouse_Mili NM_021308, Human_Hili NM_001135721, Sponge_EfPiwiA AB533505, Ascidian_BpPiwi1 AB455103, Mouse_Miwi2 NM_177905, Human_Hiwi2 NM_152431, Zebrafish_Ziwi NM_183338, Mouse_Miwi NM_021311, Human_Hiwi NM_004764, Cnidarian_ClytiaPiwi, Sponge_EfPiwiB AB533506, Planaria_Smedwi1 DQ186985, Planaria_Smedwi-2 DQ186986, Tetrahymena_Twi11 EU183125, Tetrahymena_Twi1 AB084111, Sponge_EfAgoA AB533507, Drosophila_Ago1 NM_166020, Zebrafish_Ago2 XM_694134, Mouse_Ago2 AY135688, Human_Ago2 NM_001164623, Mouse_Ago3 AY135689, Zebrafish_Ago3 XM_690923, Mouse_Ago4 AY135690, Mouse_Ago1 AY135687, Drosophila_Ago2 NM_140518, D.simulans_Ago2 ADQ27056, A.aegypti_Ago2 ACR56327, T.castaneum_Ago2 NP_001107842, N.vectensis_Piwi XP_001641994, N.vectensis_Ago XP_001638494, C.elegans_Wago-1 CAA95839, C.elegans_ Wago-2 CCD65432, C.elegans_ Wago-5 CCD72518, C.elegans_ Wago-9 CAA86741, C.elegans_Wago-6 CCD71915, C.elegans_Wago-7 CCD65142, C.elegans_Wago-12 NP_508092, C.elegans_Ergo-1 NP_503362, C.elegans_Csr-1 CCD67715, C.elegans_Rde-1 CAB05546. Scale bar, amino acid substitutions per site.

