Comparative Genomics of Flatworms (Platyhelminthes) Reveals Shared Genomic Features of Ecto- and Endoparasitic Neodermata

Christoph Hahn[1,*], Bastian Fromm[1], and Lutz Bachmann[1]

[1]Department for Research and Collections, Natural History Museum, University of Oslo, Oslo, Norway

*Corresponding author: E-mail: christoph.hahn@nhm.uio.no.

Accepted: April 7, 2014

Data deposition: The whole genome sequence of *G. salaris* has been deposited at DDBJ/EMBL/GenBank under the accession JJOG00000000 (BioProject ID PRJNA244375). The version described in this paper is version JJOG01000000.

Abstract

The ectoparasitic Monogenea comprise a major part of the obligate parasitic flatworm diversity. Although genomic adaptations to parasitism have been studied in the endoparasitic tapeworms (Cestoda) and flukes (Trematoda), no representative of the Monogenea has been investigated yet. We present the high-quality draft genome of *Gyrodactylus salaris*, an economically important monogenean ectoparasite of wild Atlantic salmon (*Salmo salar*). A total of 15,488 gene models were identified, of which 7,102 were functionally annotated. The controversial phylogenetic relationships within the obligate parasitic Neodermata were resolved in a phylogenomic analysis using 1,719 gene models (alignment length of >500,000 amino acids) for a set of 16 metazoan taxa. The Monogenea were found basal to the Cestoda and Trematoda, which implies ectoparasitism being plesiomorphic within the Neodermata and strongly supports a common origin of complex life cycles. Comparative analysis of seven parasitic flatworm genomes identified shared genomic features for the ecto- and endoparasitic lineages, such as a substantial reduction of the core bilaterian gene complement, including the homeodomain-containing genes, and a loss of the *piwi* and *vasa* genes, which are considered essential for animal development. Furthermore, the shared loss of functional fatty acid biosynthesis pathways and the absence of peroxisomes, the latter organelles presumed ubiquitous in eukaryotes except for parasitic protozoans, were inferred. The draft genome of *G. salaris* opens for future in-depth analyses of pathogenicity and host specificity of poorly characterized *G. salaris* strains, and will enhance studies addressing the genomics of host–parasite interactions and speciation in the highly diverse monogenean flatworms.

Key words: *Gyrodactylus salaris*, draft genome, genomic adaptations, flatworms, parasitism, phylogenomics.

Introduction

Obligate parasitic flatworms constitute one of the three largest groups of metazoan parasites of vertebrates (the others being the nematodes and arthropods) and include many species of medical and veterinary importance. Schistosomes, for example, are responsible for about 300,000 human deaths annually (van der Werf et al. 2003), and a range of tapeworms cause morbidity and mortality in humans and domestic livestock (e.g., *Echinococcus*, *Taenia*, and *Diphyllobothrium*).

The phylogenetic relationships of the phylum Platyhelminthes and the major lineages within have been intensively discussed for decades. Initially, the main challenge was the basal position within Bilateria and the potential paraphyly with acoelomorph flatworms. Recent studies (Baguna and Riutort 2004; Hejnol et al. 2009; Philippe et al. 2011) provided strong evidence for Platyhelminthes and Acoelomorpha being separate phyla, with Acoelomorpha basal within Bilateria. Accordingly, Platyhelminthes would include the remaining free-living and parasitic flatworms and represent derived protostomian lophotrochozoans (Baguna and Riutort 2004).

Genome-wide data are currently only available for a few platyhelminth species including the planarian *Schmidtea mediterranea* (Robb et al. 2008), a model for stem cell biology (Gentile et al. 2011), and some important human parasites, including the flukes *Schistosoma mansoni* (Berriman et al. 2009), *S. japonicum* (*Schistosoma japonicum* Genome Sequencing and Functional Analysis Consortium 2009), and *Clonorchis sinensis* (Wang et al. 2011), and the tapeworms *Echinococcus multilocularis*, *E. granulosus* (Zheng et al. 2013), *Taenia solium*, and *Hymenolepis microstoma* (Tsai et al. 2013). No genome-wide data have yet been published for a

monogenean species. The obligate parasitic flatworms form the monophyletic Neodermata, a well-established lineage based on the name giving Neodermis. This larval secondary syncytial tegument is believed to be the key innovation of Neodermata that allowed for their immense radiation (Littlewood 2006). The Neodermata currently comprise the Monogenea (Monopisthocotylea and Polyopisthocotylea), Cestoda (tapeworms; Eucestoda and Cestodaria), and Trematoda (flukes; Aspidogastrea and Digenea) (reviewed in Olson and Tkach (2005)). The Monogenea are characterized by a direct ectoparasitic lifestyle, whereas the Cestoda and Trematoda engage in endoparasitic life cycles of varying complexity. Traditional views on the evolution of Neodermata based on morphology (Janicki 1920), as well as early molecular studies using 18S ribosomal DNA markers (Littlewood et al. 1999), supported an early divergence of the Trematoda and a sister group relationship of Monogenea and Cestoda (reviewed in Lockyer et al. 2003). More recently, a sister group relationship between Cestoda and Trematoda was supported by nucleotide sequences of the combined 18S and 28S ribosomal genes (Lockyer et al. 2003), mitochondrial DNA (Park et al. 2007; Perkins et al. 2010), and microRNA loci (Fromm et al. 2013).

Understanding the phylogenetic relationships within the Neodermata is an essential prerequisite for understanding the evolutionary origins of endo- and ectoparasitism as well as of the complex life cycles of flatworms. Based on the recently published genomes of the tapeworms *E. multilocularis*, *E. granulosus*, *T. solium*, and *H. microstoma*, Tsai et al. (2013) presented several genomic traits interpreted as adaptations to the parasitic lifestyle in flatworms. Their conclusions, however, must be considered preliminary because no representative of the Monogenea had been included in the analyses, hindering a comprehensive assessment of the issue. This study targets the genome of *G. salaris* Malmberg 1957, a significant pathogen of Atlantic salmon (*Salmo salar*) causing severe ecological and economic damage in Norway and Russia (Harris et al. 2011).

We present the draft genome of *G. salaris* assembled from combined Roche 454 FLX Titanium and Illumina GAII NGS reads, and provide the first genomic reference for a monogenean flatworm. The controversial phylogenetic relationships within Neodermata were addressed using a large-scale phylogenomic approach. Furthermore, recently reported gains and losses of genetic traits as either specific for parasitic flatworms or tapeworms (Tsai et al. 2013) were assessed. With the monogenean draft genome at hand, their significance for the evolution of ecto- and endoparasitism in flatworms is discussed.

Materials and Methods

Sample Collection, DNA Extraction, and Next-Generation Sequencing

Genomic DNA was extracted from a pooled sample of ~15,000 individuals of *G. salaris* obtained from an experimentally reared parasite population on Atlantic Salmon (Salte et al. 2010) using the E.Z.N.A. Tissue DNA kit (Omega Bio-Tek) following the Tissue DNA-Spin Protocol. Library preparation and the Roche 454 FLX Titanium and Illumina GAII sequencing were performed at the Norwegian Sequencing Centre (Oslo, Norway). The Illumina reads were trimmed and end-clipped to a phred score of 33 using custom scripts and subsequently error corrected using the error correction tool of the SOAPdenovo2 software (Luo et al. 2012). For the use in Overlap Layout Consensus (OLC) assemblers, the data set was digitally normalized to a k-mer coverage of 20 using tools from the khmer package (Pell et al. 2012).

