The genome of the crustacean *Parhyale hawaiensis*: a model for animal development, regeneration, immunity and lignocellulose digestion

Damian Kao[1], Alvina G. Lai[1], Evangelia Stamataki[2], Silvana Rosic[3,4], Nikolaos Konstantinides[5], Erin Jarvis[6], Alessia Di Donfrancesco[1], Natalia Pouchkina-Stantcheva[1], Marie Sémon[5], Marco Grillo[5], Heather Bruce[6], Suyash Kumar[2], Igor Siwanowicz[2], Andy Le[2], Andrew Lemire[2], Michael B. Eisen[7], Cassandra Extavour[8], William E. Browne[9], Carsten Wolff[10], Michalis Averof[5], Nipam H. Patel[6], Peter Sarkies[3,4], Anastasios Pavlopoulos[2], and A. Aziz Aboobaker[1]

[1]Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, United Kingdom

[2]Howard Hughes Medical Institute, Janelia Research Campus, Ashburn, United States

[3]MRC Clinical Sciences Centre, Imperial College London, Hammersmith Hospital Campus, Du Cane Road, London W12 0NN

[4]Institute for Clinical Sciences, Imperial College London, Hammersmith Hospital Campus, Du Cane Road, London, W12 0NN

[5]Institut de Génomique Fonctionnelle de Lyon, Centre National de la Recherche Scientifique (CNRS) and École Normale Supérieure de Lyon, Lyon, France

[6]University of California, Berkeley, Dept. of Molecular and Cell Biology, 519A LSA 3200 Berkeley, CA 94720-3200

[7]Howard Hughes Medical Institute, University of California, Berkeley, Department Of Molecular and Cell Biology

[8]Department of Organismic and Evolutionary Biology, Harvard University 16 Divinity Avenue, BioLabs Building 4109-4111 Cambridge, MA 02138

[9]Department of Invertebrate Zoology, Smithsonian National Museum of Natural History, MRC-163, P.O. Box 37012, Washington, DC 20013-7012 USA

[10]Humboldt-Universität zu Berlin, Institut für Biologie, Vergleichende Zoologie, Philippstr. 13, Haus 2, 10115 Berlin, Germany

AUTHOR INFORMATION

These authors contributed equally to this work: Damian Kao, Alvina G. Lai, Evangelia Stamataki.

These authors also contributed equally: Anastasios Pavlopoulos, A. Aziz Aboobaker

AUTHORS CONTRIBUTIONS

Damian Kao, devised assembly strategy. Assembled and analyzed the sequencing data. Annotated the genome, transcriptome and proteome. Performed orthology group analysis. Annotated small RNAs. Drafting and revising the article.

Alvina Lai, analysed the genome including major signaling pathways, polymorphisms, immunity, lignocellulose digestion, epigenetic pathways, small RNA pathways and small RNAs. Cloning of DSCAM variants. Experimental confirmation of polymorphisms. Drafting and revising the article.

Evangelia Stamataki, prepared the genomic libraries, performed CRISPR knock-out, performed CRISPR knock-in. Drafting revising the article.

Silvana Rosic, contributed bisulfite sequencing data and analysis of genome wide methylation.

Nikolaos Konstantinides, contributed transcriptome data and transcriptome assembly.

Erin Jarvis, performed in situ hybridization detection of Hox genes, and interpreted data.

Alessia Di DonFrancesco, contributed to confirmation of polymorphism and cloning of Ph-DSCAM variants.

Natalia Pouchkina-Stancheva, contributed to confirmation of polymorphism and cloning of Ph-DSCAM variants.

Marie Sémon, contributed transcriptome data and transcriptome assembly.

Marco Grillo, contributed transcriptome data and transcriptome assembly.

Heather Bruce, contributed transcriptome data.

Suyash Kumar, performed CRISPR knock-out.

Igor Siwanowicz, performed Parhyale cuticle staining.

Andy Le, performed CRISPR knock-in.

Andrew Lemire, was consulted about sequencing strategy and helped with bioinformatics.

Michael Eisen, Contributed to RNAseq data production.

Cassandra Extavour, Contributed to project planning.

William E. Browne, Established the Chicago-F line.

Carsten Wolff, Performed karyotyping.

Michalis Averof, contributed transcriptome data and transcriptome assembly.

Nipam Patel, performed in situ hybridization detection of Hox genes, and interpreted data. Contributed transcriptome data. Established the Chicago-F line.

Peter Sarkies, conceived and designed bisulfite sequencing experiments, contributed bisultfite sequencing data and analysis of genome wide methylation.

Anastasios Pavlopoulos,conceived, designed and managed the project. Contributed to data acquisition and analysis. Drafting and revising the article.

Aziz Aboobaker, Devised assembly strategy. Contributed to data analysis. Conceived, designed and managed the project. Drafting and revising the article.

All authors read and approved the manuscript.

For correspondence

Aziz.Aboobaker@zoo.ox.ac.uk (AAA) pavlopoulosa@janelia.hhmi.org (AP)

Competing interests

The authors declare no competing interests

Funding

AAA and co-workers are funded by the Biotechnology and Biological Sciences Reserarch Council (BBSRC grant number BB/K007564/1), the Medical Research Council (MRC grant number MR/M000133/1), the John Fell Fund Oxford University Press (OUP) and the Elizabeth Hannah Jenkinson Fund. AGL receives funding from the Human Frontier Science Program postdoctoral fellowship and the Elizabeth Hannah Jenkinson Research Fund. NHP and co-workers are funded by NSF grant IOS-1257379. AP and co-workers are funded by the Howard Hughes Medical Institute. PS and co-workers are funded by the Medical Research council (MRC MC-A652- 5PZ80) and an Imperial College Research Fellowship too PS. MA and colleagues received funding from the Agence Nationale de la Recherche (France), grant ANR-12-CHEX-0001-01. The funding bodies had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

ABSTRACT

The amphipod crustacean *Parhyale hawaiensis* is a blossoming model system for studies of developmental mechanisms and more recently regeneration. We have sequenced the genome allowing annotation of all key signaling pathways, transcription factors, and non-coding RNAs that will enhance ongoing functional studies. *Parhyale* is a member of the Malacostraca clade, which includes crustacean food crop species. We analysed the immunity related genes of *Parhyale* as an important comparative system for these species, where immunity related aquaculture problems have increased as farming has intensified. We also find that *Parhyale* and other species within Multicrustacea contain the enzyme sets necessary to perform lignocellulose digestion (“wood eating”), suggesting this ability may predate the diversification of this lineage. Our data provide an essential resource for further development of *Parhyale* as an experimental model. The first malacostracan genome will underpin ongoing comparative work in food crop species and research investigating lignocellulose as an energy source.

INTRODUCTION

Very few members of the Animal Kingdom hold the esteemed position of major model system for understanding living systems. Inventions in molecular and cellular biology increasingly facilitate the emergence of new experimental systems for developmental genetic studies. The morphological and ecological diversity of the phylum Arthropoda makes them an ideal group of animals for comparative studies encompassing embryology, adaptation of adult body plans and life history evolution [1–4]. While the most widely studied group are Hexapods, reflected by over a hundred sequencing projects available in the NCBI genome database, genomic data in the other three sub-phyla in Arthropoda are still relatively sparse.

Recent molecular and morphological studies have placed crustaceans along with hexapods into a pancrustacean clade (Figure 1A), revealing that crustaceans are paraphyletic [5–9]. Previously, the only available fully sequenced crustacean genome was that of the water flea *Daphnia* which is a member of the Branchiopoda [10]. A growing number of transcriptomes for larger phylogenetic analyses have led to differing hypotheses of the relationships of the major pancrustacean groups (Figure 1B) [11–14]. The genome of the amphipod crustacean *Parhyale hawaiensis* addresses the paucity of high quality non-hexapod genomes among the pancrustacean group, and will help to resolve relationships within this group as more genomes and complete proteomes become available [15, 16]. Crucially, genome sequence data is also necessary to further advance research in *Parhyale*, currently the most tractable crustacean model system. This is particularly true for the application of powerful functional genomic approaches, such as genome editing [17–22].