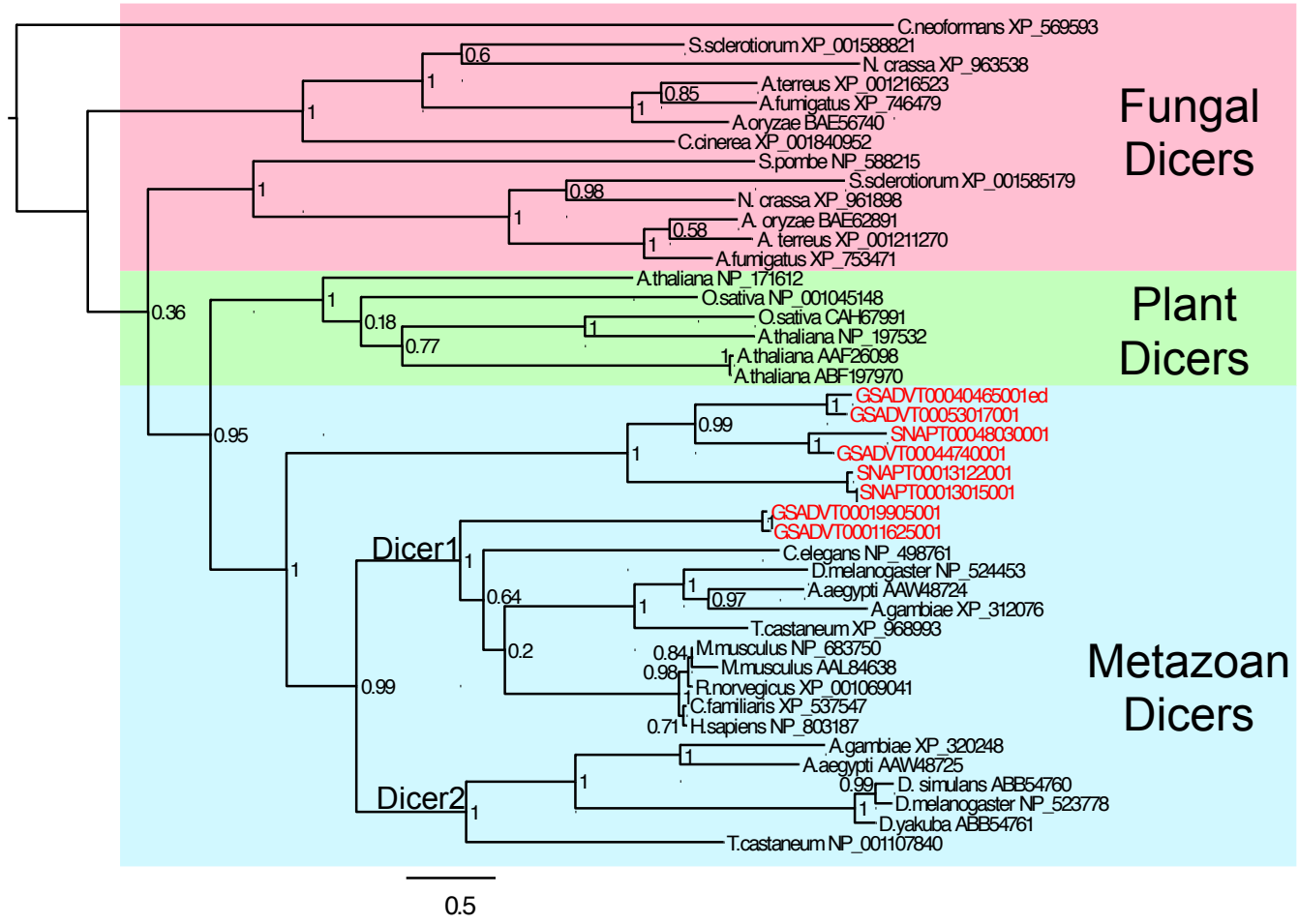


Fig. 20. Maximum-likelihood phylogenetic analysis of Dicer proteins.
Sequences corresponding to plant, fungal and metazoan species are indicated with different background color areas. Dicer1 and Dicer2 clades are labeled at the corresponding nodes. Numbers on each node represent branch support. Taxa labels are composed from the species name and the GenBank accession number: *S. pombe* NP_588215, *C. neoformans* XP_569593, *A. thaliana* NP_197532, *A. thaliana* ABF19797, *A. thaliana* AAF26098, *A. thaliana* NP_171612, *O. sativa* NP_001045148, *O. sativa* CAH67991, *C. elegans* NP_498761, *A. aegypti* AAW48724, *A. aegypti* AAW48725, *D. melanogaster* NP_524453, *D. melanogaster* NP_523778, *H. sapiens* NP_803187, *T. castaneum* NP_001107840, *T. castaneum* XP_968993, *S. sclerotiorum* XP_001585179, *S. sclerotiorum* XP_001588821, *C. cinerea* XP_001840952, *M. musculus* AAL84638, *M. musculus* NP_683750, *R. norvegicus* XP_001069041, *C. familiaris* XP_537547, *A. fumigatus* XP_746479, *A. fumigatus* XP_753471, *A. terreus* XP_001216523, *A. terreus* XP_001211270, *A. oryzae* BAE62891, *A. oryzae* BAE56740, *N. crassa* XP_961898, *N. crassa* XP_963538, *A. gambiae* XP_320248, *A. gambiae* XP_312076, *D. simulans* ABB54760, *D. yakuba* ABB54761.

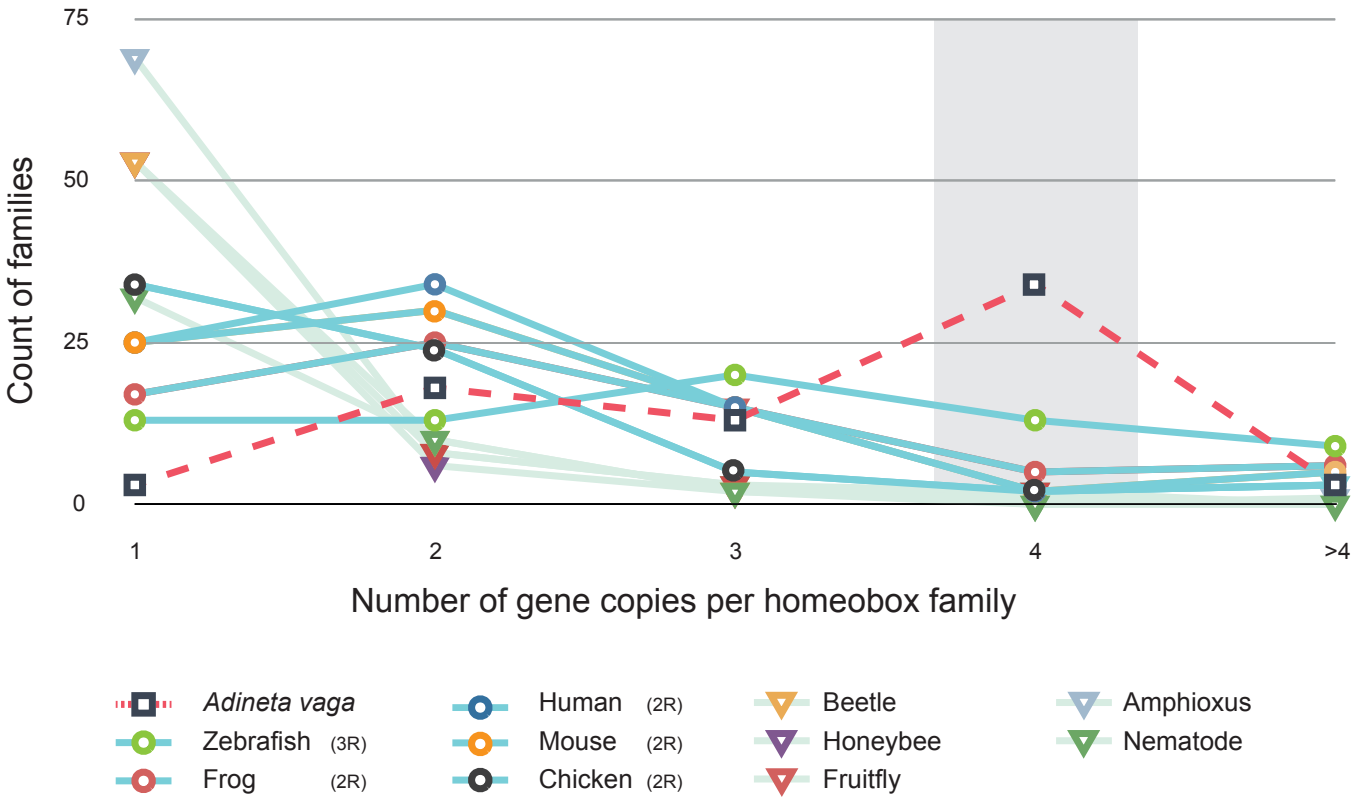


Fig. 21. Distribution of copy number of homeobox genes in metazoans.