De Novo Assembly of *G. salaris* Genomic NGS Reads and Removal of Nontarget Sequences

In order to optimally assemble the draft genome for *G. salaris* a range of de Bruijn graph (DBG) assemblers, that is, Velvet 1.2.07 (Zerbino and Birney 2008), ABySS 1.3.4 (Simpson et al. 2009), SOAPdenovo 1.0.5 (Li et al. 2010b) were utilized on the Illumina reads. Furthermore, hybrid assemblies of 454 and normalized Illumina GAII data were performed using the OLC assemblers Newbler 2.6 (Margulies et al. 2005) and Celera 7.0 (Myers et al. 2000). Sequence assemblies obtained by DBG and OLC approaches were quality assessed using standard assembly metrics such as N50, total number of contigs, and total length of the assembly (supplementary file S2, table S1, Supplementary Material online). Guanine–Cytosine (GC)-coverage scatter plots (Kumar and Blaxter 2011) were produced for the assemblies (fig. 1 and supplementary file S1, fig. S1, Supplementary Material online) and the conjunction of all information was used to choose the best assembly (see supplementary file S1, Section A for details, Supplementary Material online). Putative nontarget contigs were removed prior to subsequent analyses (supplementary file S1, Section A for details, Supplementary Material online). Completeness of the gene space in the draft was assessed using CEGMA 2.3 (Parra et al. 2007, 2009). Trimmed Illumina reads were mapped back to the conservative draft assembly using BWA (Li and Durbin 2009) in order to identify the putative *G. salaris* readpool. K-mer frequencies (20-mer) were subsequently calculated for this readpool using respective tools from the khmer package (Pell et al. 2012) in order to estimate the *G. salaris* genome size based on the k-mer frequency distribution (Li et al. 2010a) (see supplementary file S1, Section A, Supplementary Material online). Single nucleotide polymorphisms (SNPs) were called using Freebayes (Garrison and Marth 2012) and SNP density and transition/transversion ratio were calculated using VCFtools (Danecek et al. 2011). To minimize false positives, SNPs called at positions with low ($<20\times$) and high coverage ($>100\times$) were not considered.

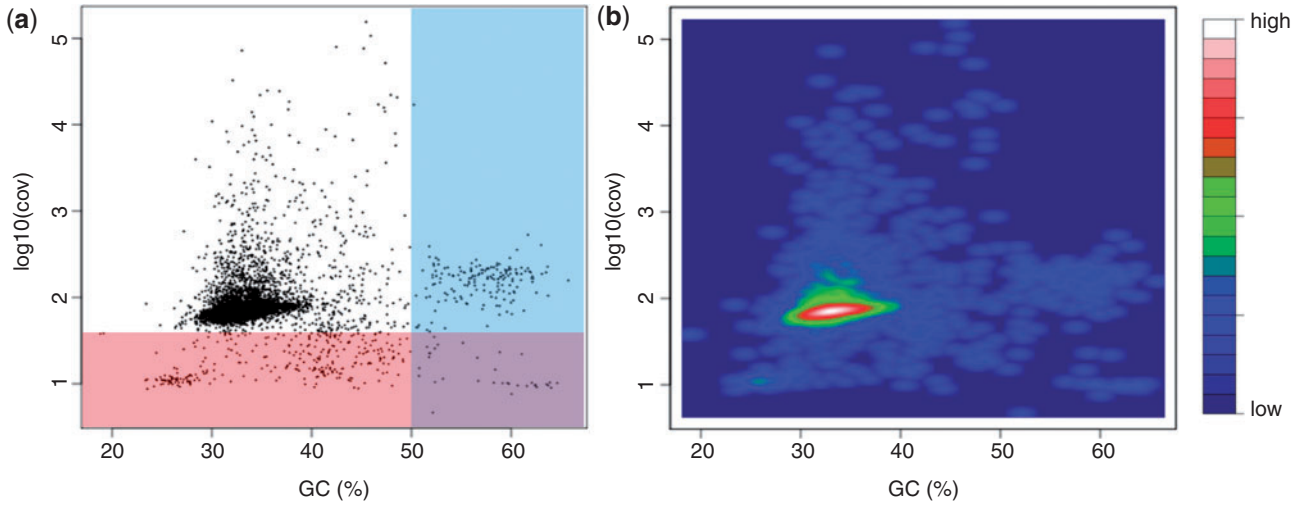

Fig. 1.—Aggregate properties (coverage and GC content) of contigs obtained by Celera 7.0 as (*a*) scatter plot and (*b*) heat map. Colored areas in (*a*) illustrate the aggressive cleaning strategy adopted in this study, that is, the removal of putative nontarget contigs of coverage <40× (red; putative host contamination) and GC content >50% (blue; putative bacterial contamination) (see supplementary file S1, Section A for details, Supplementary Material online).

Gene Prediction and Annotation

A *G. salaris* specific library of repeat families, including simple repeats, interspersed repeats, and satellite DNAs was identified from the assembled draft genome using Repeat Modeler (Smit and Hubley 2008–2010). Subsequently, the automated annotation pipeline MAKER 2.26 (Holt and Yandell 2011) was used for structural annotation (supplementary file S1, Section B for details, Supplementary Material online). Repetitive DNA was masked using RepeatMasker (Smit et al. 1996–2010) and the *G. salaris* specific repeat library. MAKER then reconciled homology-based physical evidence with the results of purely ab initio predicted gene models obtained by means of the gene predictors GeneMark (Lomsadze et al. 2005), SNAP (Korf 2004), and Augustus (Stanke and Waack 2003) (see supplementary file S1, Section B for further details, Supplementary Material online). All gene models were screened for known domains using InterProScan5RC6 (Quevillon et al. 2005) and subjected to a BLAST search (Altschul et al. 1997) against a custom build of NCBI's nr protein database (restricted to Metazoan proteins). B2G4pipe 2.5 (Conesa et al. 2005) was used to establish functional annotations, that is, to assign putative biological functions and Gene Ontology terms (Ashburner et al. 2000) to the *G. salaris* gene models. The initial functional annotations were reconciled with the InterProScan results and mapped onto metabolic pathways included in the KEGG database (Kanehisa et al. 2012) using blast2GO v. 2.6.6 (Conesa et al. 2005; Gotz et al. 2008). For comparative reasons, the same procedure was also performed for the gene sets of *S. mediterranea*, *E. multilocularis*, *S. mansoni*, *Caenorhabditis elegans*, *Lottia gigantea*, and *Helobdella robusta*.

Analyses of Gene Models

Basic statistics for obtained *G. salaris* gene models, including average intron/exon lengths and number of introns, were calculated using custom scripts. The functionally annotated gene models were screened for the presence/absence of selected genomic traits that were specifically addressed in a recent study on cestode genomes (Tsai et al. 2013) (see supplementary file S1, Section D for details, Supplementary Material online).

Gene Orthology Search, Cluster Selection, and Phylogenomic Analyses

The protein complements of 16 metazoan taxa (supplementary file S2, table S2, Supplementary Material online) were searched for reciprocal best hits using BLAST (Altschul et al. 1997), and the results were subsequently used for identifying orthologous gene clusters by using the MCL algorithm (van Dongen 2000) following Fischer et al. (2011). An inflation parameter of 2.1 was used as previously described (Smith et al. 2011). For subsequent phylogenetic analyses of metazoan relationships, a set of putative orthologous gene clusters was selected based on criteria used in previous studies (Dunn et al. 2008; Hejnol et al. 2009) with minor modifications: Clusters of putative orthologous genes were required to 1) include at least eight taxa (i.e., ≥50%), 2) include at least one taxon from each of the four platyhelminth groups, 3) have a mean of <5 sequences per taxon, and 4) have a median of <2 sequences per taxon. Sequences of each accepted cluster were aligned using Clustal Omega (Sievers et al. 2011). Ambiguously aligned regions were removed using