Parhyale is a member of the diverse Malacostraca clade with thousands of extant species including economically and nutritionally important groups such as shrimps, crabs, crayfish and lobsters, as well as common garden animals like woodlice. They are found in all marine, fresh water, and higher humidity terrestrial environments. Apart from attracting research interest as an economically important food crop, this group of animals has been used to study developmental biology and the evolution of morphological diversity (for example with respect to Hox genes) [19, 23–25], stem cell biology [26, 27], innate immunity processes [28, 29] and recently the cellular mechanisms of regeneration [26, 27]. In addition, members of the Malacostraca, specifically both Amphipods and Isopods, are thought to be capable of “wood eating” or lignocellulose digestion and to have microbiota-free digestive systems [30–33].

The life history of *Parhyale* makes it a versatile model organism amenable to experimental manipulations (Figure 1C)[34]. Gravid females lay eggs every 2 weeks upon reaching sexual maturity and hundreds of eggs can be easily collected at all stages of embryogenesis. Embryogenesis takes about 10 days at 26°C and has been described in detail with an accurate staging system [35]. Early embryos display an invariant cell lineage with each blastomere at the 8-cell stage contributing to a specific germ layer (Figure 1D)[35, 36]. Embryonic and post-embryonic stages are amenable to experimental manipulations and direct observation *in vivo* [36–48]. These can be combined with transgenic approaches [25, 45, 48, 49], RNA interference (RNAi) [24] and morpholino-mediated gene knockdown [50], and transgene-based lineage tracing [26]. Most recently the utility of the clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated (Cas) system for targeted genome editing has been elegantly demonstrated during the systematic study of *Parhyale* Hox genes [18, 19]. This arsenal of experimental tools (Table 1) has already established *Parhyale* as an attractive model system for biological research.

So far, work in *Parhyale* has been constrained by the lack of a reference genome and other standardized genome-wide resources. To address this limitation, we have sequenced, assembled and annotated the genome. At an estimated size of 3.6 Gb, this genome represents one of the largest animal genomes tackled to date. The large size has not been the only challenge of the *Parhyale* genome, that also exhibits some of the highest levels of sequence repetitiveness and polymorphism reported among published genomes. We provide information in our assembly regarding polymorphism to facilitate functional genomic approaches sensitive to levels of sequence similarity, particularly homology-dependent genome editing approaches. We analysed a number of key features of the genome as foundations for new areas of research in *Parhyale*, including innate immunity in crustaceans, lignocellulose digestion, non-coding RNA biology, and epigenetic control of the genome. Our data bring *Parhyale* to the forefront of developing model systems for a broad swathe of important bioscience research questions.

RESULTS AND DISCUSSION

Genome assembly, annotation, and validation

The *Parhyale* genome contains 23 pairs (2n=46) of chromosomes (Figure 2) and with an estimated size of 3.6 Gb, it is currently the second largest reported arthropod genome after the locust genome [51, 52]. Sequencing was performed on genomic DNA isolated from a single adult male taken from a line derived from a single female and expanded after two rounds of sib-mating. We performed k-mer analyses of the trimmed reads to assess the impact of repeats and polymorphism on the assembly process. We analyzed k-mer frequencies (Figure 3A) and compared k-mer representation between our different sequencing libraries. We observed a 93% intersection of unique k-mers among sequencing libraries, indicating that the informational content was consistent between libraries (Supplemental Data 6). The k-mer analysis revealed a bimodal distribution of error-free k-mers (Figure 3A). The higher-frequency peak corresponded to k-mers present on both haplotypes (i.e. homozygous regions), while the lower-frequency peak had half the coverage and corresponded to k-mers present on one haplotype (i.e. heterozygous regions) [53]. We concluded that the single sequenced adult *Parhyale* exhibits very high levels of heterozygosity, similar to the highly heterozygous oyster genome (see below).

In order to quantify global heterozygosity and repeat content of the genome we assessed the de-Bruijn graphs generated from the trimmed reads to observe the frequency of both variant and repeat branches [54] (Figure 3B and C). We found that the frequency of the variant branches was 10x higher than that observed in the human genome and very similar to levels in the highly polymorphic genome of the oyster *Crassostrea gigas* [55]. We also observed a frequency of repeat branches approximately 4x higher than those observed in both the human and oyster genomes (Figure 3C), suggesting that the big size of the *Parhyale* genome can be in large part attributed to the expansion of repetitive sequences.

These metrics suggested that both contig assembly and scaffolding with mate-pair reads were likely to be challenging due to high heterozygosity and repeat content. After an initial contig assembly we remapped reads to assess coverage of each contig. We observed a major peak centered around 75 x coverage and a smaller peak at 150x coverage. Contigs with lower 75x coverage represent regions of the genome that assembled into separate haplotypes and had half the frequency of mapped sequencing reads, reflecting high levels of heterozygosity. This resulted in independent assembly of haplotypes for much of the genome (Figure 3D).

One of the prime goals in sequencing the *Parhyale* genome was to achieve an assembly that could assist functional genetic and genomic approaches in this species. Different strategies have been employed to sequence highly heterozygous diploid genomes of non-model and wild-type samples [56]. We aimed for an assembly representative of different haplotypes, allowing manipulations to be targeted to different allelic variants in the assembly. This could be particularly important for homology dependent strategies that are likely to be sensitive to polymorphism. However, the presence of alternative haplotypes could lead to poor scaffolding between contigs as many mate-pair reads may not map uniquely to one contig and distinguish between haplotypes in the assembly. To alleviate this problem we used a strategy to conservatively identify pairs of allelic contigs and proceeded to use only one in the scaffolding process. First, we estimated levels of similarity (identity and alignment length) between all assembled contigs to identify independently assembled allelic regions (Figure 3E). We then kept the longer contig of each pair for scaffolding using our mate-pair libraries (Figure 3F), after which we added back the shorter allelic contigs to produce the final genome assembly (Figure 4A).

RepeatModeler and RepeatMasker were used on the final assembly to find repetitive regions, which were subsequently classified into families of transposable elements or short tandem repeats (Supplemental Data 7). We found 1,473 different repeat element sequences representing 57% of the assembly (Figure 4, Supplemental Table 1). The *Parhyale* assembly comprises of 133,035 scaffolds (90% of assembly), 259,343 unplaced contigs (4% of assembly), and 584,392 shorter, potentially allelic contigs (6% of assembly), with a total length of 4.02 Gb (Table 2). The N50 length of the scaffolds is 81,190bp. The final genome assembly was annotated with Augustus trained with high confidence gene models derived from assembled transcriptomes, gene homology, and *ab initio* predictions. This resulted in 28,155 final gene models (Figure 4B; Supplemental Data 8) across 14,805 genic scaffolds and 357 unplaced contigs with an N50 of 161,819, bp and an N90 of 52,952 bp.

Parhyale has a mean coding gene size (introns and ORFs) of 20kb (median of 7.2kb), which is longer than *D. pulex* (mean: 2kb, median: 1.2kb), while shorter than genes in *Homo sapiens* (mean: 52.9kb, median: 18.5kb). This difference in gene length was consistent across reciprocal blast pairs where ratios of gene lengths revealed *Parhyale* genes were longer than *Caenorhabditis elegans*, *D. pulex*, and *Drosophila melanogaster* and similar to *H. sapiens*. (Figure 5A). The mean intron size in *Parhyale* is

5.4kb, similar to intron size in *H. sapiens* (5.9kb) but dramatically longer than introns in *D. pulex* (0.3kb), *D. melanogaster* (0.3kb) and *C. elegans* (1kb) (Figure 5B).

For downstream analyses of *Parhyale* protein coding content, a final proteome consisting of 28,666 proteins was generated by combining candidate coding sequences identified with TransDecoder [57] from mixed stage transcriptomes. Almost certainly the high number of predicted gene models and proteins is an overestimation due to fragmented genes, very different isoforms or unresolved alleles, that will be consolidated as annotation of the *Parhyale* genome improves. We also included additional high confidence gene predictions that were not found in the transcriptome (Figure 4C). The canonical proteome dataset was annotated with both Pfam, KEGG, and BLAST against Uniprot. Assembly quality was further evaluated by alignment to core eukaryotic genes defined by the Core Eukaryotic Genes Mapping Approach (CEGMA) database [58]. We identified 244/248 CEGMA orthology groups from the assembled genome alone and 247/248 with a combination of genome and mapped transcriptome data (Figure 4, Supplemental Figure 2). Additionally, 96% of over 280,000 identified transcripts, most of which are fragments that do not contain a large ORF, also mapped to the assembled genome. Together these data suggest that our assembly is close to complete with respect to protein coding genes and transcribed regions that are captured by deep RNA sequencing.