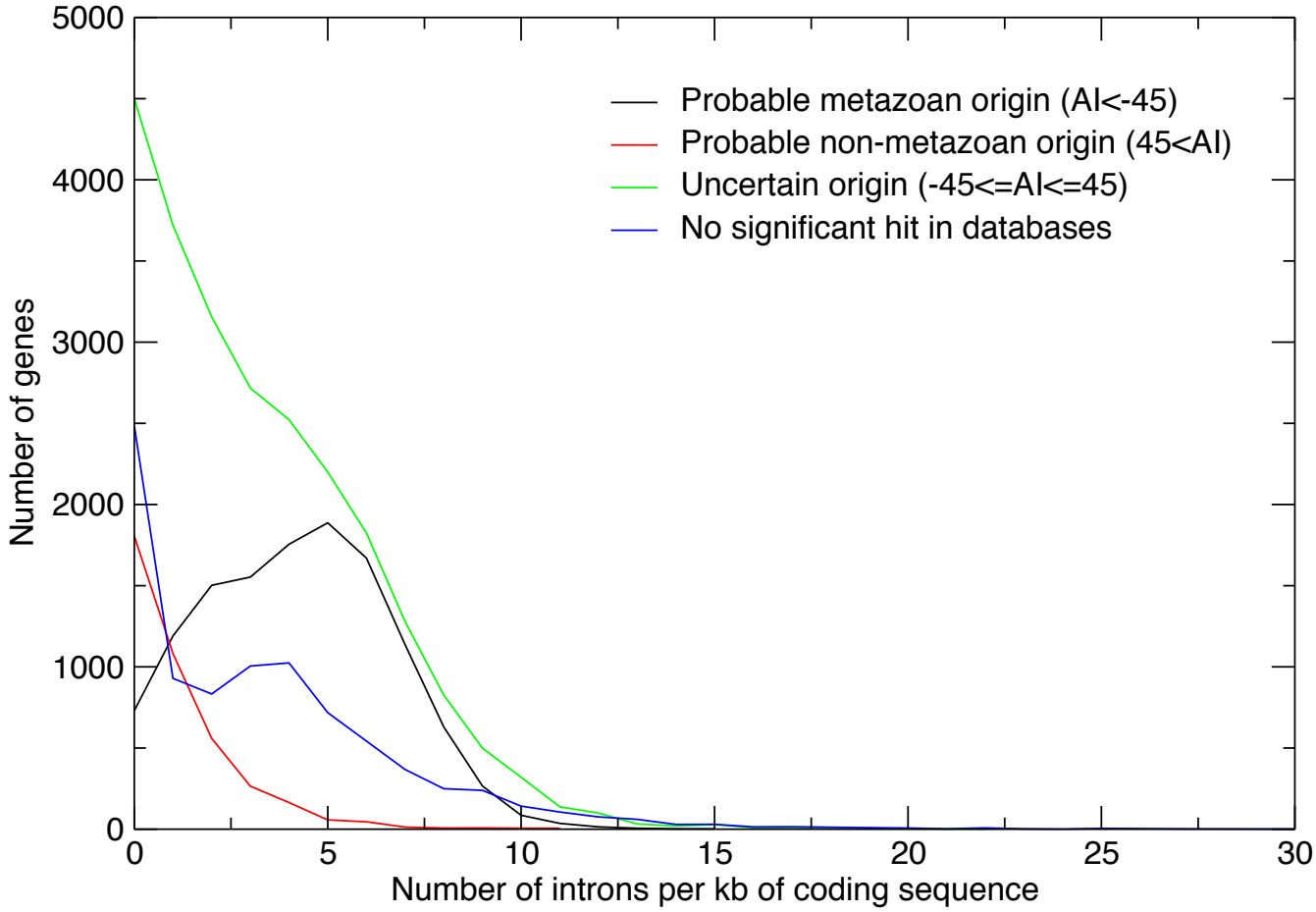


Fig. 22. Distribution of density of introns according to origin of genes.

Supplementary Tables

Table 1. Sequencing statistics.

	Material	Library type and sequencing platform	Sequencing laboratory	Million reads	Genomic coverage
Genomic DNA	Genoscope_A	454 Titanium single reads	Genoscope	4,7	7X
	Genoscope_B	454 GS-FLX mate pairs 3 Kb insert	Genoscope	6,9	7X
	Genoscope_C	454 Titanium mate pairs 8 Kb insert	Genoscope	3,4	4X
	Genoscope_F	454 Titanium mate pair 20 Kb insert	Genoscope	1,9	3X
	Genoscope_F	Illumina HiSeq2000 paired ends 2X100 bp	Genoscope	100	96X
	Genoscope_G	Illumina HiSeq2000 mate pairs 11 Kb insert 2X101 bp	Genoscope	169	157X
	MBL_Lib1	Illumina GAIIx paired ends 690 bp insert 2X160 bp	U. Michigan	85,6	126X
	MBL_Lib2	Illumina HiSeq1000 paired ends 450bp insert 2X101 bp	MBL	66,8	30X
cDNA	A	454 Titanium single reads	Genoscope	0,8	N/A
	A-S (total from 12 libraries)	Ilumina GAIIX single reads 76bp	Genoscope	350	N/A
	MBL1	454 Titanium single reads	MBL	0,9	N/A

Table 2. Assembly statistics and gene predictions.

Assembly	Number	Mean size (Kb)	N50 (Kb)	N90 (Kb)	Longest (Kb)	Total size (Mb)	Percentage of the assembly
Contigs	48,994	4.4	46.7	2.5	563.3	213.8	98%
Scaffolds	38,875	5.6	259.9	7.6	1 087.3	218.1	100%

Annotated features	Number	Mean size (bp)	%GC	Total size (Mb)	Percentage of the assembly
Genes	49,300	1,816	31.0	89.5	41.1%
Exons CDS	252,985	225	33.0	56.9	26.1%
Introns CDS	203,685	99	26.3	20.2	9.3%
Intergenic	51,792	2081	29.4	128.6	58.9%
Transposons	14,343	452	34.0	5.5	2.52%

Gene category	All	Probable metazoan origin (AI≤-45)	Uncertain origin (-45<AI≤45)	Probable non-metazoan origin (45<AI)	No significant hit in GenBank
Number	49,300	12,658	23,708	4,019	8,915
% of all genes	100%	25.7%	48.1%	8.2%	18.0%
Mean GC%	33.3	33.1	33.5	34.6	32.4

Table 3. Inventory of known TE families in *A. vaga*.