Aliscore and Alicut (Kück 2009; Misof and Misof 2009). Maximum-likelihood (ML) phylogenetic trees were inferred individually for each gene cluster with RAxML 7.3.2 (Stamatakis 2006) using the best-fitting amino acid substitution model as determined by the RAxML amino acid substitution model selection Perl script. Individual tree robustness was evaluated with 100 bootstrap pseudo replicates. Clusters of putative orthologous genes were then divided into classes 1) containing no paralogs, 2) containing only monophyletic paralogs (i.e., putative in-paralogs), or 3) containing paraphyletic paralogs (i.e., putative out-paralogs), using custom scripts. The latter clusters were omitted from further analyses. Clusters containing in-paralogs were subjected to monophyly masking, that is, all but one representative paralog per taxon were removed from the alignment. For each such cluster, multiple alignments including only one paralog at a time were constructed using Clustal Omega (Sievers et al. 2011) and trimmed using Aliscore and Alicut (Kück 2009; Misof and Misof 2009). Custom scripts were used to assess all alignments and select the final alignment for each cluster using only the particular paralog that yielded the longest alignment after trimming, that is, caused the smallest number of ambiguously aligned positions in the alignment. If several paralogs ranked equally, one was chosen randomly. Clusters without paralogs (Class 1) and the monophyly masked clusters (Class 2) were concatenated to a supermatrix using FASconCAT (Kuck and Meusemann 2010), after removing any clusters yielding alignments shorter than 100 amino acid positions. ML analyses were performed on the final matrix using RAxML 7.8.3 (Stamatakis 2006). The data set was partitioned according to the individual genes and the best-fitting amino acid substitution model, as determined by the RAxML amino acid substitution model selection Perl script, was applied to each partition. ML tree robustness was assessed using 100 bootstrap pseudo replicates. Bayesian inference (BI) was performed using the CAT-GTR model (Lartillot and Philippe 2004) as implemented in Phylobayes-MPI 1.3b (Lartillot et al. 2013). BI was run for 2,000 generations in five parallel chains. The initial 500 generations were discarded as burn in, well after convergence (MaxDiff values below 0.3). A 50% consensus tree was computed from the remaining 1,500 trees from each chain. BI tree robustness was assessed using posterior probabilities. In addition, ML and BI were inferred for a concatenated alignment of a reduced set of genes with the strongest phylogenetic signal (average bootstrap support $\geq$80). This method was recently proposed to reduce potential incongruence in phylogenomic analyses (Salichos and Rokas 2013). The same study emphasized that bootstrap values can be misleading in large phylogenomic data sets and introduced internode certainty (IC) as a more sensitive metric for the detection of incongruence. However, only genes that are represented in all taxa can currently be taken into account for the calculation of IC. For the above described data set of 16 metazoan taxa, IC would be based on <50% of the genes. To overcome this limitation, we compiled a further data set targeting exclusively lophotrochozoan taxa. Clusters of orthologous genes were required to 1) include all 10 lophotrochozoan taxa, 2) have a mean of <5 sequences per taxon, and 3) have a median of <2 sequences per taxon. Nonlophotrochozoan taxa were removed and individual alignments were subsequently processed (aligned, trimmed, and monophyly masked) as described above. The data set was finally reduced to exclusively genes with strongest phylogenetic signal (in this case an average bootstrap support threshold of 90 was feasible), and phylogenetic analyses were performed as described above. IC was inferred using RAxML 8.0.3 (Stamatakis 2006). See supplementary file S1, Section C for more details, Supplementary Material online. The scripts used for the preparation of data were made freely available (https://github.com/chrishah/phylog/, last accessed April 30, 2014).

Results

Genome Sequencing and Gene Prediction

Five de novo assemblers were tested and the result obtained by Celera 7.0 (Myers et al. 2000) was selected as the best conservative draft assembly for *G. salaris* (fig. 1) based on basic assembly metrics (e.g., N50) and the inferred robustness against putative nontarget reads (see supplementary file S1, Section A, and supplementary file S2, table S1 for details, Supplementary Material online). Furthermore, unlike other assemblers Celera 7.0 used both Roche 454 FLX Titanium and Illumina GAII reads in a simultaneous hybrid assembly. The shotgun sequencing strategy combined 0.6 Gb Roche 454 FLX Titanium (average read length 514 bp) and 19.8 Gb Illumina GAII (read length 76 bp; paired end insert size 167 $\pm$ 40 bp) reads. After removal of putative nontarget contigs, the draft genome comprised 6,075 contigs (>200 bp; N50 18.4 kb; average coverage ~126×) totaling 67.4 Mb (GC content 33.84%). The gene space was estimated to be >89.11% complete using CEGMA (Parra et al. 2007). The two most conservative CEGMA reference gene sets were recovered at a rate of >95% in the draft genome of *G. salaris*. A total of 25.68% of the assembly was identified as repetitive, and the genome size was estimated as ~120 Mb (see supplementary file S1, Section A, Supplementary Material online). The assembled sequences of the *G. salaris* draft genome were deposited in GenBank (BioProject ID PRJNA244375) and can furthermore be obtained from the *Gyrodactylus salaris* genome project's web page (http://invitro.titan.uio.no/gyrodactylus/, last accessed April 30, 2014). We identified a total of 93,185 SNPs (SNP density 1.55 SNPs/kb) with a transition/transversion ratio of 0.92. Structural gene annotation produced 15,488 gene models, of which 8,637 contained functional domains. Functional annotation was achieved for 7,102 gene models (supplementary file S1, Section D and fig. S4, Supplementary Material online). The average *G. salaris* gene

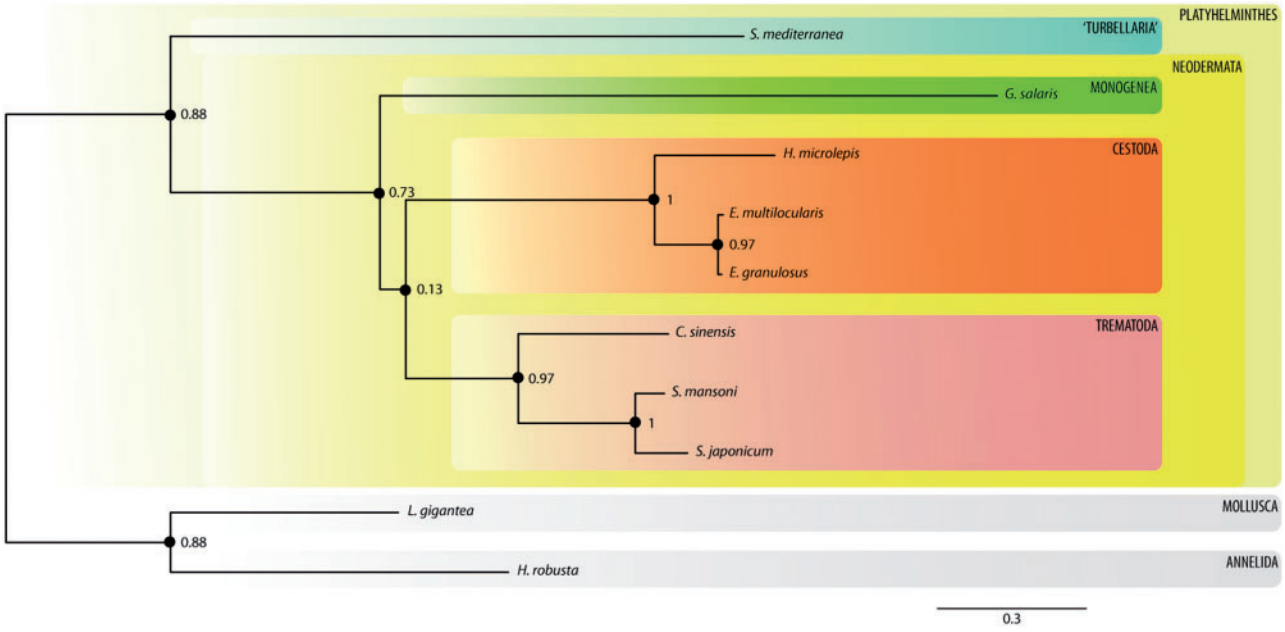

Fig. 2.—Phylogenetic relationships among the major groups of parasitic flatworms. BI topology (CAT-GTR model) based on a supermatrix containing the 312 gene models (135,085 amino acid positions) with the strongest phylogenetic signal (average bootstrap ≥90). Full circles represent nodes with posterior probability = 1 and boostrap support of 100. Numbers next to internodes represent IC, and are based on a subset of 311 genes available for the full taxon set.

model was 2.7 kb long, contained four exons with an average length of 289-bp coding for a 906-bp transcript. Overall, the average intron length was 659 bp. However, when looking at introns according to their position within genes, it turned out that the average intron length decreased significantly (Pearson's product-moment correlation $\rho = -0.88$, $P = 8.19 \times 10^{-9}$) with increasing ordinal position (see supplementary file S1, fig. S11, Supplementary Material online).