High levels of heterozygosity and polymorphism in the *Parhyale* genome

To estimate the level of heterozygosity in genes we first identified transcribed regions of the genome by mapping back transcripts to the assembly. Where these regions appeared in a single contig in the assembly, heterozygosity was calculated using information from mapped reads. Where these regions appeared in more than one contig, because haplotypes had assembled independently, heterozygosity was calculated using an alignment of the genomic sequences corresponding to mapped transcripts and information from mapped reads. This allowed us to calculate heterozygosity for each gene within the sequenced individual (Supplemental Data 9). We then calculated the genomic coverage of all transcribed regions in the genome and found, as expected, they fell broadly into two categories with higher and lower read coverage (Figure 6A; Supplemental Data 9). Genes that fell within the higher read coverage group had a lower mean heterozygosity (1.09% of bases displaying polymorphism), which is expected as more reads were successfully mapped. Genes that fell within the lower read coverage group had a higher heterozygosity (2.68%), as reads mapped independently to each haplotype (Figure 6B) [54]. Thus, we conclude that heterozygosity that influences read mapping and assembly of transcribed regions, and not just non-coding parts of the assembly.

The assembled *Parhyale* transcriptome was derived from various laboratory populations, hence we expected to see additional polymorphism beyond that detected in the two haplotypes of the individual male we sequenced. Analysing all genes using the transcriptome we found additional variations in transcribed regions not found in the genome of the sequenced individual. In addition to polymorphisms that agreed with heterozygosity in the genome sequence we observed that the rate of additional variations is not substantially different between genes from the higher (0.88%) versus lower coverage group genes (0.73%; Figure 6C). This analysis suggests that within captive laboratory populations of *Parhyale* there is considerable additional polymorphism distributed across genes, irrespective of whether or not they have relatively low or high heterozygosity in the individual male we sequenced. In addition the single male we have sequenced provides an accurate reflection of polymorphism of the wider laboratory population and the established Chicago-F strain does not by chance contain unusually divergent haplotypes. We also performed an assessment of polymorphism on previously cloned *Parhyale* developmental genes, and found some examples of startling levels of variation. (Supplemental Data 2, Figure 6, Supplemental Figure 1). For example, we found that the cDNAs of the germ line determinants, *nanos* (78 SNPS, 34 non-synonymous substitutions and one 6bp indel) and *vasa* (37 SNPs, 7 non-synonymous substitutions and a one 6bp indel) can have more variability within laboratory *Parhyale* populations than might be observed for orthologs between closely related species.

To further evaluate the extent of polymorphism across the genome, we mapped the genomic reads to a set of previously Sanger-sequenced BAC clones of the *Parhyale* Hox cluster from the same Chicago-F line from which we sequenced the genome of an adult male. [18]. We detected SNPs at a rate of 1.3 to 2.5% among the BACs (Table 3) and also additional sequence differences between the BACs and genomic reads, confirming that additional polymorhism exists in the Chicago-F line beyond that detected between in the haplotypes of the individual male we sequenced.

Overlapping regions of the contiguous BACs gave us the opportunity to directly compare Chicago-F haplotypes and accurately observe polynucleotide polymorphisms, that are difficult to detect with short reads that do not map when polymorphisms are large, but are resolved by longer Sanger reads. (Figure 7A). Since the BAC clones were generated from a pool of Chicago-F animals, we expected each sequenced BAC to be representative of one haplotype. Overlapping regions between BAC clones could potentially represent one or two haplotypes. We found that the genomic reads supported the SNPs observed between the overlapping BAC regions. We found relatively few base positions with evidence supporting the existence of a third allele. This analysis revealed many insertion/deletion (indels) with some cases of indels larger than 100 base pairs (Figure 7B). The finding that polynucleotide polymorphisms are prevalent between the haplotypes of the Chicago-F is another reason, in addition to regions of high SNP heterozygosity in the genome sequence, for the extensive independent assembly of haplotypes. Taken togther these data mean that special attention will have to be given to those functional genomic approaches that are dependent on homology, such as CRISPR/Cas9 based knock in strategies.

A comparative genomic analysis of the *Parhyale* genome

Assessment of conservation of the proteome using BLAST against a selection of metazoan proteomes was congruent with broad phylogenetic expectations. These analyses included crustacean proteomes likely to be incomplete as they come from limited transcriptome datasets, but nonetheless highlighted genes likely to be specific to the Malacostraca (Figure 5C). To better understand global gene content evolution we generated clusters of orthologous and paralogous gene families comparing the *Parhyale* proteome with other complete proteomes across the Metazoa using Orthofinder [59] (Figure 5D; Supplemental Data 10). Amongst proteins conserved in protostomes and deuterostomes we saw no evidence for widespread gene duplication in the lineage leading to *Parhyale*. We identified orthologous and paralogous protein groups across 16 species with 2,900 and 2,532 orthologous groups containing proteins found only in Panarthropoda and Arthropoda respectively. We identified 855 orthologous groups that were shared exclusively by Mandibulata, 772 shared by Pancrustacea and 135 shared by Crustacea. There were 9,877 *Parhyale* proteins that could not be assigned to an orthologous group, potentially representing rapidly evolving or lineage specific proteins. Amongst these proteins we found 609 proteins (2.1% of the proteome) that had paralogs within *Parhyale*, suggesting that younger and/or more divergent *Parhyale* genes have undergone some considerable level of gene duplication events.

Our analysis of shared orthologous groups was equivocal with regard to alternative hypotheses on the relationships among pancrustacean subgroups: 44 groups of orthologous proteins are shared among the multicrustacea clade (uniting the Malacostraca, Copepoda and Thecostraca), 37 groups are shared among the Allocarida (Branchiopoda and Hexapoda) and 49 groups are shared among the Vericrustacea (Branchiopoda and Multicrustacea)(Supplemental Data 17).

To further analyse the evolution of the *Parhyale* proteome we examined protein families that appeared to be expanded (z-score >2), compared to other taxa (Figure 5, Supplemental Figure 1, Supplemental Data 10, Supplemental Data 15). We conservatively identified 29 gene families that are expanded in *Parhyale*. Gene family expansions include the Sidestep (55 genes) and Lachesin (42) immunoglobulin superfamily proteins as well as nephrins (33 genes) and neurotrimins (44 genes), which are thought to be involved in immunity, neural cell adhesion, permeability barriers and axon guidance [60–62]. Other *Parhyale* gene expansions include *APN* (aminopeptidase N) (38 genes) and cathepsin-like genes (30 genes), involved in proteolytic digestion [63].

Major signaling pathways and transcription factors in *Parhyale*

Components of all common metazoan cell-signalling pathways are largely conserved in *Parhyale*. At least 13 *Wnt* subfamilies were present in the cnidarian-bilaterian ancestor. *Wnt3* has been lost in protostomes that retain 12 *Wnt* genes [64–66]. Some sampled ecdysozoans have undergone significant *Wnt* gene loss, for example *C. elegans* has only 5 *Wnt* genes [67]. At most 9 *Wnt* genes are present in any individual hexapod species [68], with *wnt2* and *wnt4* potentially lost before the hexapod radiation [69]. The *Parhyale* genome encodes 6 of the 13 *Wnt* subfamily genes; *wnt1, wnt4, wnt5, wnt10, wnt11* and *wnt16* (Figure 8). *Wnt* genes are known to have been ancestrally clustered [70]. We observed that *wnt1* and *wnt10* are linked in a single scaffold (phaw_30.0003199); given the loss of *wnt6* and *wnt9*, this may be the remnant of the ancient *wnt9-1-6-10* cluster conserved in some protostomes.