No.	Class I: LTR	Full-length[a]	Fragments[b]	Total kb[c]	No.	Class II: TIR	Full-length	Fragments	Total kb	No.	Class II: TIR	Full-length	Fragments	Total kb
1	Vesta1a	2	22	47.5	79	Avmar1	25	70	79.0	168	hAT16	1	7	7.2
2	Vesta1b	1	7	15.5	80	Avmar1a	5	36	32.5	169	hAT17	1	6	8.2
3	Vesta2	2	20	44.2	81	Avmar2	3	3	4.9	170	hAT18	0	2	1.3
4	Vesta3	2	18	23.8	82	Avmar3	1	3	1.9	171	hAT19	1	9	9.1
5	Vesta4a	1	4	17.8	83	Avmar4	1	4	2.8	172	hAT20	1	21	22.1
6	Vesta4b	1	5	10.2	84	Avmar5	1	6	4.8	173	hAT21	1	5	11.4
7	Vesta5	2	9	20.3	85	Avmar5a	1	7	3.7	174	hAT22	1	8	12.6
8	Vesta6a	2	5	11.4	86	Avmar6	2	4	8.4	175	hAT23	1	7	9.8
9	Vesta6b	1	7	7.7	87	Avmar7	2	18	7.8	176	hAT24	1	3	7.4
10	Vesta6c	4	9	26.7	88	Avmar8	1	13	8.2	177	hAT25	1	9	10.0
11	Vesta7	1	12	8.7	89	Avmar9	2	7	7.1	178	hAT26	1	7	7.4
12	Juno1	5	74	75.9	90	Avmar10	1	1	1.6	179	hAT27	1	1	3.4
13	Juno2	3	54	90.3	91	Avmar11	1	13	5.9	180	hAT28	1	11	11.5
14	Juno3	1	6	10.0	92	Tc1	3	16	20.3	181	hAT29	1	48	27.7
15	Juno4a	2	0	11.4	93	Tc2	1	5	4.5	182	hAT30	1	12	6.7
16	Juno4b	1	14	9.2	94	Tc3	1	0	1.3	183	hAT31	1	3	4.1
17	TelKA1	1	20	20.6	95	Tc4	6	34	38.3	184	hAT32	1	3	7.1
18	TelKA1a	3	23	33.5	96	Tc5	2	10	11.0	185	CACTA1	1	12	13.9
19	TelKA2	2	7	21.1	97	Tc6	2	16	8.0	186	CACTA2	1	5	3.2
20	TelKA3a	1	18	20.0	98	Tc7	1	9	4.3	187	CACTA3	1	2	5.3
21	TelKA3b	1	10	15.5	99	Tc8	1	4	6.6	188	CACTA4	1	2	7.1
22	TelKA4	2	10	24.5	100	Tc9	0	6	4.0	189	CACTA5	1	3	4.8
23	Mag	1	10	13.8	101	Tc10	1	8	5.7	190	CACTA6	1	5	7.4
	All LTR	**42**	**364**	**579.7**	102	Tc11	2	2	4.7	191	CACTA8	1	47	44.5
	Class I: non-LTR				103	Tc12	3	21	23.9	192	EnSpm1	1	42	60.1
24	Hebe	14	116	218.4	104	Tc13	1	20	11.2	193	EnSpm2	1	26	26.8
25	Hebe2	3	50	68.9	105	Tc14	2	8	6.8	194	EnSpm3	1	11	18.2
26	Tx1	2	25	31.5	106	Tc15	2	6	6.9	195	EnSpm4	1	32	35.0
27	RTE1	4	42	62.8	107	Tc16	4	1	6.2	196	EnSpm5	1	11	12.0
28	RTE2	3	15	27.0	108	Tc17	1	7	6.8	197	EnSpm6	0	2	2.0
29	RTE2a	2	3	9.5	109	Tc18	3	2	4.6	198	MuDR1	2	11	11.0
30	RTE3	3	30	40.5	110	Tc19	1	18	22.6	199	MuDR2	2	18	15.8
31	RTE4	1	8	7.9	111	Tc20	2	20	16.2	200	MuDR3	2	46	37.7
32	RTE5_BovB[d]	0	31	26.8	112	Tc21	2	3	4.8	201	MuDR4	1	5	7.0
33	Soliton1	1	4	5.7	113	Tc22	0	2	1.2	202	MuDR5	2	10	12.2
34	Soliton2	1	13	11.4	114	pogo1	1	3	5.5	203	MuDR6	2	12	6.9
35	Soliton3	1	8	17.0	115	pogo2	1	2	3.4	204	MuDR7	1	6	12.8
36	Soliton4	1	1	5.7	116	pogo3	1	12	15.3	205	MuDR8	2	11	8.7
37	Soliton5	1	2	5.3	117	pogo4	4	35	50.7	206	MuDR9	0	5	1.6
38	Soliton6	1	10	7.4	118	pogo5	3	10	14.0	207	MuDR10	0	5	5.9
39	Soliton6a	1	8	17.8	119	pogo6	1	22	21.5	208	MuDR11	2	3	4.9
40	Soliton6b	1	11	15.3	120	pogo7	2	18	22.1	209	MuDR12	1	7	4.4
41	Soliton7	1	4	14.6	121	pogo8	2	18	17.1	210	MuDR13	1	6	4.5
42	Soliton8	1	14	21.3	122	pogo9	1	11	18.2	211	MuDR14	0	7	2.8
43	Soliton9a	2	11	14.8	123	pogo10	1	6	6.0	212	MuDR15	0	13	5.2
44	Soliton9b	1	15	23.2	124	pogo11	1	6	7.0	213	MuDR16	0	5	4.2
45	R9	4	10	16.9	125	pogo12	2	7	7.5	214	MULE1	1	11	26.7