Phylogenomic Analysis

The orthology search for the metazoan data set identified 38,568 orthologous gene clusters, out of which 3,121 passed the selection criteria (see Materials and Methods section). A total of 548 orthologous gene clusters contained no paralogs, and 1,303 clusters contained only monophyletic paralogs. The supermatrix for phylogenomic analyses of the metazoan data set contained 1,719 gene clusters ≥100 amino acids (aa), and comprised a total of 519,023 aa positions (supplementary files S3 and S2, table S4, Supplementary Material online). Within the Platyhelminthes, the monophyly of the Neodermata with a basal position of the ectoparasitic Monogenea and a sister group relationship of the endoparasitic Trematoda and Cestoda was highly supported in all analyses (fig. 2). With respect to deep metazoan relationships, the phylogenetic tree obtained using BI (supplementary fig. S6, Supplementary Material online) is in agreement with previously published results and recovers the monophyly of Chordata, Ecdysozoa, Lophotrochozoa, and Platyhelminthes with high statistical support. The phylogenomic reconstructions using a subset of the metazoan data set, that is, including exclusively the 173 gene models with the strongest phylogenetic signal (average bootstrap support of gene trees ≥80; supplementary file S2, table S5 and supplementary file S4, Supplementary Material online) consistently recovered the same tree topology (see supplementary file S1, fig. S5*b* for ML and fig. S6 for BI, Supplementary Material online). The ML tree for the full metazoan data set differed from the BI tree in the position of the nematode *C. elegans*, which was recovered in a sister group relationship to the Platyhelminthes. This is likely an artifact caused by long branch attraction (Felsenstein 1987; see supplementary file S1, Section C for more detailed discussion, Supplementary Material online). The lophotrochozoan data set of genes with strongest phylogenetic signal (average bootstrap support ≥90) comprised a total of 312 genes and 135,085 aa positions (supplementary files S5 and S2, table S6, Supplementary Material online). Both BI and ML analyses of this data set strongly supported a basal position of the Monogenea and a sister relationship of Cestoda and Trematoda (fig. 2). IC values are based on 311 individual gene trees.

Genome-Wide Reduction of Bilaterian Genes

We identified a total of 1,490 orthologous gene clusters, which contained representatives of all 15 bilaterian taxa

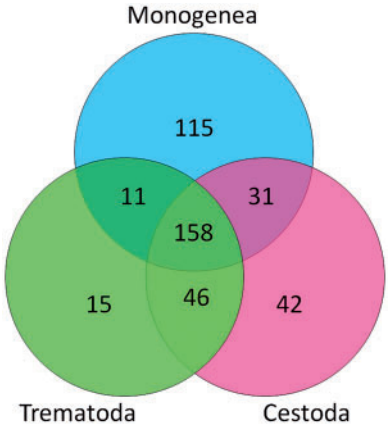

Fig. 3.—Venn diagram illustrating the number of gene clusters devoid of orthologous genes of one to three parasitic flatworm groups, while present in all other nonneodermatan Bilateria in the data set.

(supplementary file S2, table S2, Supplementary Material online). A total of 315 clusters were found to lack *G. salaris* genes, while containing genes of each of the eight nonneodermatan bilaterians. Similarly, 277 and 230 gene clusters were lacking gene models of Cestoda and Trematoda, respectively. Integrating these results indicates that 158 bilaterian core genes have been lost throughout the Neodermata (fig. 3). For the other bilaterians included in the data set much lower numbers of lineage-specific losses were observed, that is, 23 in *L. gigantea* (Mollusca), 36 in *H. robusta* (Annelida), 99 in *S. mediterranea* (Turbellaria), 110 in *C. elegans* (Nematoda), 46 in *Daphnia pulex* (Arthropoda), and 8 in *Homo sapiens* (Chordata).

Specific Genomic Features in Parasitic Flatworms

The gene complement of *G. salaris* was analyzed with respect to recently proposed genomic adaptations to parasitism in flatworms (Tsai et al. 2013). A total of 55 DEAD/DEAH box helicase containing gene models were identified for *G. salaris*. However, the ubiquitous stem-cell-specific RNA helicase *vasa* (DDX4) was absent. One gene model was found orthologous to the structurally very similar *PL10* (DDX3) gene, and clustered within a clade consisting exclusively of parasitic flatworm *PL10* (supplementary file S1, fig. S8, Supplementary Material online). The *G. salaris* homeodomain-containing gene complement comprised 79 gene models, which could be assigned to 52 families within nine classes (supplementary file S1, fig. S7 and supplementary file S2, table S7, Supplementary Material online). All classes reported earlier for tapeworms and flukes were represented, and 23 of the 24 homeodomain genes previously described as lost in flukes and tapeworms (Tsai et al. 2013) were absent in *G. salaris* as well; however, *Vsx* was represented in the draft genome. A further 10 losses were detected in *G. salaris*, three of them shared with tapeworms (*Gbx*, *Bari*, and *Rax*), four with flukes (*Mkx*, *Zeb*, *Cux*, and *Hox3*), and three appeared specific to the monogenean lineage (*Meox*, *En*, and *Pou2*) (fig. 4). The *piwi* subfamily was not detected in the *G. salaris* genome. However, two argonaute gene models were found in the *G. salaris* gene complement, which clustered within the *AGO* family and a further clade comprising exclusively flatworm sequences, respectively (supplementary file S1, fig. S9, Supplementary Material online).

The intron length distribution of *G. salaris* gene models had two prominent modes at ~32 bp (subsequently referred to as short mode) and ~58 bp (long mode) (supplementary file S1, fig. S10 compares the intron length distributions of four flatworm taxa, Supplementary Material online). A total of 14,342 introns (31%) were shorter than 100 bp, whereas the remaining 31,312 introns (69%) exceeded 100 bp (ranging up to ~20,000 bp in length), which resulted in an average intron length of 659 bp. The ratio of the occurrence of introns of the two modes changed with their ordinal position within the genes. Ordinal intron positions 1 and 2 were dominated by short mode introns, whereas the long mode introns dominated at ordinal positions >3 (supplementary file S1, fig. S12, Supplementary Material online). *Echinococcus multilocularis* orthologs of *G. salaris* genes, which contained introns falling into either the short or the long length mode, were more likely themselves to contain introns of the corresponding short/long mode in *E. multilocularis* than expected by chance (see supplementary file S1, Section D and supplementary file S2, table S8, Supplementary Material online).

In total, eight heat shock protein (*Hsp*) 70 gene models were identified in the *G. salaris* draft genome. All *Hsp70* subfamilies previously reported for flatworms (Tsai et al. 2013), that is, *Hsp110*, ER *Hsp70*, mitochondrial *Hsp70*, and flatworm cytosolic *Hsp70* were present, and each subfamily was represented by one to two gene copies (supplementary file S1, fig. S16, Supplementary Material online). None of the *Hsp70* genes detected for *G. salaris* clustered with tapeworm-specific *Hsp70* genes reported by Tsai et al. (2013).

The four peroxisome marker proteins *PEX3*, *PEX10*, *PEX12*, and *PEX19*, proposed earlier as unequivocal in silico indicators for the presence of peroxisomes (Schluter et al. 2006), were not detected in the *G. salaris* draft genome, nor in the gene complement of *S. mansoni* or *E. multilocularis*. However, we detected indication of all four marker proteins in the genomes of the free-living flatworm *S. mediterranea*, as well as the mollusc *L. gigantea*, the annelid *H. robusta*, and the nematode *C. elegans*.

The majority of genes encoding key enzymes necessary for fatty acid synthesis was absent from the *G. salaris* gene complement. Only the acetyl-CoA carboxylase, involved in fatty acid precursor production, could be detected (supplementary file S1, fig. S13, Supplementary Material online). The enzyme complement of fatty acid metabolism and -elongation, however, was found to be largely complete (supplementary file S1, figs. S14 and S15, Supplementary Material online).

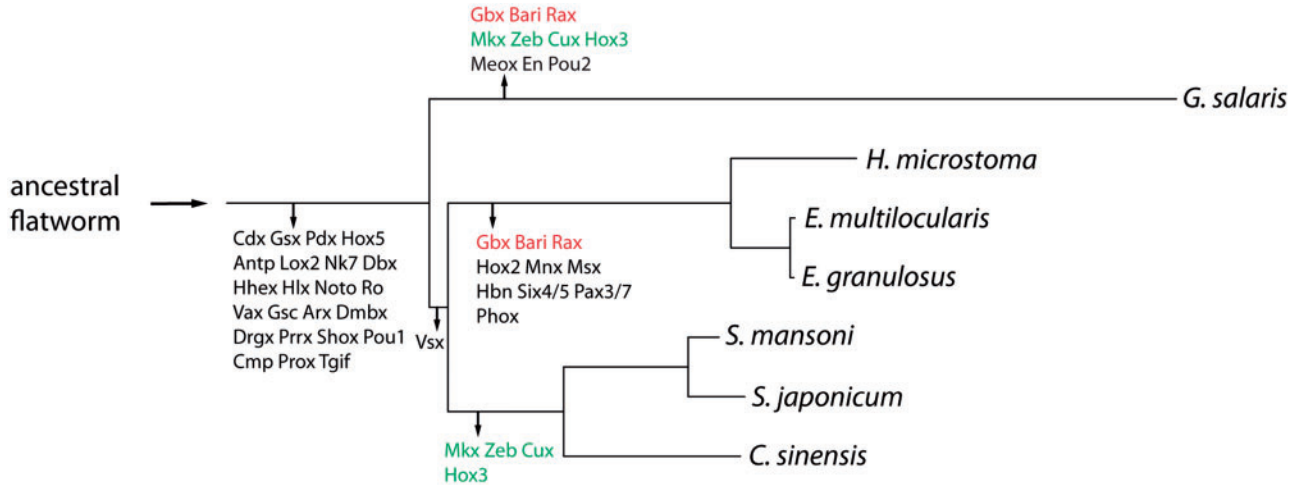

Fig. 4.—Pattern of inferred homeobox gene family loss during the evolution of obligate parasitism in flatworms. Putative convergent losses are indicated in red (Monogenea and Cestoda) and green (Monogenea and Trematoda).