We could identify 2 Fibroblast Growth Factor (*FGF*) genes and only a single FGF receptor (*FGFR*) in the *Parhyale* genome, suggesting one *FGFR* has been lost in the malacostracan lineage (Figure 8, Supplemental Figure 1). Within the Transforming Growth Factor beta (*TGF-β*) signaling pathway we found 2 genes from the activin subfamily (an activin receptor and a myostatin), 7 genes from the Bone Morphogen Protein (*BMP*) subfamily and 2 genes from the inhibin subfamily. Of the *BMP* genes, *Parhyale* has a single decapentaplegic homologue (Supplemental Data 2). Other components of the TGF-β pathway were identified such as the neuroblastoma suppressor of tumorigenicity (NBL1/DAN), present in *Aedes aegypti* and *Tribolium castaneum* but absent in *D. melanogaster* and *D. pulex*, and TGFB-induced factor homeobox 1 (*TGIF1*) which is a Smad2-binding protein within the pathway present in arthropods but absent in nematodes (*C. elegans* and *Brugia malayi*;Supplemental Data 2). We identified homologues of *PITX2*, a downstream target of the TGF-β pathway involved in endoderm and mesoderm formation present in vertebrates and crustaceans (*Parhyale* and *D. pulex*) but not in insects and nematodes [71]. With the exception of *SMAD7* and *SMAD8/9*, all other *SMADs* (*SMAD1, SMAD2/3, SMAD4, SMAD6*) are found in arthropods sampled, including *Parhyale*. Components of other pathways interacting with TGF-β signaling like the *JNK, Par6, ROCK1/RhoA, p38* and *Akt* pathways were also recovered and annotated in the *Parhyale* genome (Supplemental Data 2). We identified major Notch signaling components including Notch, Delta, Deltex, Fringe and modulators of the Notch pathway such as *Dvl* and *Numb*. Members of the gamma-secretase complex (Nicastrin, Presenillin, and *APH1*) were also present (Supplemental Data 4) as well as to other co-repressors of the Notch pathway such as Groucho and *CtBP* [72].

A genome wide survey to annotate all potential transcription factors (TFs) discovered a total of 1,143 proteins with DNA binding domains that belonged to all the major families previously identified. Importantly, we observed a large expansion of TFs containing the zinc-finger (ZF)-C2H2 domain, that was previously observed in a trancriptomic study of *Parhyale* [73]. *Parhyale* has 699 ZF-C2H2-containing genes [74], which is comparable to the number found in *H. sapiens* [75], but significantly expanded compared to other arthropod species like *D. melanogaster* encoding 326 members (Figure 8, Supplemental Table 2).

The *Parhyale* genome contains 126 homeobox-containing genes (Figure 9; Supplemental Data 2), which is higher than the numbers reported for other arthropods (104 genes in *D. melanogaster*, 93 genes in the honey bee *Apis melllifera*, and 113 in the centipede *Strigamia maritima*) [76]. We identified a *Parhyale* specific expansion in the Ceramide Synthase (*CERS*) homeobox proteins, which include members with divergent homeodomains [77]. *H. sapiens* have six *CERS* genes, but only five with homeodomains [78]. We observed an expansion to 12 *CERS* genes in *Parhyale*, compared to 1-4 genes found in other arthropods [79] (Figure 8, Supplemental Figure 3). In phylogenetic analyses all 12 *CERS* genes in *Parhyale* clustered together with a *CERS* from another amphipod *Echinogammarus veneris*, suggesting that this is recent expansion in the amphipod lineage.

Parhyale contains a complement of 9 canonical Hox genes that exhibit both spatial and temporal colinearity in their expression along the anterior-posterior body axis [18]. Chromosome walking experiments had shown that the Hox genes *labial* (*lab*) and *proboscipedia* (*pb*) are linked and that

Deformed (*Dfd*), *Sex combs reduced* (*Scr*), *Antennapedia* (*Antp*) and *Ultrabithorax* (*Ubx*) are also contiguous in a cluster [18]. Previous experiments in *D. melanogaster* had shown that the proximity of nascent transcripts in RNA fluorescent *in situ* hybridizations (FISH) coincide with the position of the corresponding genes in the genomic DNA [80, 81]. Thus, we obtained additional information on Hox gene linkage by examining nascent Hox transcripts in cells where Hox genes are co-expressed. We first validated this methodology in *Parhyale* embryos by confirming with FISH, the known linkage of *Dfd* with *Scr* in the first maxillary segment where they are co-expressed (Figure 10A-A"). As a negative control, we detected no linkage between *engrailed1* (*en1*) and *Ubx* or *abd-A* transcripts (Figure 10B - B" and C - C"). We then demonstrated the tightly coupled transcripts of *lab* with *Dfd* (co-expressed in the second antennal segment, Figure 10D - D"), *Ubx* and *abd-A* (co-expressed in the posterior thoracic segments, Figure 10E - E"), and *abd-A* with *Abd-B* (co-expressed in the anterior abdominal segments, (Figure 10F - F"). Collectively, all evidence supports the linkage of all analysed Hox genes into a single cluster as shown in (Figure 10G - G"). The relative orientation and distance between certain Hox genes still needs to be worked out. So far, we have not been able to confirm that *Hox3* is also part of the cluster due to the difficulty in visualizing nascent transcripts for *Hox3* together with *pb* or *Dfd*. Despite these caveats, *Parhyale* provides an excellent arthropod model system to understand these still enigmatic phenomena of Hox gene clustering and spatio-temporal colinearity, and compare the underlying mechanisms to other well-studied vertebrate and invertebrate models [82].

The ParaHox and *NK* gene clusters encode other *ANTP* class homeobox genes closely related to Hox genes [83]. In *Parhyale*, we found 2 caudal (*Cdx*) and 1 *Gsx* ParaHox genes. Compared to hexapods, we identified expansions in some NK-like genes, including 5 Bar homeobox genes (*BarH1/2*), 2 developing brain homeobox genes (*DBX*) and 6 muscle segment homeobox genes (*MSX/Drop*). Evidence from several bilaterian genomes suggests that *NK* genes are clustered together [84–87]. In the current assembly of the *Parhyale* genome, we identified an *NK2-3* gene and an *NK3* gene on the same scaffold (phaw_30.0004720) and the tandem duplication of an *NK2* gene on another scaffold (phaw_30.0004663). Within the *ANTP* class, we also observed 1 mesenchyme homeobox (*Meox*), 1 motor neuron homeobox (*MNX/Exex*) and 3 even-skipped homeobox (*Evx*) genes.

The *Parhyale* genome encodes glycosyl hydrolase enzymes consistent with lignocellulose digestion ("wood eating")

Lignocellulosic (plant) biomass is the most abundant raw material on our planet and holds great promise as a source for the production of bio-fuels [88]. Understanding how some animals and their symbionts achieve lignocellulose digestion is a promising research avenue for exploiting lignocellulose-rich material [89, 90]. Amongst Metazoans, research into the ability to depolymerize plant biomass into useful catabolites is largely restricted to terrestrial species such as ruminants, termites and beetles. These animals rely on mutualistic associations with microbial endosymbionts that provide cellulolytic enzymes known as glycosyl hydrolases (GHs) [91, 92] (Figure 11). Much less studied is lignocellulose digestion in aquatic animals despite the fact that lignocellulose represents a major energy source in aquatic environments, particularly for benthic invertebrates [93]. Recently, it has been suggested that the marine wood-boring Isopod *Limnoria quadripunctata* and the amphipod *Chelura terebrans* may have sterile microbe-free digestive systems and they produce all required enzymes for lignocellulose digestion [30, 31, 94]. Significantly, these species have been shown to have endogenous GH7 family enzymes with cellobiohydrolase (beta-1,4-exoglucanase) activity, previously thought to be absent from animal genomes. From an evolutionary perspective, it is likely that GH7 coding genes were acquired by these species via horizontal gene transfer from a protist symbiont.

Parhyale is a detrivore that can be sustained on a diet of carrots (Figure 11C), suggesting that they too may be able to depolymerize lignocellulose for energy (Figure 11A and B). We searched for GH family genes in *Parhyale* using the classification system of the CAZy (Carbohydrate-Active enZYmes) database [95] and the annotation of protein domains in predicted genes with PFAM [96]. We identified 73 GH genes with complete GH catalytic domains that were classified into 17 families (Supplemental Data 2) including 3 members of the GH7 family. Phylogenetic analysis of *Parhyale* GH7s show high sequence similarity to the known GH7 genes in *L. quadripunctata* and the amphipod *C. terebrans* [31] (Figure 12A; Figure 12, Supplemental Figure 1). GH7 family genes were also identified in the transcriptomes of three more species spanning the multicrustacea clade: *Echinogammarus veneris* (amphipod), *Eucyclops serrulatus* (copepod) and *Calanus finmarchicus* (copepod) (Supplemental Data 2). As previously reported, we also discovered a closely related GH7 gene in the branchiopod *Daphnia* (Figure 12A) [90]. This finding supports the grouping of Branchiopoda with Multicrustacea (rather than with Hexapoda) and the acquisition of a GH7 gene by a vericrustacean ancestor. Alternatively, this suggests an even earlier acquisition of a GH7 gene by a crustacean ancestor with subsequent loss of the GH7 family gene in the lineage leading to insects.