46	R9a	2	11	15.8	126	pogo13	1	4	6.6	215	MULE2	2	52	49.5
47	R4	4	15	29.4	127	Tigger	2	8	12.4	216	MULE3	1	36	43.6
	All non-LTR	**55**	**457**	**714.7**	128	Tigger2	2	32	15.1	217	MULE4	0	3	5.8
	Class I: PLE				129	AvPB1	1	15	8.3	218	MULE5	1	4	3.6
48	Penelope1	1	17	10.3	130	AvPB2	1	18	19.1	219	Sola2	1	19	28.5
49	Penelope1a	1	27	10.3	131	AvPB3	2	41	40.0	220	Sola2a	1	5	7.0
50	Penelope2	1	13	10.9	132	AvPB4	2	8	9.8	221	Sola2b	1	29	25.8
51	Penelope3	2	36	37.3	133	piggyBac1	2	7	4.7	222	Sola2c	2	7	15.6
52	Penelope3a	1	7	4.7	134	piggyBac2	1	18	14.6	223	Sola2d	1	11	13.3
53	Penelope4	1	10	10.9	135	piggyBac3	2	13	17.1	224	Sola2e	1	6	12.6
54	Penelope5	1	17	12.2	136	piggyBac4	0	7	4.7	225	Sola2f	1	15	10.9
55	Penelope6	0	8	7.9	137	piggyBac5	1	15	12.7	226	Sola2g	2	5	7.4
56	AthenaT	2	98	179.3	138	piggyBac6	1	16	17.0	227	Sola2h	1	32	17.3
57	AthenaS	1	14	29.8	139	piggyBac7	0	2	1.3	228	Sola2i	0	6	1.9
58	AthenaQ	2	15	30.8	140	piggyBac8	2	4	5.9	229	Sola3a	1	6	14.9
59	AthenaR	2	8	17.3	141	piggyBac9	0	3	1.4	230	Sola3b	1	11	15.6
60	AthenaP	2	10	19.2	142	Pokey1	2	10	8.5	231	Sola3c	0	2	1.6
61	AthenaK	1	10	13.4	143	Pokey2	1	13	16.2	232	Ginger2	2	5	8.4
62	AthenaN	1	34	41.3	144	Pokey3	3	32	33.8	233	Ginger2a	0	6	7.3
63	AthenaJ	2	24	57.0	145	Pokey4	1	3	5.4	234	Ginger2b	1	8	7.4
64	AthenaM	2	20	23.7	146	Uribo1	2	25	26.1	235	Ginger2c	2	6	8.5
65	AthenaO	2	32	35.9	147	Uribo2	2	24	20.8	236	Ginger2d	1	10	15.1
66	AthenaL1	2	23	31.5	148	Looper	1	3	3.5	237	Ginger2e	2	14	9.9
67	AthenaL2	1	40	40.5	149	Looper2	1	0	2.3	238	IS5	1	2	2.8
68	AthenaL3	2	15	27.6	150	Looper3	2	25	19.0	239	ISL2EU1	1	33	34.4
69	AthenaV	1	27	48.7	151	Looper4	1	13	11.1	240	ISL2EU2	1	12	13.6
70	AthenaW	2	34	24.9	152	hAT1	2	24	14.5	241	ISL2EU3	2	2	3.1
71	AthenaX	0	14	9.7	153	hAT2	1	3	5.1	242	Merlin1	1	9	3.2
	All PLE	**32**	**532**	**735.1**	154	hAT3	1	4	8.6	243	Merlin2	1	2	1.9
	All Class I	**129**	**1353**	**2029.5**	155	hAT3a	2	11	7.9	244	Merlin3	1	1	0.8
	Class II: Helitron				156	hAT4	1	2	3.5	245	Merlin4	1	12	5.6
72	Heli1	2	110	65.8	157	hAT5	1	17	14.8	246	Merlin5	0	6	1.5
73	Heli2	2	102	198.0	158	hAT6	0	1	2.8	247	Harbinger1	1	0	2.7
74	Heli3	0	26	83.9	159	hAT7	1	26	20.6	248	Harbinger2	1	6	4.2
75	Heli3a	1	62	42.6	160	hAT8	1	3	6.4	249	Harbinger3	1	4	3.3
76	Heli4	0	2	1.8	161	hAT9	1	27	18.4	250	Kolobok	1	44	32.7
77	Helentron	2	60	102.4	162	hAT10	1	1	2.5	251	Kolobok2	1	17	19.0
	All Helitron	**7**	**362**	**494.5**	163	hAT11	1	19	19.6	252	Kolobok3	1	7	14.8
	Class II: MITE[e]	**n/a**	**151**	**46.5**	164	hAT12	1	4	7.1	253	Kolobok4	0	7	10.3
78	Class II: Crypton	**2**	**8**	**8.5**	165	hAT13	1	7	4.5	254	Zator1	1	20	31.1
	All Class II	**259**	**2639**	**2738.3**	166	hAT14	1	21	21.1	255	Zator2	2	4	10.1
	All Class I+II	**388**	**3992**	**4767.8**	167	hAT15	1	7	11.4		**All TIR**	**250**	**2118**	**2188.8**

[a]Number of full-length copies as assessed by BLAT, using full-length sequences as queries.
[b]Number of fragmented copies longer than 50 bp.
[c]Total length of genomic DNA in kb occupied by the family.
[d]RTE5 belongs to the RTE BovB clade, which is particularly prone to horizontal transfers[88,89].
[e]MITEs are TIR families for which the transposase source is not known.

Table 4. Characteristics of LTR retrotransposon families.

LTR family	Full-length	Fragments not solo	Solo LTRs	LTR length/ mismatch
Vesta1a	2	10	12	509/1
Vesta1b	1	2	5	508/4
Vesta2	2	18	2	166/3
Vesta3	2	8	10	159/0
Vesta4a	1	2	2	196/0
Vesta4b	1	4	1	306/0
Vesta5	2	4	5	225/2
Vesta6a	2	3	2	273/0
Vesta6b	1	3	4	255/0
Vesta6c	4	5	4	167/0
Vesta7	1	4	8	213/0
Juno1	5	54	20	302/0
Juno2	3	52	2	175/2
Juno3	1	4	2	387/1
Juno4a	2	0	0	369/0
Juno4b	1	5	9	284/0
TelKA1	1	10	10	551/2
TelKA1a	3	19	4	325/4
TelKA2	2	4	3	170/0
TelKA3a	1	11	7	509/7
TelKA3b	1	8	2	550/0
TelKA4	2	9	1	266/3
Mag	1	8	2	275/1

The second column is the same as in Supplementary Table 3, and the third column represents the difference between the total number of fragments in Supplementary Table 3 and the number of solo LTRs. Also shown are the LTR length and the number of mismatches between LTRs for each family. For Juno1, the 5‘ and 3‘ segments that were located at scaffold ends and flanked by the same target site duplication were counted as full-length, as were the two circularly permuted contigs with the LTR in the middle consisting only of Juno1 sequences. Copies with internal gaps of Ns introduced during scaffolding were also counted as full-length.

Table 5. Comparison of the CAZyme repertoire in 7 metazoan species, including *A. vaga*.