Furthermore, a range of fatty acid binding and -transport proteins were identified (supplementary file S2, table S12, Supplementary Material online).

Discussion

In recent years, several genomes of platyhelminths have become available. Species causing severe human diseases, such as the blood flukes *S. mansoni* (Berriman et al. 2009) and *S. japonicum* (*Schistosoma japonicum* Genome Sequencing and Functional Analysis Consortium 2009), were the first to be studied at a genomic level, and recently four tapeworm genomes of public health interest were published (Tsai et al. 2013; Zheng et al. 2013), forming the basis for comparative genomic analyses in order to identify specific evolutionary adaptations to the obligate parasitic lifestyle in flatworms. Accordingly, Tsai et al. (2013) proposed a number of genomic features for parasitic flatworms; however, their conclusions must be considered preliminary because the phylogenetic relationships within Neodermata were not resolved, and there was no representative of the Monogenea, the third major parasitic group, included in their comparative genomic analysis.

The *G. salaris* Genome

The size of the *G. salaris* genome, calculated as ~120 Mb, was found in the same order of magnitude as other platyhelminth genomes, including the sequenced tapeworms, which range between 115 and 151 Mb (Tsai et al. 2013; Zheng et al. 2013). Using shallow sequencing data, the genome size of the closely related sibling species *Gyrodactylus thymalli* was recently estimated at ~170 Mb (Hahn et al. 2013). This estimate, however, can only be interpreted as an upper limit because the authors had only little information on the ratio of host and other nontarget DNA in the analyzed total DNA extract. Although genome assembly for metazoans with relatively small genome size has improved greatly over the past years to the point where it has become a democratized, bottom-up enterprise (Kumar et al. 2012), a number of challenges remain for small invertebrates. The very small amounts of genomic DNA that can be extracted from single individuals may be insufficient for the needs of some NGS applications. Low DNA yield is also an issue for *Gyrodactylus* as individual parasites contain only about 1,000 cells (Bakke et al. 2007). Whole-genome amplification (WGA) approaches may overcome this obstacle (Rodrigue et al. 2009), but in turn may imply other problems such as amplification of nontarget DNA, which can confuse genome assembly. Furthermore, bias in the WGA may leave large regions of the genome unsequenced (Rodrigue et al. 2009). In order to provide enough genomic DNA for NGS sequencing of *G. salaris*—in particular for the Roche 454 FLX Titanium pyrosequencing—we chose the strategy of extracting DNA from a pooled sample. The sampled population was reared experimentally under controlled conditions (Salte et al. 2010), and genetic heterozygosity was expected to be low. Indeed, the observed SNP density was less than half of the recently described 1/288 bp (3.47 SNPs/kb) for *S. japonicum* (Liu et al. 2006). As *Gyrodactylus* feeds on host mucus and epithelial cells approximately every 15–30 min (Cable et al. 2002), host cells and bacteria of the host skin epiflora will always be present in the gut of the parasites. The inevitable coextraction of this nontarget DNA further complicates genome assembly of *G. salaris*. As the host genome is about one order of magnitude larger (estimated to 3×10^9 bp (Davidson et al. 2010)) than that of *G. salaris*, even minute amounts of host cells in the total DNA sample could lead to significant contamination on the NGS read level. To overcome this challenge, we applied an in silico

cleaning of the NGS data downstream to sequencing. We adopted a recently proposed strategy for the separation of genomes of symbiotic organisms (Kumar and Blaxter 2011) that is based on differing aggregate properties of parasite and nontarget genomes. The bioinformatic cleaning of the *G. salaris* assembly was performed aggressively to minimize nontarget contigs in the draft genome data set, and we consider the removal of ~50% (see supplementary file S1, Section A, Supplementary Material online) of all reads appropriate in the light of the potential sample contamination. An assembly N50 of around the median gene length of an organism is sufficient to expect about 50% of all genes to be uninterruptedly present in a draft (Yandell and Ence 2012)—a simple rule of thumb suggests the median gene length in the genome of *G. salaris* (estimated as ~120 Mb) to be about 2.5 kb (interpolated after fig. 1 in Yandell and Ence (2012)). The actual N50 of >18 kb thus suggests an uninterrupted representation of the majority of genes. The gene space in the draft genome of *G. salaris* was assessed as being roughly 90% complete based on the presence of a set of core genes of varying level of evolutionary conservatism (Parra et al. 2007, 2009). However, it is very likely that the gene space is even more complete because CEGMA tends to underestimate the completeness of divergent genomes. It has thus been suggested earlier that only the most conservative groups of genes should be used as indicators for completeness in such cases (Parra et al. 2009). In line with this reasoning, the two most conservative CEGMA reference gene sets were recovered at a rate of >95% in the draft genome of *G. salaris*. The 15,488 predicted gene models for *G. salaris* were well within the range reported for other parasitic flatworms, for example, 10,345 for *E. multilocularis* (Tsai et al. 2013), 11,325 for *E. granulosus* (Zheng et al. 2013), 10,852 for *S. mansoni* (Berriman et al. 2009), and 16,258 for *C. sinensis* (Wang et al. 2011). Although the published information on general gene statistics for previously sequenced flatworms is slightly ambiguous and rather fragmentary, the results for *G. salaris* gene models indicated a reduction in average gene length and number of exons per gene when compared with other parasitic flatworms (supplementary file S2, table S3, Supplementary Material online). The average intron size calculated for *G. salaris* was similar to that previously reported for tapeworms, whereas average protein- and exon length appeared similar across flatworms in general. The prediction of gene models for *G. salaris* was based on the information available in public databases (see supplementary file S2, table S2, Supplementary Material online) combined with ab initio gene prediction. Although purely ab initio gene predictors are prone to overprediction, MAKER2 appeared relatively robust in respect to false-positive predictions (Holt and Yandell 2011). Whether the set of genes predicted for *G. salaris* is indeed inflated remains to be clarified by means of transcriptome sequencing and expression studies. However, the largely concordant gene statistics when compared with other parasitic flatworms (see above), and the identification of functional domains and functional annotation in 55.8% and 45.9% of the predicted gene models, respectively, imply the inference of a high-quality set of gene models for *G. salaris*. In comparison, in the recently released genome of *E. granulosus* 40.3% of the gene models could be functionally annotated (Zheng et al. 2013).

Phylogenomics of Platyhelminthes

The phylum Platyhelminthes comprises the free-living Turbellaria and the obligate parasitic Monogenea, Cestoda, and Trematoda, the latter three groups forming the Neodermata. Although there was little dispute about the monophyly of Neodermata in the phylum, the phylogenetic relationships within the Neodermata have been discussed for many years. The phylogenomic analyses unambiguously supported the monophyly of the Platyhelminthes and the obligate parasitic Neodermata with high statistical support. This study provides the first genome wide approach to assess the interrelationships of the obligate parasitic flatworms, and within the monophyletic Neodermata all analyses consistently recovered an early divergence of the ectoparasitic Monogenea and a sister group relationship of the endoparasitic Trematoda and Cestoda (more detailed discussion of the IC values can be found in supplementary file S1, Section C, Supplementary Material online). The data therefore soundly reject a sister group relationship between Monogenea and Cestoda, previously united as the Cercomeromorphae (Janicki 1920), and support recent results obtained by analyzing combined ribosomal 18S and 28S data sets (Lockyer et al. 2003), mitochondrial genomes (Park et al. 2007; Perkins et al. 2010), and microRNA complements (Fromm et al. 2013). We cannot comment on a possible paraphyly of the Monogenea in respect to the two main lineages within, the Monopisthocotylea and Polyopisthocotylea (Mollaret et al. 1997; Justine 1998). Nevertheless, the draft genome of the monopisthocotylean *G. salaris* provides a paramount foundation for future studies targeting this question. The basal position of the Monogenea within the Neodermata implies that the endoparasitism of Cestoda and Trematoda is a monophyletic but derived lifestyle within Platyhelminthes and that the ectoparasitism in Monogenea represents the plesiomorphic state of parasitism in Neodermata. As pointed out by Park et al. (2007), a sister group relationship between Cestoda and Trematoda also suggests that the transition from direct life cycles on a single host (as in Monogenea) to complex life cycles with multiple hosts (as in Cestoda and Trematoda) was a single evolutionary event.