GH families 5, 9, 10, and 45 encode beta-1,4-endoglucanases which are also required for lignocellulose digestion and are commonly found across Metazoa. We found 3 GH9 family genes with complete catalytic domains in the *Parhyale* genome as well as in the other three multicrustacean species (Figure 12B). These GH9 enzymes exhibited a high sequence similarity to their homologues in the isopod *Limnoria* and in a number of termites. Beta-glucosidases are the third class of enzyme required for digestion of lignocellulose. They have been classified into a number of GH families: 1, 3, 5, 9 and 30, with GH1 representing the largest group [95]. In *Parhyale*, we found 7 beta-glucosidases from the GH30 family and 3 from the GH9 family, but none from the GH1 family.

Understanding lignocellulose digestion in animals using complex mutualistic interactions with microbes has proven to be a difficult task. The study of “wood-eating” in *Parhyale* can offer new insights into lignocellulose digestion in the absence of gut microbes, and the unique opportunity to apply molecular genetic approaches to understand the activity of glycosyl hydrolases in the digestive system.

Lignocellulose digestion may also have implications for gut immunity in some crustaceans, since these reactions have been reported to take place in a sterile gut [32, 33].

Characterisation of the innate immune system in a Malacostracan

Immunity research in Malacostracans has attracted interest due to the rapid rise in aquaculture related problems [28, 29, 97]. Malacostracan food crops represent a huge global industry (>$40 Billion at point of first sale), and reliance on this crop as a source of animal protein is likely to increase in line with human population growth [97]. Here we provide an overview of immune-related genes in *Parhyale* that were identified by mapping proteins to the ImmunoDB database [98] (Supplemental Data 2). The ability of the innate immune system to identify pathogen-derived molecules is mediated by pattern recognition receptors (PRRs) [99]. Several groups of invertebrate PRRs have been characterized, i.e. thioester-containing proteins (*TEP*), Toll-like receptors (*TLR*), peptidoglycan recognition proteins (*PGRP*), C-type lectins, galectins, fibrinogen-related proteins (*FREP*), gram-negative binding proteins (*GNBP*), Down Syndrome Cell Adhesion Molecules (*Dscam*) and lipopolysaccharides and beta-1, 3-glucan binding proteins (*LGBP*).

The functions of *PGRPs* have been described in detail in insects like *D. melanogaster* [100] and the PGRP family has also been reported in Vertebrates, Molluscs and Echinoderms [101, 102]. Surprisingly, we found no PGRP genes in the *Parhyale* genome. *PGRPs* were also not found in other sequence datasets from Branchiopoda, Copepoda and Malacostraca (Figure 13A), raising the possibility of their close phylogenetic relationship (like the GH7 genes). In the absence of *PGRPs*, the freshwater crayfish *Pacifastacus leniusculus* relies on a Lysine-type peptidoglycan and serine proteinases, *SPH1* and *SPH2* that forms a complex with *LGBP* during immune response [103]. In *Parhyale*, we found one LGBP gene and two serine proteinases with high sequence identity to *SPH1/2* in *Pacifastacus*. The *D. pulex* genome has also an expanded set of Gram-negative binding proteins (proteins similar to *LGBP*) suggesting a compensatory mechanism for the lost *PGRPs* [104]. Interestingly, we found a putative *PGRP* in the Remipede *Speleonectes tulumensis* (Figure 13A) providing further support for sister group relationship of Remipedia and Hexapoda [14].

Innate immunity in insects is transduced by three major signaling pathways: the Immune Deficiency (*Imd*), Toll and Janus kinase/signal transducer and activator of transcription (*JAK/STAT*) pathways [105, 106]. We found 16 members of the Toll family in *Parhyale* including 10 Toll-like receptors (TLRs) (Figure 13B). Some TLRs have been also implicated in embryonic tissue morphogenesis in *Parhyale* and other arthropods [107]. Additionally, we identified 7 Imd and 25 JAK/STAT pathway members including two negative regulators: suppressor of cytokine signaling (*SOCS*), and protein inhibitor of activated *STAT* (*PIAS*) [108].

The blood of arthropods (hemolymph) contains hemocyanin which is a copper-binding protein involved in the transport of oxygen, and circulating blood cells called hemocytes for the phagocytosis of pathogens. Phagocytosis by hemocytes is facilitated by the evolutionarily conserved gene family, the thioester-containing proteins (*TEPs*) [109]. Previously sequenced Pancrustacean species contained between 2 to 52 *TEPs*. We find 5 *TEPs* in the *Parhyale* genome. Arthropod hemocyanins themselves are structurally related to phenoloxidases (PO; [110]) and can be converted into POs by conformational changes under specific conditions [111]. POs are involved in several biological processes (like the melanization immune response, wound healing and cuticle sclerotization) and we identified 7 PO genes in *Parhyale*. Interestingly, hemocyanins and PO activity have been shown to be highly abundant together with glycosyl hydrolases in the digestive system of Isopods and Amphipods, raising a potential mechanistic link between gut sterility and degradation of lignocellulose [30, 33].

Another well-studied transmembrane protein essential for neuronal wiring and adaptive immune responses in insects is the immunoglobulin (*Ig*)-superfamily receptor Down syndrome cell adhesion molecule (*Dscam*) [112, 113]. Alternative splicing of *Dscam* transcripts can result in thousands of different isoforms that have a common architecture but have sequence variations encoded by blocks of alternative spliced exons. The *D. melanogaster Dscam* locus encodes 12 alternative forms of exon 4 (encoding the N-terminal half of Ig2), 48 alternative forms of exon 6 (encoding the N-terminal half of Ig3), 33 alternative forms of exon 9 (encoding Ig7), and 2 alternative forms of exon 17 (encoding transmembrane domains) resulting in a total of 38,016 possible combinations. The *Dscam* locus in *Parhyale* (and in other crustaceans analysed) has a similar organization to insects; tandem arrays of multiple exons encode the N-terminal halves of Ig2 (exon 4 array with at least 13 variants) and Ig3 (exon 6 array with at least 20 variants) and the entire Ig7 domain (exon 14 array with at least 13 variants) resulting in at least 3,380 possible combinations (Figure 13C-E). The alternative splicing of hypervariable exons in *Parhyale* was confirmed by sequencing of cDNA clones amplified with Dscam-specific primers. Almost the entire *Dscam* gene is represented in a single genomic scaffold and exhibits high amino-acid sequence conservation with other crustacean *Dscams* (Figure 13, Supplemental Figure 1). The number of *Dscam* isoforms predicted in *Parhyale* is similar to that predicted for Daphnia species [114]. It remains an open question whether the higher number of isoforms observed in insects coincides with the evolution of additional Dscam functions compared to crustaceans.

From a functional genomics perspective, the *Parhyale* immune system appears to be a good representative of the malacostrocan or even multicrustacean clade that can be studied in detail with existing tools and resources.

Non-coding RNAs and associated proteins in the *Parhyale* genome

Non-coding RNAs are a central, but still a relatively poorly understood part of eukaryotic genomes. In animal genomes, different classes of small RNAs are key for genome surveillance, host defense against viruses and parasitic elements in the genome, and regulation of gene expression through transcriptional, post-transcriptional and epigenetic control mechanisms [115–123]. The nature of these non-coding RNAs, as well as the proteins involved in their biogenesis and function, can vary between animals. For example, some nematodes have Piwi-interacting short RNAs (piRNAs), while others have replaced these by alternate small RNA based mechanisms to compensate for their loss [124].

As a first step, we surveyed the *Parhyale* genome for known conserved protein components of the small interfering RNA (siRNA/RNAi) and the piRNA pathways (Table 4). We found key components of all major small RNA pathways, including 4 argonaute family members, 2 PIWI family members, and orthologs of *D. melanogaster Dicer-1* and *Dicer-2*, *drosha* and *loquacious*, (Figure 14, Supplemental Figure 1). Among Argonaute genes, *Parhyale* has 1 *AGO-1* ortholog and 3 *AGO-2* orthologs, which is presumably a malacostraca-specific expansion. While *Parhyale* only has 2 PIWI family members, other crustacean lineages have clearly undergone independent expansions of this protein family. Unlike in *C. elegans*, many mammals, fish and insects (but not *D. melanogaster*), we did not find any evidence in the *Parhyale* genome for the *SID-1* (systemic RNA interference defective) transmembrane protein that is essential for systemic RNAi [125–127]. Species without a *SID-1* ortholog can silence genes only in a cell-autonomous manner [128]. This feature has important implications for future design of RNAi experiments in *Parhyale*.