Species / Class	GH	fam	GT	fam	PL	fam	CE	fam	CBM	fam
C. elegans	111	23	267	37	0	0	1	1	80	7
M. incognita	91	23	173	25	30	1	0	0	32	5
D. plexippus	151	24	170	40	0	0	2	2	100	9
D. melanogaster	103	23	150	41	0	0	2	2	260	9
A. vaga	**623**	**54**	**412**	**45**	**13**	**3**	**27**	**5**	**154**	**15**
O. dioica	108	25	219	35	1	1	1	1	36	6
H. sapiens	98	29	234	43	0	0	2	2	39	6

Abundance and number of families (fam) found in each CAZyme class for each metazoan. GH = Glycoside Hydrolase, GT = Glycosyl Transferase, PL = Polysaccharide Lyase, CE = Carbohydrate Esterase, CBM = Carbohydrate Binding Module, following the nomenclature of http://www.cazy.org

Table 6. AI indexes per CAZyme class in *A. vaga.*

CAZy family	Average AI	SD	%AI>0	%AI>45	%AI>100
GT	-37	126	24%	**17%**	9%
GH	46	140	55%	**43%**	28%
PL	85	67	92%	**61%**	54%
CE	87	82	85%	**85%**	40%
CBM	3	163	49%	**30%**	21%

SD = Standard deviation; %AI>0 = percentage of CAZymes with AI>0; %AI>45 = percentage of CAZymes with AI>45; %AI>100 = percentage of CAZymes with AI>100

Table 7. CAZymes degrading chitin found in *A. vaga*.

	Degradation			
Family/Species	GH18	GH19	GH20	**Total**
C. elegans	43	6	5	54
M. incognita	2	2	4	8
D. plexippus	21	0	9	30
D. melanogaster	22	0	4	26
A. vaga	53	15	28	**96**
O. dioica	7	0	8	15
H. sapiens	8	0	3	11

Table 8. Comparison of selected candidate antioxidant genes between *A. vaga* and 7 other species.

*overabundance if >2,2

Family name	*A. vaga*	*C. elegans*	*D. melanogaster*	*H. sapiens*	*M. incognita*	*N. vectensis*	*O. dioica*	*S. purpuratus*	Average metazoan	*A. vaga* / average metazoan*	PFAM domain associated with family name
Super-oxide dismutase	**27**	8	7	5	4	10	12	10	8,0	3,4	PF00080, PF00081, PF02777
Glutathione peroxidase	**9**	8	2	10	2	12	7	4	6,4	1,4	PF00255
Peroxiredoxin	**169**	49	45	33	51	49	29	34	41,4	4,1	PF10417, PF00578, PF00141, PF01328, PF08534
Glutathione s-transferase	**283**	190	184	118	32	82	48	170	117,7	2,4	PF00043, PF02798, PF13409, PF13410, PF13417
Catalase	**13**	3	2	1	4	1	0	1	1,7	7,6	PF00199
Aldehyde dehydrogenase	**45**	13	11	19	14	30	15	27	18,4	2a,4	PF00171
Thioredoxin	**225**	93	72	80	70	105	90	79	84,1	2,7	PF00085, PF13192, PF13743, PF13848, PF13899, PF13098, PF13192, PF13462
Aldo keto reductase	**93**	15	13	19	5	12	11	14	12,7	7,3	PF00248
Isocitrate dehydrogenase	**14**	6	7	5	4	9	6	5	6,0	2,3	PF00180
Glutaredoxin	**25**	14	9	15	7	17	9	16	12,4	2,0	PF00462, PF04399, PF05768

Table 9. Comparison of selected candidate antioxidant genes between *A. vaga* and *C. elegans*

Family name	Number of genes in *A. vaga*	Number of genes in *C. elegans* according to public databases	% AI>45 in *A. vaga*	% genes expressed in hydrated *A. vaga*
Superoxide dismutase	24	5	0	95,8
Glutathione peroxidase	11	6	0	100
Peroxiredoxin	12	3	0	91,7
Glutathione s-transferase	75	49	22,6	84
Catalase	14	3	14,3	57,1
Glutathione reductase	3	2	66,7	66,7
Aldehyde dehydrogenase	23	11	0	69,6
Thioredoxin	45	13	0	86,7
Aldo keto reductase	56	13	60,7	48,2
Isocitrate dehydrogenase	15	6	0	93,3

Supplementary Data

Data 1_blocks_alleles.tab

This file lists all allelic colinear blocks.

Data 2_blocks_ohnologs.tab

This file lists all ohnologous colinear blocks.

Data 3_pairs_alleles.tab

This file lists all allelic pairs genes.

Data 4_pairs_ohnologs.tab

This file lists all ohnologous pairs of genes.

Data 5_AI.tab

This file provides AI value and description of best score blast result against nrprot for each *A. vaga* gene.

Data 6_KaKs.tab

Curated list of 104 pairs of genes with Ka Ks values

Data 7_meiosis_genes.tab

List of meiosis genes found or not found in the genome of *A. vaga*

Data 8_homeobox_genes.zip

Phylogenetic trees of the homeobox genes of *A. vaga*

Data 9_TE_defense_genes.tab
Annotation and classification of candidate TE defence genes

Data 10_PFAM.abundancies.tab
This file provides statistics on each PFAM domain in *A. vaga* and 12 other species.