Genomic Features of Obligate Parasitism in Flatworms

With the phylogenetic relationships within Neodermata resolved the draft genome of *G. salaris* allows for assessing whether recently proposed adaptive genomic traits for flukes and tapeworms (Tsai et al. 2013) are related to the

evolution of endoparasitism or represent in fact synapomorphies of the Neodermata. The list of the adaptive genomic features included among others 1) the absence of *vasa*, 2) a substantially reduced homeobox gene complement, 3) the loss of the *piwi* subfamily and the emergence of a fluke- and tapeworm-specific subfamily of argonaute proteins, 4) the innovation of a bimodal intron length distribution in tapeworms, 5) an expansion of the *Hsp70* family in tapeworms, 6) the absence of essential peroxisome enzymes, and possibly the absence of peroxisomes in general, and 7) the lack of key enzymes for fatty acids synthesis.

The ubiquitous stem-cell specific RNA helicase *vasa* (DDX4) was absent in tapeworms and flukes. Instead two copies of the structurally very similar *PL10* (DDX3) gene were found in both lineages, and it was speculated that a duplication of *PL10* (DDX3) in free-living flatworms facilitated a functional takeover of *vasa* by *PL10*, followed by a subsequent loss of the then functionally redundant *vasa* gene in the common ancestor of tapeworms and flukes (Tsai et al. 2013). Both *vasa* as well as the parasite-specific *PL10* have been reported for the monogenean *Neobenedenia girellae* (Ohashi et al. 2007), an ectoparasite of marine fish, which, like *G. salaris*, belongs to the Monopisthocotylea. However, in *G. salaris* we found only one copy of the parasitic flatworm-specific *PL10* protein. *Vasa* was absent, and the loss of the gene is thus not a specific feature of tapeworms and flukes. Whether *PL10* has functionally taken over for *vasa* in *G. salaris* remains to be tested, but it is noteworthy that knock-down experiments in *N. girellae* indicated a more important role of *PL10* in germ cell formation and fertilization than *vasa* (Ohashi et al. 2007). The reason why two monopisthocotylean monogeneans differ in this respect is unclear but may reflect the suggested paraphyly of the group (Perkins et al. 2010).

The homeodomain complement of invertebrates is very variable. Recently substantial expansions were reported in the lophotrochozoans *H. robusta* (Annelida, Hirudinea), *Capitella telata* (Annelida, Polychaeta), and *L. gigantea* (Mollusca, Gastropoda) with 181, 121, and 111 homeodomain-containing proteins, respectively (Simakov et al. 2013). In contrast, a shared loss of 24 homeobox gene families in flukes and tapeworms, and an additional loss of 10 further families unique to tapeworms, when comparing the respective homeodomain complements to a set of 96 homeobox gene families expected for the most recent common ancestor of the Bilateria, were reported (Tsai et al. 2013). A similar reduction in the homeodomain-containing gene complement was confirmed for *G. salaris*. The data imply that the majority of homeodomain gene losses must have occurred early in the evolution of Neodermata. There was, however, also an indication of a partly convergent history of homeobox gene family losses in the three neodermatan lineages (fig. 4). Although direct experimental evidence is still lacking, the substantial loss of homeodomain genes may be an adaptation to an obligate parasitic lifestyle.

The *piwi* gene subfamily was also reported lost in cestodes and trematodes. Along with the gain of a specific argonaute subfamily, these have been proposed as tapeworm- and fluke-specific adaptations to parasitism (referred to as group 4 argonaute clade by Tsai et al. 2013). However, the *piwi* gene subfamily was absent also from the draft genome of *G. salaris* and we conclude therefore that the loss of this particular gene subfamily is a further synapomorphy of Neodermata. In addition, genes belonging to the fluke- and tapeworm-specific group 4 argonaute clade (Tsai et al. 2013) were also present in the genome of *G. salaris*. As reported earlier (Zheng 2013), we confirmed a *S. mediterranea* argonaute protein clustering within this clade (supplementary file S1, fig. S9, Supplementary Material online), and thus conclude that this represents a flatworm-specific expansion, rather than a phenomenon restricted to endoparasitic neodermatans. The loss of *piwi* in Neodermata raises of course many questions with respect to the gene's canonical functions. *Piwi* and argonaute proteins are essential for maintaining genome integrity in stem cells in a process referred to as *piwi*-interacting RNA (piRNA) pathway; in short they repress the mobilization of transposable elements (TEs). The question how flatworms keep up genome integrity without *piwi* and *vasa* was addressed recently (Skinner et al. 2014). The authors suggested that parasitic platyhelminths may have evolved a germline, stem cell-specific endogenous, noncanonical siRNA pathway for TE silencing that took over the piRNA pathway (Skinner et al. 2014).

Intron lengths in eukaryotes usually vary within a broad range. However, the length distribution is frequently characterized by at least one distinct mode commonly below 100 bp: These putative optima for minimal intron length vary between species. Despite the occurrence of such microintron modes, the average intron length is usually substantially higher, for example, 663 in *E. multilocularis* (Tsai et al. 2013), 1,411 in *Drosophila melanogaster*, and 372 in *C. elegans* (Yu et al. 2002). Yu et al. (2002) argued that such minimal introns may enhance the rate at which messenger RNA is exported from the cell nucleus. A bimodal intron length distribution was earlier suggested as an evolutionary innovation, that is, a newly gained genomic trait, in tapeworms (Tsai et al. 2013), and the authors speculated about modifications of the splicing machinery and/or the nucleus export machinery as being causal adaptations for this phenomenon. However, a bimodal distribution was also found in *G. salaris* with modes at ~32 and ~58 bp, respectively (supplementary file S1, Section D and fig. S10, Supplementary Material online), indicating that this feature is more common than previously suggested.

Although the mode positions vary between the species, we found evidence for congruence between *G. salaris* and *E. multilocularis* gene models in respect to the gene composition in their modes. *Gyrodactylus salaris* gene models containing introns falling either into the short (2,232) or the long intron length modes (2,199) were more likely to be

orthologous to *E. multilocularis* gene models containing the corresponding short/long mode introns (1,793 and 2,156, respectively) than expected if gene models were randomly distributed in the two species (see supplementary file S1, Section D for more details, Supplementary Material online). In *G. salaries*, we furthermore found remarkable differences in the number of short/long mode introns depending on their ordinal position within the gene models. Although the short mode introns are most prevalent at intron positions 1 and 2 of the gene models, the long mode introns are more evenly distributed with respect to ordinal position, but become dominant in introns of ordinal position ≥3 (see supplementary file S1, fig. S12, Supplementary Material online). However, average intron length in *G. salaris* gene models was found to significantly decrease with ordinal position (see supplementary file S1, fig. S11, Supplementary Material online). Such a negative correlation between intron length and intron ordinal position was reported earlier for a range of organisms (Marais et al. 2005; Gazave et al. 2007; Zhang and Edwards 2012). For primates and *Drosophila*, it has furthermore been reported that levels of evolutionary constraint vary across classes of introns of different length. Although longer introns appeared more divergent in primates (Gazave et al. 2007), there was a negative correlation between intron length and the level of divergence in *Drosophila* (Haddrill et al. 2005). Whether or not intron length in *G. salaris* is correlated with evolutionary constraints needs to be addressed in future studies. Such more comprehensive studies may also investigate whether the bimodal intron length distribution and the biased occurrence of short and long mode introns with respect to their ordinal position are adaptive. In this context, it also needs to be assessed whether the bimodal intron length distribution is a synapomorphy of Neodermata, a hypothesis that currently implies a secondary loss of this genomic trait in Trematoda as observed for *S. mansoni* (Tsai et al. 2013) (see also supplementary file S1, Section D and fig. S10, Supplementary Material online).