We also assessed the miRNA and putative long non-coding RNAs (lncRNA) content of *Parhyale* using both MiRPara and Rfam [129, 130]. We annotated 1405 homologues of known non-coding RNAs using Rfam. This includes 980 predicted tRNAs, 45 rRNA of the large ribosomal subunit, 10 rRNA of the small ribosomal subunit, 175 snRNA components of the major spliceosome (U1, U2, U4, U5 and U6), 5 snRNA components of the minor spliceosome (U11, U12, U4atac and U6atac), 43 ribozymes, 38 snoRNAs, 71 conserved cis-regulatory element derived RNAs and 42 highly conserved miRNA genes (Supplemental Data 5; Supplemental Data 11). *Parhyale* long non-coding RNAs (lncRNAs) were identified from the transcriptome using a series of filters to remove coding transcripts producing a list of 220,284 putative lncRNAs (32,223 of which are multi-exonic). Only one *Parhyale* lncRNA has clear homology to another annotated lncRNA, the sphinx lncRNA from *D. melanogaster* [131].

We then performed a more exhaustive search for miRNAs using MiRPara (Supplemental Data 11) and a previously published *Parhyale* small RNA read dataset [132]. We identified 1,403 potential miRNA precursors represented by 100 or more reads. Combining MiRPara and Rfam results, we annotated 31 out of the 34 miRNA families found in all Bilateria, 12 miRNAs specific to Protostomia, 4 miRNAs specific to Arthropoda and 5 miRNAs previously found to be specific to Mandibulata (Figure 14). We did not identify *mir-125, mir-283* and *mir-1993* in the *Parhyale* genome. The absence of *mir-1993* is consistent with reports that this miRNA was lost during Arthropod evolution [133]. While we did not identify *mir-125*, we observed that *mir-100* and *let-7* occurred in a cluster on the same scaffold (Figure 14, Supplemental Figure 2), where *mir-125* is also present in other animals. The absence of *mir-125* has been also reported for the centipede genome [76]. *mir-100* is one of the most primitive miRNAs shared by Bilateria and Cnidaria [133, 134]. The distance between *mir-100* and *let-7* genes within the cluster can vary substantially between different species. Both genes in *Parhyale* are localized within a 9.3kb region (Figure 14, Supplemental Figure 2) as compared to 3.8kb in the mosquito *Anopheles gambiae* and 100bp in the beetle *Tribolium* [135]. Similar to *D. melanogaster* and the polychaete *Platynereis dumerilii*, we found that *Parhyale mir-100* and let-7 are co-transcribed as a single, polycistronic lncRNA. We also found another cluster with *miR-71* and *mir-2* family members which is conserved across many invertebrates [136] (Figure 14, Supplemental Figure 2).

Conserved linkages have also been observed between miRNAs and Hox genes in Bilateria [137–141]. For example, the phylogenetically conserved *mir-10* is present within both vertebrate and invertebrate Hox clusters between Hoxb4/*Dfd* and *Hoxb5/Scr* [142]. In the *Parhyale* genome and Hox BAC sequences, we found that *mir-10* is also located between *Dfd* and *Src* on BAC clone PA179-K23 and scaffold phaw_30.0001203 (Figure 14, Supplemental Figure 2). However, we could not detect *mir-iab-4* near the *Ubx* and *AbdA* genes in *Parhyale*, the location where it is found in other arthropods/insects [143]. Preliminary evidence regarding the presence of PIWI proteins and other piRNA pathway proteins also suggests that the piRNA pathway is likely active in *Parhyale*, although piRNAs themselves await to be surveyed. The opportunity to study these piRNA, miRNA and siRNA pathways in a genetically tractable crustacean system will shed further light into the regulation and evolution of these pathways and their contribution to morphological diversity.

Methylome analysis of the *Parhyale* genome

Methylation of cytosine residues (m5C) in CpG dinucleotides in animal genomes is regulated by a conserved multi-family group of DNA methyltransferases (DNMTs) with diverse roles in the epigenetic control of gene expression, genome stability and chromosome dynamics [144–146]. The phylogenetic distribution of DNMTs in Metazoa suggests that the bilaterian ancestor had at least one member of the Dnmt1 and Dnmt3 families (involved in *de novo* methylation and maintenance of DNA methylation) and the Dnmt2 family (involved in tRNA methylation), as well as additional RNA methyltransferases [147, 148]. Many animal groups have lost some of these DNA methyltransferases, for example *DNMT1* and 3 are absent from *D. melanogaster* and flatworms [149, 150], while *DNMT2* is absent from nematodes *C. elegans* and *C. briggsae*. The *Parhyale* genome encodes members of all 3 families *DNMT1, DNMT3* and *DNMT2*, as well as 2 orthologs of conserved methyl-CpG-binding proteins and a single orthologue of *Tet2*, an enzyme involved in DNA demethylation [151] (Figure 15A).

We used genome wide bisulfite sequencing to confirm the presence and also assess the distribution of CpG dinucleotide methylation. Our results indicated that 20-30% of *Parhyale* DNA is methylated at CpG dinucleotides (Figure 15B). The *Parhyale* methylation pattern is similar to that observed in vertebrates, with high levels of methylation detected in transposable elements and other repetitive elements, in promoters and gene bodies (Figure 15C). A particular class of rolling-circle transposons are very highly methylated in the genome, potentially implicating methylation in silencing these elements. For comparison, about 1% or less of CpG-associated cytosines are methylated in insects like *Drosophila, Apis, Bombyx and Tribolium*. [144, 152, 153]. These data represent the first documentation of a crustacean methylome. Considering the utility of *Parhyale* for genetic and genomic research, we anticipate future investigations to shed light on the functional importance and spatiotemporal dynamics of epigenetic modifications during normal development and regeneration, as well as their relevance to equivalent processes in vertebrate systems.

Parhyale **genome editing using homology-independent approaches**

Parhyale has already emerged as a powerful model for developmental genetic research where the expression and function of genes can be studied in the context of stereotyped cellular processes and with a single-cell resolution. Several experimental approaches and standardized resources have been established to study coding and non-coding sequences (Table 1). These functional studies will be enhanced by the availability of the assembled and annotated genome presented here. As a first application of these resources, we tested the efficiency of the CRISPR/Cas system for targeted genome editing in *Parhyale* [17–22]. In these studies, we targeted the *Distal-less* patterning gene (called *PhDll-e*) [24] that has a widely-conserved and highly-specific role in animal limb development [154].

We first genotyped our wild-type laboratory culture and found two *PhDll-e* alleles with 23 SNPs and 1 indel in their coding sequences and untranslated regions. For *PhDll-e* knock-out, two sgRNAs targeting both alleles in their coding sequences downstream of the start codon and upstream of the DNA-binding homeodomain were injected individually into 1-cell-stage embryos (G0 generation) together with a transient source of Cas9 (Figure 16, Supplemental Figure 1 A-B). Both sgRNAs gave rise to animals with truncated limbs (Figure 16A and B); the first sgRNA at a relatively low percentage around 9% and the second one at very high frequencies ranging between 53% and 76% (Figure 16, Supplemental Figure 1). Genotyping experiments revealed that injected embryos carried *PhDll-e* alleles modified at the site targeted by each sgRNA (Figure 16, Supplemental Figure 1 B-D). The number of modified *PhDll-e* alleles recovered from G0s varied from two, in cases of early bi-allelic editing at the 1-cell-stage, to three or more, in cases of later-stage modifications by Cas9 (Figure 16, Supplemental Figure 1 C). We isolated indels of varying length that were either disrupting the open reading frame, likely producing loss-of-function alleles or were introducing in-frame mutations potentially representing functional alleles (Figure 16, Supplemental Figure 1 C-D). In one experiment with the most efficient sgRNA, we raised the injected animals to adulthood and set pairwise crosses between 17 fertile G0s (10 male and 7 female): 88% (15/17) of these founders gave rise to G1 offspring with truncated limbs, presumably by transmitting *PhDll-e* alleles modified by Cas9 in their germlines. We tested this by genotyping individual G1s from two of these crosses and found that embryos bearing truncated limbs were homozygous for loss-of-function alleles with out-of-frame deletions, while their wild-type siblings carried one loss-of-function allele and one functional allele with an in-frame deletion (Figure 16, Supplemental Figure 1 D).