Supplementary References

1. Wallace, R. L. & Snell, T. W. Rotifera. in *Ecology and Classification of North American Freshwater Invertebrates* (eds. Thorp, J. H. & Covich, A. P.) 187-248 (Academic Press, New York, 2001).
2. Leasi, F., Rouse, G. W. & Sørensen, M. V. A new species of *Paraseison* (Rotifera: Seisonacea) from the coast of California, USA. Journal of the Marine Biological Association of the United Kingdom 92, 959-965 (2012).
3. Sørensen, M. V., Segers, H. & Funch, P. On a new Seison Grube, 1861 from coastal waters of Kenya, with a reappraisal of the classification of the Seisonida (Rotifera). Zoological Studies 44, 34-43 (2005).
4. Hsu, W. S. Oogenesis in *Habrotrocha tridens* (Milne). Biological Bulletin 111, 364-374 (1956).
5. Hsu, W. S. Oogenesis in the Bdelloidea rotifer *Philodina roseola* Ehrenberg. La Cellule 57, 283-296 (1956).
6. Garey, J. R., Schmidt-Rhaesa, A., Near, T. J. & Nadler, S. A. The evolutionary relationships of rotifers and acanthocephalans. Hydrobiologia 387-388, 83-91 (1998).
7. Mark Welch, D. B. Evidence from a protein-coding gene that acanthocephalans are rotifers. Invertebrate Biology 119, 17-26 (2000).
8. Witek, A., Herlyn, H., Ebersberger, I., Mark Welch, D. B. & Hankeln, T. Support for the monophyletic origin of Gnathifera from phylogenomics. Molecular Phylogenetics and Evolution 53, 1037-1041 (2009).
9. Sørensen, M. V. & Giribet, G. A modern approach to rotiferan phylogeny: Combining morphological and molecular data. Molecular Phylogenetics and Evolution 40, 585-608 (2006).
10. Segers, H. Annotated checklist of the rotifers (Phylum Rotifera), with notes on nomenclature, taxonomy and distribution. Zootaxa 1564, 1-104 (2007).
11. van Leeuwenhoek, A. Part of a Letter from Mr Antony van Leeuwenhoek, F. R. S. concerning Green Weeds Growing in Water, and Some Animalcula Found about Them. Philosophical Transactions (1683-1775) 23, 1304-1311 (1702).
12. Waggoner, B. M. & Poinar, G. O. Fossil habrotrochid rotifers in Dominican amber. Experientia 49, 354-357 (1993).
13. Ricci, C. Anhydrobiotic capabilities of bdelloid rotifers. Hydrobiologia 387-388, 321-326 (1998).
14. Gladyshev, E. & Meselson, M. Extreme resistance of bdelloid rotifers to ionizing radiation. Proceedings of the National Academy of Sciences 105, 5139-5144 (2008).
15. Mattimore, V. & Battista, J. R. Radioresistance of *Deinococcus radiodurans*: functions necessary to survive ionizing radiation are also necessary to survive prolonged desiccation. Journal of Bacteriology 178, 633-637 (1996).
16. Gladyshev, E. A., Meselson, M. & Arkhipova, I. R. Massive horizontal gene transfer in bdelloid rotifers. Science 320, 1210-1213 (2008).
17. Lee, J.-S. et al. Sequence analysis of genomic DNA (680 Mb) by GS-FLX-Titanium sequencer in the monogonont rotifer, *Brachionus ibericus*. Hydrobiologia 662, 65-75 (2011).
18. Hur, J. H., Van Doninck, K., Mandigo, M. L. & Meselson, M. Degenerate tetraploidy was established before bdelloid rotifer families diverged. Molecular Biology and Evolution 26, 375-383 (2009).
19. Van Doninck, K. et al. Phylogenomics of unusual histone H2A variants in bdelloid rotifers. PLoS Genetics 5, e1000401 (2009).

20. Gladyshev, E. A., Meselson, M. & Arkhipova, I. R. A deep-branching clade of retrovirus-like retrotransposons in bdelloid rotifers. Gene 390, 136-145 (2007).
21. Gladyshev, E. A. & Arkhipova, I. R. Telomere-associated endonuclease-deficient *Penelope*-like retroelements in diverse eukaryotes. Proceedings of the National Academy of Sciences 104, 9352-9357 (2007).
22. Gregory, T. R. Animal Genome Size Database. http://www.genomesize.com/result_species.php?id=5369 (2012)
23. Mark Welch, J. & Meselson, M. Karyotypes of bdelloid rotifers from three families. Hydrobiologia 387-388, 403-407 (1998).
24. Chevreux, B., Wetter, T. & Suhai, S. Genome sequence assembly using trace signals and additional sequence information. Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB) 99, 45-56 (1999).
25. Aury, J.-M. et al. High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. BMC Genomics 9, 603 (2008).
26. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27, 578-579 (2011).
27. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079 (2009).
28. Gilles, A. et al. Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. BMC Genomics 12, 245 (2011).
29. Huse, S., Huber, J., Morrison, H., Sogin, M. & Welch, D. M. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biology 8, R143 (2007).
30. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nature Methods 9, 357-359 (2012).
31. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841-842 (2010).
32. Denoeud, F. et al. Annotating genomes with massive-scale RNA sequencing. Genome Biology 9, R175 (2008).
33. Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics Chapter 4, Unit 4 10 (2004).
34. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27, 573-80 (1999).
35. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. Bioinformatics 21 Suppl 1, i351-8 (2005).
36. Bairoch, A. et al. The Universal Protein Resource (UniProt). Nucleic Acids Res 33, D154-9 (2005).
37. Birney, E. & Durbin, R. Using GeneWise in the Drosophila annotation experiment. Genome Res 10, 547-8 (2000).
38. Kent, W. J. BLAT--the BLAST-like alignment tool. Genome Res 12, 656-64 (2002).
39. Li, R. et al. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25, 1966-1967 (2009).
40. Mott, R. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. Comput Appl Biosci 13, 477-8 (1997).
41. Korf, I. Gene finding in novel genomes. BMC Bioinformatics 5, 59 (2004).
42. Howe, K. L., Chothia, T. & Durbin, R. GAZE: a generic framework for the integration of gene-prediction data by dynamic programming. Genome Res 12, 1418-27 (2002).
43. Putnam, N. H. et al. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 317, 86-94 (2007).
44. Hellsten, U. et al. The genome of the western clawed frog *Xenopus tropicalis*. Science 328, 633-636 (2010).
45. Jurka, J. et al. Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and Genome Research 110, 462-467 (2005).

46. Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H. Considering transposable element diversification in *de novo* annotation approaches. PLoS ONE 6, e16526 (2011).
47. Li, R. et al. ReAS: recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. PLoS Computational Biology 1, e43 (2005).
48. Arkhipova, I. R. & Meselson, M. Diverse DNA transposons in rotifers of the class Bdelloidea. Proceedings of the National Academy of Sciences of the United States of America 102, 11781-11786 (2005).
49. Gladyshev, E. A. & Arkhipova, I. R. Rotifer rDNA-specific R9 retrotransposable elements generate an exceptionally long target site duplication upon insertion. Gene 448, 145-150 (2009).
50. Gladyshev, E. & Arkhipova, I. A subtelomeric non-LTR retrotransposon *Hebe* in the bdelloid rotifer *Adineta vaga* is subject to inactivation by deletions but not 5' truncations. Mobile DNA 1, 12 (2010).
51. Arkhipova, I. & Meselson, M. Transposable elements in sexual and ancient asexual taxa. Proceedings of the National Academy of Sciences of the United States of America 97, 14473-14477 (2000).
52. Edwards, J. H. The Oxford Grid. Annals of Human Genetics 55, 17-31 (1991).
53. Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Research 40, e49 (2012).
54. Birky, C. W., Jr. Heterozygosity, heteromorphy, and phylogenetic trees in asexual eukaryotes. Genetics 144, 427-437 (1996).
55. Schwartz, S. et al. Human-mouse alignments with BLASTZ. Genome Research 13, 103-107 (2003).
56. Blanchette, M. et al. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Research 14, 708-715 (2004).
57. Easteal, S. & Collet, C. Consistent variation in amino-acid substitution rate, despite uniformity of mutation rate: protein evolution in mammals is not neutral. Molecular Biology and Evolution 11, 643-647 (1994).
58. Chen, J.-M., Cooper, D. N., Chuzhanova, N., Férec, C. & Patrinos, G. P. Gene conversion: mechanisms, evolution and human disease. Nat Rev Genet 8, 762-775 (2007).
59. Smith, C. E., Llorente, B. & Symington, L. S. Template switching during break-induced replication. Nature 447, 102-105 (2007).
60. Malkova, A., Naylor, M. L., Yamaguchi, M., Ira, G. & Haber, J. E. RAD51-dependent break-induced replication differs in kinetics and checkpoint responses from RAD51-mediated gene conversion. Molecular and Cellular Biology 25, 933-944 (2005).
61. LaFave, M. C. & Sekelsky, J. Mitotic recombination: Why? When? How? Where? PLoS Genetics 5, e1000411 (2009).
62. Connallon, T. & Clark, A. G. Gene duplication, gene conversion and the evolution of the Y chromosome. Genetics 186, 277-286 (2010).
63. Higgs, P. G. & Woodcock, G. The accumulation of mutations in asexual populations and the structure of genealogical trees in the presence of selection. Journal of Mathematical Biology 33, 677-702 (1995).
64. Gordo, I. & Charlesworth, B. The degeneration of asexual haploid populations and the speed of Muller's ratchet. Genetics 154, 1379-1387 (2000).
65. Etheridge, A., Pfaffelhuber, P. & Wakolbinger, A. How often does the ratchet click? Facts, heuristics, asymptotics. arXiv 0709.2775 (2007).
66. Rouzine, I. M., Brunet, É. & Wilke, C. O. The traveling-wave approach to asexual evolution: Muller's ratchet and speed of adaptation. Theoretical Population Biology 73, 24-46 (2008).
67. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular Biology and Evolution 3, 418-426 (1986).
68. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Research 13, 2178-2189 (2003).

69. Mirkin, B., Fenner, T., Galperin, M. & Koonin, E. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evolutionary Biology 3, 2 (2003).
70. Gouret, P., Thompson, J. & Pontarotti, P. PhyloPattern: regular expressions to identify complex patterns in phylogenetic trees. BMC Bioinformatics 10, 298 (2009).
71. Levasseur, A. et al. The chordate proteome history database. Evolutionary Bioinformatics 8, 437-447 (2012).
72. Cantarel, B. L. et al. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Research 37, D233-D238 (2009).
73. Altschul, S. F., Gish, W., Miller, W., Meyers, E. W. & Lipman, D. J. Basic local alignment search tool. Journal of Molecular Biology 215, 403-410 (1990).
74. Eddy, S. R. Profile hidden Markov models. Bioinformatics 14, 755-763 (1998).
75. Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. Biological sequence analysis. Probabilistic models of proteins and nucleic acids (Cambridge University Press, 1998).
76. Bateman, A. et al. The Pfam protein families database. Nucleic Acids Research 32, D138-D141 (2004).
77. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688-2690 (2006).
78. Zhong, Y.-f. & Holland, P. W. H. HomeoDB2: functional expansion of a comparative homeobox gene database for evolutionary developmental biology. Evolution & Development 13, 567-568 (2011).
79. Juliano, C., Wang, J. & Lin, H. Uniting germline and stem cells: the function of Piwi proteins and the piRNA pathway in diverse organisms. Annual Review of Genetics 45, 447-469 (2011).
80. Zong, J., Yao, X., Yin, J., Zhang, D. & Ma, H. Evolution of the RNA-dependent RNA polymerase (RdRP) genes: Duplications and possible losses before and after the divergence of major eukaryotic groups. Gene 447, 29-39 (2009).
81. de Jong, D. et al. Multiple *Dicer* genes in the early-diverging Metazoa. Molecular Biology and Evolution 26, 1333-1340 (2009).
82. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32, 1792-1797 (2004).
83. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic Biology 59, 307-321 (2010).
84. Rambaut, A. FigTree: Tree figure drawing tool, version 1.0. (2006)
85. Yigit, E. et al. Analysis of the *C. elegans* Argonaute family reveals that distinct Argonautes act sequentially during RNAi. Cell 127, 747-757 (2006).
86. Librado, P. & Rozas, J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25, 1451-1452 (2009).
87. Cerutti, H. & Casas-Mollano, J. On the origin and functions of RNA-mediated silencing: from protists to man. Current Genetics 50, 81-99 (2006).
88. Piskurek, O. & Okada, N. Poxviruses as possible vectors for horizontal transfer of retroposons from reptiles to mammals. Proceedings of the National Academy of Sciences 104, 12046-12051 (2007).
89. Piskurek, O. & Jackson, D. J. Transposable elements: from DNA parasites to architects of metazoan evolution. Genes 3, 409-422 (2012).