Peroxisomes are largely ubiquitous organelles in eukaryotes involved in a variety of metabolic processes and essential in fatty acid—and antioxidant metabolism (Lazarow and Fujiki 1985; Schluter et al. 2007). The absence of many genes associated with peroxisomes including four unequivocal peroxisome marker proteins (Schluter et al. 2006) led to the conclusion that tapeworms and flukes may lack peroxisomes (Tsai et al. 2013), a feature also reported for the Apicomplexan parasites *Plasmodium falciparum*, *P. yoelii*, *Cryptosporidium parvum*, *Toxoplasma gondii*, and *Theileria parva* (Schluter et al. 2006, 2007). Our results indicate the absence of peroxisomes also in *G. salaris* and confirm the findings for endoparasitic flatworms, whereas the detection of all four marker proteins in *S. mediterranea*, *H. robusta*, *L. gigantea*, and *C. elegans*, respectively, denotes the presence of functional peroxisomes in the latter species. The inability to synthesize fatty acids de novo has been previously reported for tapeworms (Tsai et al. 2013) and flukes (Berriman et al. 2009; Wang et al. 2011) as several key enzymes were lacking, and only a few proteins involved in fatty acid precursor production were detected. It was therefore proposed that endoparasitic flatworms scavenge essential fatty acids from their hosts, and the high expression of fatty acid binding proteins (Tsai et al. 2013) seemed in line with the hypothesis. However, a similar strategy might also be adopted by the ectoparasitic *G. salaris*. The genomic data indicate that the loss of peroxisomes and the substantial reduction in the fatty acid biosynthesis pathway are characteristic for an adaptive strategy to the obligate parasitic lifestyle of Neodermata. Further studies will be needed to test this hypothesis and to unravel the underlying biochemical details.

The reductions in specific genes/gene families described above are in line with the more general observation that a large number of expected bilaterian gene models are absent in Neodermata (fig. 3). Although such comparisons rely on many assumptions and may in addition suffer from a somewhat uncertain definition of the set of expected gene models, the losses of genes in *G. salaris*, as well as in tapeworms and flukes, appeared much higher than in other bilaterian lineages. A similar loss of expected loci has recently been reported for conserved microRNAs (Fromm et al. 2013). Whether or not the loss of genes correlates with the loss of further specific biochemical pathways, such as the fatty acid biosynthesis, or represents a genome-wide trend needs to be addressed in more detail in future studies. However, this study identified a large number of genomic traits shared in obligate parasitic flatworms in spite of the obvious ecological differences of ecto- and endoparasitic lifestyles. These traits are thus likely to have arisen as an adaptation to parasitism, already in the common ancestor of neodermatan flatworms. The high-quality draft genome of *G. salaris* facilitates now to include also a representative of the Monogenea in future comparative genomics studies addressing important biological questions such as the genetic basis for ecto- and endoparasitic lifestyles as well as the direct (single-host) and the complex (multiple-hosts) lifecycles represented in the Monogenea and the Trematoda+Cestoda sistergroup, respectively.

Impact of the *G. salaris* Genome

Notably, *G. salaris* is a notorious parasite of Atlantic salmon and other European salmonids and is frequently referred to as "the Salmon killer." The parasite causes significant ecological and economic damage in infected watercourses. In recent years, substantial effort has been directed toward an understanding of strain-specific pathogenicity and transmission behavior in *G. salaris* (reviewed in Bakke et al. 2007). However, most studies have relied on intergenic ribosomal spacers and/or mitochondrial genes as molecular markers (Hansen et al. 2003, 2007; Meinila et al. 2004), which have proven insufficient for resolving these issues, and the lack of diagnostic markers thus remains a major limitation in *G. salaris* strain

identification and characterization. The provided draft genome of *G. salaris* can now facilitate the development of diagnostic markers, and provides a platform for future investigations of the genomic basis of the varying pathogenicity, host specificity, and -preference of various strains of *G. salaris*. In general, the first monogenean reference genome is expected to enhance future studies on the genomics of host–parasite interactions and speciation in the highly diverse monogenean flatworms, and will provide a platform for the development of drug-based control strategies specifically targeting essential biochemical functions in monogenean parasites. The gene models presented in this study may also allow the identification of "speciation genes," that is, genes that retain reproductive isolation (Wu and Ting 2004) after rapid evolution following the initial utilization of a new host species. Host switching is assumed to play a major role in the diversification of monogenean species (Huyse and Volckaert 2005) and other parasitic organisms in general (Sorenson et al. 2003).

Supplementary Material

Supplementary files S1–S5 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors are thankful to T. A. Bakke for providing *G. salaris* samples, P. D. Harris for comments on the manuscript, G. Koutsovoulos for advice concerning genome assembly and gene prediction, M. M. Worren for script contributions, and M. Johansen for implementing the web page. Access to high-performance computing facilities granted by NOTUR (project nn9201k) is gratefully acknowledged. This study was supported by internal funding of the Natural History Museum, University of Oslo, Norway.

Literature Cited

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402.

Ashburner M, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 25:25–29.

Baguna J, Riutort M. 2004. Molecular phylogeny of the Platyhelminthes. Can J Zool. 82:168–193.

Bakke TA, Cable J, Harris PD. 2007. The biology of gyrodactylid monogeneans: the "Russian-doll killers.". Adv Parasitol. 64:161–376.

Berriman M, et al. 2009. The genome of the blood fluke *Schistosoma mansoni*. Nature 460:352–365.

Cable J, Tinsley RC, Harris PD. 2002. Survival, feeding and embryo development of *Gyrodactylus gasterostei* (Monogenea: Gyrodactylidae). Parasitology 124:53–68.

Conesa A, et al. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674–3676.

Danecek P, et al. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158.

Davidson WS, et al. 2010. Sequencing the genome of the Atlantic salmon (*Salmo salar*). Genome Biol. 11:403.

Dunn CW, et al. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452:745–749.

Felsenstein J. 1987. Cases in which parsimony or compatibility methods will be positively misleading. Syst Biol. 27:401–410.

Fischer S, et al. 2011. Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. Chapter 6. Curr Protoc Bioinformatics. Unit 6.12:1–19.

Fromm B, Worren MM, Hahn C, Hovig E, Bachmann L. 2013. Substantial loss of conserved and gain of novel microRNA families in flatworms. Mol Biol Evol. 30:2619–2628.

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907.

Gazave E, Marques-Bonet T, Fernando O, Charlesworth B, Navarro A. 2007. Patterns and rates of intron divergence between humans and chimpanzees. Genome Biol. 8:R21.

Gentile L, Cebria F, Bartscherer K. 2011. The planarian flatworm: an in vivo model for stem cell biology and nervous system regeneration. Dis Model Mech. 4:12–19.

Gotz S, et al. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 36: 3420–3435.

Haddrill PR, Charlesworth B, Halligan DL, Andolfatto P. 2005. Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC content. Genome Biol. 6:R67.

Hahn C, Bachmann L, Chevreux B. 2013. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. Nucleic Acids Res. 41:e129.

Hansen H, Bachmann L, Bakke TA. 2003. Mitochondrial DNA variation of *Gyrodactylus* spp. (Monogenea, Gyrodactylidae) populations infecting Atlantic salmon, grayling, and rainbow trout in Norway and Sweden. Int J Parasitol. 33:1471–1478.

Hansen H, Bakke TA, Bachmann L. 2007. DNA taxonomy and barcoding of monogenean parasites: lessons from *Gyrodactylus*. Trends Parasitol. 23:363–367.

Harris PD, Bachmann L, Bakke TA. 2011. The parasites and pathogens of the Atlantic salmon: lessons from *Gyrodactylus salaris*. In: Aas O, Einum S, Klemetsen A, Skurdal J, editors. Atlantic salmon ecology. Chichester (United Kingdom): Wiley-Blackwell. p. 221–252.

Hejnol A, et al. 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. Proc Biol Sci. 276:4261–4270.

Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12:491.

Huyse T, Volckaert FA. 2005. Comparing host and parasite phylogenies: gyrodactylus flatworms jumping from goby to goby. Syst Biol. 54: 710–718.