The non-homologous end joining (NHEJ) repair mechanism operating in the injected cells can be exploited not only for gene knock-out experiments described above, but also for CRISPR knock-in approaches where an exogenous DNA molecule is inserted into the targeted locus in a homology-independent manner. This homology-independent approach could be particularly useful for *Parhyale* that exhibits high levels of heterozygosity and polymorphisms in the targeted laboratory populations, especially in introns and intergenic regions. To this end, we co-injected into 1-cell-stage embryos the Cas9 protein together with the strongest sgRNA and a tagging plasmid. The plasmid was designed in such a way that upon its linearization by the same sgRNA and Cas9 and its integration into the *PhDll-e* locus in the appropriate orientation and open reading frame, it would restore the endogenous *PhDll-e* coding sequence in a bicistronic mRNA also expressing a nuclear fluorescent reporter. Among injected G0s, about 7% exhibited a nuclear fluorescence signal in the distal (telopodite and exopodite) parts of developing appendages (Figure 16C and Figure 16, Supplemental Figure 1 E), which are the limb segments that were missing in the knock-out experiments (Figure 16B). Genotyping of one of these embryos demonstrated that the tagged *PhDll-e* locus was indeed encoding a functional *PhDll-e* protein with a small in-frame deletion around the targeted region (Figure 16, Supplemental Figure 1 F).

These results, together with the other recent applications of the CRISPR/Cas system to study Hox genes in *Parhyale* [18, 19], demonstrate that the ability to manipulate the fertilized eggs together with the slow tempo of early cleavages can result in very high targeting frequencies and low levels of mosaicism for both knock-out and knock-in approaches. Considering the usefulness of the genome-wide resources described in this report, we anticipate that the *Parhyale* embryo will prove an extremely powerful system for fast and reliable G0 screens of gene expression and function.

CONCLUSION

In this article we described the first complete genome of a malacostracan crustacean species, the genome of the marine amphipod *Parhyale hawaiensis*. At an estimated size of 3.6 Gb, it is among the largest genomes submitted to NCBI. The *Parhyale* genome reported here is that of a single adult male from a sib-bred line called Chicago-F. We find *Parhyale* has an abundance of repetitive sequence and high levels of heterozygosity in the individual sequenced. Combined with analysis of available transcriptome sequences and independently sequenced genomic BAC clones, we conclude high levels of heterozygosity are representative of high levels of single and polynucleotide polymorphisms in the broader laboratory population. Our comparative bioinformatics analyses suggest that the expansion of repetitive sequences and the increase in gene size due to an expansion of intron size have contributed to the large size of the genome. Despite these challenges, the *Parhyale* genome and associated transcriptomic resources reported here provide a useful assembly of most genic regions in the genome and a comprehensive description of the *Parhyale* transcriptome and proteome.

Parhyale has emerged since the early 2000's as an attractive animal model for developmental genetic and molecular cell biology research. It fulfills several desirable biological and technical requirements as an experimental model, including a relatively short life-cycle, year-round breeding under standardized laboratory conditions, availability of thousands of eggs for experimentation on a daily basis, and amenability to various embryological, cellular, molecular genetic and genomic approaches. In addition, *Parhyale* has stereotyped cell lineages and cell behaviors, a direct mode of development, a remarkable appendage diversity and the capacity to regenerate limbs post-embryonically. These qualities can be utilized to address fundamental long-standing questions in developmental biology, like cell fate specification, nervous system development, organ morphogenesis and regeneration [155]. Research on these topics will benefit enormously from the standardized genome-wide resources reported here.

Forward and reverse genetic analyses using both unbiased screens and candidate gene approaches have already been devised successfully in *Parhyale* (Table 1). The availability of coding and non-coding sequences for all identified signaling pathway components, transcription factors and various classes of non-coding RNAs will dramatically accelerate the study of the expression and function of genes implicated in the aforementioned processes.

Equally importantly, our analyses highlight additional areas where *Parhyale* could serve as a new experimental model to address other questions of broad biomedical interest. From a functional genomics perspective, the *Parhyale* immune system appears to be a good representative of the malacostracan or even the multicrustacean clade that can be studied in detail with existing tools and resources. Besides the evolutionary implications and the characterization of alternative strategies used by arthropods to defend against pathogens, a deeper mechanistic understanding of the *Parhyale* immune system will be relevant to aquaculture. Some of the greatest setbacks in the crustacean farming industry are caused by severe disease outbreaks. *Parhyale* is closely related to farmed crustaceans (primarily shrimps, prawns and crayfish) and the knowledge acquired from studying its innate immunity could help enhance the sustainability of this industry by preventing or controlling infectious diseases [97, 156–159].

An immune-related problem that will be also interesting to explore in *Parhyale* concerns the possibility of a sterile digestive tract similar to that proposed for limnoriid Isopods [30]. *Parhyale*, like limnoriid Isopods, encodes and expresses all enzymes required for lignocellulose digestion, suggesting that it is able to "digest wood" by itself without symbiotic microbial partners. Of course, a lot of work still needs to be invested in the characterization of the cellulolytic system in *Parhyale* before any comparisons can be made with other well-established symbiotic digestion systems of lignocellulose. Nevertheless, the possibility of an experimentally tractable animal model that serves as a living bioreactor to convert lignocellulose into simpler metabolites, suggests that future research in *Parhyale* may also have a strong biotechnological potential, especially for the production of biofuels from the most abundant and cheapest raw material, plant biomass.

Although more high-quality genomes with a broader phylogenetic coverage are still needed for meaningful evolutionary comparisons, our observations from analysing the *Parhyale* genome and other crustacean data sets also contribute to the ongoing debate on the relationships between crustacean groups. While the analysis of shared orthologous groups did not provide clear support for either the Allotriocarida hypothesis (uniting Branchiopoda with Hexapoda) or the Vericrustacea hypothesis (uniting Branchiopoda with Malacostraca), we noted the presence of GH7 genes and the absence of PGRP genes in branchiopod and multicrustacean genomes supporting the Vericrustacea hypothesis. It still remains to be proven how reliable these two characters will be to distinguish between these alternative phylogenetic affinities.

Finally, *Parhyale* was introduced recently as a new model for limb regeneration [26]. In some respects, including the segmented body plan, the presence of a blood system and the contribution of lineage-committed adult stem cells to newly formed tissues, regeneration in *Parhyale* may resemble the process in vertebrates more than other established invertebrate models (e.g. planarians, hydra). Regenerative research in *Parhyale* has been founded on transgenic approaches to label specific populations of cells and will be further assisted by the resources presented here. Likewise, we expect that the new genomic information and CRISPR-based genome editing methodologies together with all other facets of *Parhyale* biology will open other new research avenues not yet imagined.

ACKNOWLEDGMENTS

We are grateful to Serge Picard for sequencing the genome libraries, and Frantisek Marec and Peer Martin for useful advice on *Parhyale* karyotyping.

MATERIALS AND METHODS

Raw genomic reads are deposited at NCBI with the project accession: PRJNA306836. A list of software and external datasets used are provided in Supplemental Data 1. Detailed methodology and codes for each section are provided as supplementary IPython notebooks in HTML format viewable with a web browser. All supplemental data including IPython notebook can be downloaded from this figshare link:
`https://figshare.com/articles/supplemental_data_for_Parhyale_hawaniensis_genome/3498104`

Alternatively, the IPython notebooks can also be viewed at the following github repository:
`https://github.com/damiankao/phaw_genome`

Genome library preparation and sequencing

About 10 µg of genomic DNA were isolated from a single adult male from the Chicago-F isofemale line established in 2001 [51]. The animal was starved for one week and treated for 3 days with penicillin-streptomycin (100x, Gibco/Thermo Fisher Scientific), tetracycline hydrochloride (20 µg/ml, Sigma-Aldrich) and amphotericin B (200x, Gibco/Thermo Fisher Scientific). It was then flash frozen in liquid nitrogen, homogenized manually with a pestle in a 1.5 ml microtube (Kimble Kontes) in 600 µl of Lysis buffer (100 mM Tris-HCl pH 8, 100 mM NaCl, 50 mM EDTA, 0.5% SDS, 200 µg/ml Proteinase K, 20 µg/ml RNAse A). The lysate was incubated for 3 hours at 37°C, followed by phenol/chloroform extractions and ethanol precipitation. The condensed genomic DNA was fished out with a Pasteur pipette, washed in 70% ethanol, air-dried, resuspended in nuclease-free water and analysed on a Qubit fluorometer (Thermo Fisher Scientific) and on a Bioanalyzer (Agilent Technologies). All genome libraries were prepared from this sample: 1 µg of genomic DNA was used to generate the shotgun libraries using the TruSeq DNA Sample Prep kit (Illumina) combined with size-selection on a LabChip XT fractionation system (Caliper Life Sciences Inc) to yield 2 shotgun libraries with average fragment sizes 431 bp and 432 bp, respectively; 4 µg of genomic DNA were used to generate 4 mate-pair libraries with average fragment sizes 5.5 kb, 7.3 kb, 9.3 kb and 13.8 kb using the Nextera Mate Pair Sample Preparation kit (Illumina) combined with agarose size selection. All libraries were sequenced on a HiSeq

2500 instrument (Illumina) using paired-end 150 nt reads.