Janicki C. 1920. Grundinien einer ''Cercomer Theorie'' zur Morphologie der Trematoden und Cestoden. Festschr Zschokke. 30:1–22.

Justine JL. 1998. Non-monophyly of the monogeneans? Int J Parasitol. 28: 1653–1657.

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. 2012. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 40:D109–D114.

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

Kuck P, Meusemann K. 2010. FASconCAT: convenient handling of data matrices. Mol Phylogenet Evol. 56:1115–1118.

Kumar S, Blaxter ML. 2011. Simultaneous genome sequencing of symbionts and their hosts. Symbiosis 55:119–126.

Kumar S, Schiffer PH, Blaxter M. 2012. 959 Nematode Genomes: a semantic wiki for coordinating sequencing projects. Nucleic Acids Res. 40:D1295–D1300.

Kück P. 2009. ALICUT: a Perlscript which cuts ALISCORE identified RSS, version 2.0 edn [Internet]. [cited 2014 April 30]. Available from: http://www.zfmk.utilities.de.
Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol. 21:1095–1109.
Lartillot N, Rodrigue N, Stubbs D, Richer J. 2013. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Syst Biol. 62:611–615.
Lazarow PB, Fujiki Y. 1985. Biogenesis of peroxisomes. Annu Rev Cell Biol. 1:489–530.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760.
Li R, et al. 2010a. The sequence and de novo assembly of the giant panda genome. Nature 463:311–317.
Li R, et al. 2010b. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 20:265–272.
Littlewood DTJ. 2006. The evolution of parasitism in flatworms. In: Maule AG, Marks NJ, editors. Parasitic flatworms: molecular biology, biochemistry, immunology and physiology. Wallingford (United Kingdom): Cabi Publishing-C a B Int. p. 1–36.
Littlewood DTJ, Rohde K, Clough KA. 1999. The interrelationships of all major groups of Platyhelminthes: phylogenetic evidence from morphology and molecules. Biol J Linn Soc Lond. 66:75–114.
Liu F, et al. 2006. New perspectives on host–parasite interplay by comparative transcriptomic and proteomic analyses of *Schistosoma japonicum*. PLoS Pathog. 2:268–281.
Lockyer AE, Olson PD, Littlewood DTJ. 2003. Utility of complete large and small subunit rRNA genes in resolving the phylogeny of the Neodermata (Platyhelminthes): implications and a review of the cercomer theory. Biol J Linn Soc Lond. 78:155–171.
Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 33:6494–6506.
Luo R, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1:18.
Marais G, Nouvellet P, Keightley PD, Charlesworth B. 2005. Intron size and exon evolution in *Drosophila*. Genetics 170:481–485.
Margulies M, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380.
Meinila M, Kuusela J, Zietara MS, Lumme J. 2004. Initial steps of speciation by geographic isolation and host switch in salmonid pathogen *Gyrodactylus salaris* (Monogenea: Gyrodactylidae). Int J Parasitol. 34:515–526.
Misof B, Misof K. 2009. A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. Syst Biol. 58:21–34.
Mollaret I, et al. 1997. Phylogenetic analysis of the Monogenea and their relationships with Digenea and Eucestoda inferred from 28S rDNA sequences. Mol Biochem Parasitol. 90:433–438.
Myers EW, et al. 2000. A whole-genome assembly of *Drosophila*. Science 287:2196–2204.
Ohashi H, et al. 2007. Expression of *vasa* (vas)-related genes in germ cells and specific interference with gene functions by double-stranded RNA in the monogenean, *Neobenedenia girellae*. Int J Parasitol. 37:515–523.
Olson PD, Tkach VV. 2005. Advances and trends in the molecular systematics of the parasitic platyhelminthes. Adv Parasitol. 60:165–243.
Park JK, et al. 2007. A common origin of complex life cycles in parasitic flatworms: evidence from the complete mitochondrial genome of *Microcotyle sebastis* (Monogenea: Platyhelminthes). BMC Evol Biol. 7:11.
Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23:1061–1067.
Parra G, Bradnam K, Ning ZM, Keane T, Korf I. 2009. Assessing the gene space in draft genomes. Nucleic Acids Res. 37:289–297.
Pell J, et al. 2012. Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. Proc Natl Acad Sci U S A. 109:13272–13277.
Perkins EM, Donnellan SC, Bertozzi T, Whittington ID. 2010. Closing the mitochondrial circle on paraphyly of the Monogenea (Platyhelminthes) infers evolution in the diet of parasitic flatworms. Int J Parasitol. 40:1237–1245.
Philippe H, et al. 2011. Acoelomorph flatworms are deuterostomes related to *Xenoturbella*. Nature 470:255–258.
Quevillon E, et al. 2005. InterProScan: protein domains identifier. Nucleic Acids Res. 33:W116–W120.
Robb SMC, Ross E, Alvarado AS. 2008. SmedGD: the *Schmidtea mediterranea* genome database. Nucleic Acids Res. 36:D599–D606.
Rodrigue S, et al. 2009. Whole genome amplification and de novo assembly of single bacterial cells. PLoS One 4:e6864.
Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497:327–331.
Salte R, et al. 2010. Prospects for a genetic management strategy to control *Gyrodactylus salaris* infection in wild Atlantic salmon (*Salmo salar*) stocks. Can J Fish Aquat Sci. 67:121–129.
Schistosoma japonicum Genome Sequencing, Functional Analysis Consortium. 2009. The *Schistosoma japonicum* genome reveals features of host–parasite interplay. Nature 460:345–351.
Schluter A, et al. 2006. The evolutionary origin of peroxisomes: an ER-peroxisome connection. Mol Biol Evol. 23:838–845.
Schluter A, et al. 2007. PeroxisomeDB: a database for the peroxisomal proteome, functional genomics and disease. Nucleic Acids Res. 35:D815–D822.
Sievers F, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 7:539.
Simakov O, et al. 2013. Insights into bilaterian evolution from three spiralian genomes. Nature 493:526–531.
Simpson JT, et al. 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19:1117–1123.
Skinner DE, Rinaldi G, Koziol U, Brehm K, Brindley PJ. 2014. How might flukes and tapeworms maintain genome integrity without a canonical piRNA pathway? Trends Parasitol. 30:123–129.
Smit AFA, Hubley R. 2008–2010. RepeatModeler Open-1.0. [cited 2014 April 30]. Available from: http://www.repeatmasker.org.
Smit AFA, Hubley R, Green P. 1996–2010. RepeatMasker Open-3.0. [cited 2014 April 30]. Available from: http://www.repeatmasker.org.
Smith SA, et al. 2011. Resolving the evolutionary relationships of molluscs with phylogenomic tools. Nature 480:364–367.
Sorenson MD, Sefc KM, Payne RB. 2003. Speciation by host switch in brood parasitic indigobirds. Nature 424:928–931.
Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.
Stanke M, Waack S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19(Suppl. 2):ii215–ii225.
Tsai IJ, et al. 2013. The genomes of four tapeworm species reveal adaptations to parasitism. Nature 496:57–63.
van der Werf MJ, et al. 2003. Quantification of clinical morbidity associated with schistosome infection in sub-Saharan Africa. Acta Trop. 86:125–139.
van Dongen S. 2000. Graph clustering by flow simulation [PhD thesis]. [Utrecht (Holland)]: University of Utrecht.
Wang XY, et al. 2011. The draft genome of the carcinogenic human liver fluke *Clonorchis sinensis*. Genome Biol. 12:R107.
Wu CI, Ting CT. 2004. Genes and speciation. Nat Rev Genet. 5:114–122.
Yandell M, Ence D. 2012. A beginner's guide to eukaryotic genome annotation. Nat Rev Genet. 13:329–342.

Yu J, et al. 2002. Minimal introns are not "junky.". Genome Res. 12: 1185–1189.
Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18:821–829.
Zhang Q, Edwards SV. 2012. The evolution of intron size in amniotes: a role for powered flight? Genome Biol Evol. 4:1033–1043.
Zheng HJ, et al. 2013. The genome of the hydatid tapeworm *Echinococcus granulosus*. Nat Genet. 45:1168–1175.
Zheng Y. 2013. Phylogenetic analysis of the Argonaute protein family in platyhelminths. Mol Phylogenet Evol. 66:1050–1054.

Associate editor: Andreas Wagner