Karyotyping

For chromosome spreads, tissue was obtained from embryos at stages 14-18 [35]. Eggs were taken from the mother and incubated for 1–2 h in isotonic colchicine solution (0.05% colchicine, artificial sea water). After colchicine incubation, the embryonic tissue was dissected from the egg and placed in hypotonic solution (0.075 M KCl) for 25 min. For tissue fixation, we replaced the hypotonic solution with freshly prepared ice-chilled Carnoy's fixative (six parts ethanol, three parts methanol and one part anhydrous acetic acid) for 25 min. The fixed tissue was minced with a pair of fine tungsten needles in Carnoy's solution and the resulting cell suspension was dropped with a siliconized Pasteur pipette from a height of about 5 cm onto a carefully cleaned ice-chilled microscopic slide. After partial evaporation of the Carnoy's fixative the slides were briefly exposed a few times to hot water vapors to rehydrate the tissue. The slides were then dried on a 75°C metal block in a water bath. Finally, the slides with prepared chromosomes were aged overnight at 60°C. After DNA staining either with Hoechst (H33342, Molecular Probes) or with DAPI (Invitrogen), chromosomes were counted on a Zeiss Axioplan II Imaging equipped with C-Apochromat 63x/1.2 NA objective and a PCO pixelfly camera. FIJI was used to improve image quality (contrast and brightness) and FIJI plugin 'Cell Counter' was used to determine the number of chromosomes.

Analysis of polymorphism and repetitiveness

The *Parhyale* raw data and assembled data are available on the NCBI website. Genome assembly was done with Abyss [160] at two different k-mer settings (70, 120) and merged with GAM-NGS. Scaffolding was performed with SSPACE [161]. We chose cut-offs of >95% overlap length and >95% identity when removing shorter allelic contigs before scaffolding as these gave better scaffolding results as assessed by assembly metrics. Transcriptome assembly was performed with Trinity [57]. The completeness of the genome and transcriptome was assessed by blasting against CEGMA genes [58] and visualized by plotting the orthologue hit ratio versus e-value. K-mer analysis of variant and repetitive branching was performed with String Graph Assembler's preqc module [54]. K-mer intersection analysis was performed using jellyfish2 [162]. An in-depth description of the assembly process is detailed in Supplemental Data 6.

Transcriptome library preparation, sequencing and assembly

Parhyale transcriptome assembly was generated from Illumina reads collected from diverse embryonic stages (Stages 19, 20, 22, 23, 25, and 28), and adult thoracic limbs and regenerating thoracic limbs (3 and 6 days post amputation). For the embryonic samples, RNA was extracted using Trizol; PolyA+ libraries were prepared with the Truseq V1 kit (Illumina), starting with 0.6 - 3.5ug of total mRNA, and sequenced on the Illumina Hiseq 2000 as paired-end 100 base reads, at the QB3 Vincent J. Coates Genomics Sequencing Laboratory. For the limb samples, RNA was extracted using Trizol; PolyA+ libraries were prepared with the Truseq V2 kit (Illumina), starting with 1ug of total mRNA, and sequenced on the Illumina Hiseq 2500 as paired-end 100 base reads, at the IGBMC Microarray and Sequencing platform. 260 million reads from embryos and 180 million reads from limbs were used for the transcriptome assembly. Prior to the assembly we trimmed adapter and index sequences using cutadapt [163]. We also removed spliced leader sequences: GAATTTTCACTGTTCCCTTTACCACGTTTTACTG, TTACCAATCACCCCTTTACCAAGCGTTTACTG, CCCTTTACCAACTCTTAACTG, CCCTTTACCAACTTTACTG using cutadapt with 0.2 error allowance to remove all potential variants [164]. To assemble the transcriptome we used Trinity (version trinityrnaseq_r20140413) [57] with settings: -min_kmer_cov 2, -path_reinforcement_distance 50.

Gene model prediction and canonical proteome dataset generation

Gene prediction was done with a combination of Evidence Modeler [165] and Augustus [166]. The transcriptome was first mapped to the genome using GMAP [167]. A secondary transcriptome reference assembly was performed with STAR/Cufflinks [168, 169]. The transcriptome mapping and Cufflinks assembly was processed through the PASA pipeline [165] to consolidate the annotations. The PASA dataset, a set of Exonerate [170] mapped Uniprot proteins, and Ab inito GeneMark [171] predictions were consolidated with Evidence Modeler to produce a set of gene annotations. A high confidence set of gene models from Evidence Modeler containing evidence from all three sources was used to train Augustus. Evidence from RepeatMasker [172], PASA and Exonerate were then used to generate Augustus gene predictions. A final list of genes for down-stream analysis was generated using both transcriptome and gene predictions (canonical proteome dataset). Detailed methods are described in Supplemental Data 8.

Polymorphism analysis on genic regions and BAC clones

For variant analysis on the BAC clones, the short shot-gun library genomic reads were mapped to the BAC clones individually. GATK was then used to call variants. For variant analysis on the genic regions, transcript sequences used to generate the canonical proteome dataset were first aligned to the genome assembly. Genome alignments of less than 30 base pairs were discarded. The possible genome alignments were sorted based on number of mismatches with the top alignment having the least amount of mismatches. For each transcript, the top two genome aligments were used to call potential variants. Trascripts or parts of transcripts where there were more than five genomic mapping loci were discarded as potentially highly conserved domains or repetitive regions. Detailed methods of this process are described in Supplemental Data 9.

Polymorphisms in *Parhyale* developmental genes

Parhyale genes (nucleotide sequences) were downloaded from GenBank. Each gene was used as a query for blastn against the *Parhyale* genome using the Geneious software [173]. In each case two reference contig hits were observed where both had E values of close to zero. A new sequence called geneX_snp was created and this sequence was annotated with the snps and/or indels present in the alternative genomic contigs. To determine the occurrence of synonymous and non-synonymous substitutions, the original query and the newly created sequence (with polymorphisms annotated) were in silico translated into protein sequences followed by pairwise alignment. Regions showing amino acid changes were annotated as non-synonymous substitutions. Five random genes from the catalogue were selected for PCR, cloning and Sanger sequencing to confirm genomic polymorphisms and assess further polymorphism in the lab popultaion. Primers for genomic PCR designed to capture and amplify exon regions are listed as the following: dachshund (PH1F = 5'- GGTGCGCTAAATTGAAGAAATTACG-3' and PH1R = 5'- ACTCAGAGGGTAATAGTAACAGAA-3'), distalless exon 2 (PH2F = 5'-CACGGCCCGGCACTAACTATCTC-3' and PH2R = 5'-GTAATATATCTTACAACAACGACTGAC-3'), distalless exon 3 (PH3F = 5'-GGTGAACGGGCCGGAGTCTC-3' and PH3R = 5'-GCTGTGGGTGCTGTGGGT-3'), homothorax (PH4F = 5'-TCGGGGTGTAAAAAGGACTCTG-3' and PH4R = 5'-AACATAGGAACTCACCTGGTGC-3'), orthodenticle (PH5F = 5'-TTTGCCACTAACACATATTTCGAAA-3' and PH5R = 5'-TCCCAAGTAGATGATCCCTGGAT-3') and prospero (PH6F = 5'-TACACTGCAACATCCGATGACTTA-3' and PH6R = 5'-CGTGTTATGTTCTCTCGTGGCTTC-3').

Evolutionary analyses of orthologous groups

Evolutionary analyses and comparative genomics were performed with 16 species: *D. melanogaster, A. gambiae, D. pulex, L. salmonis, S.maritima, S. mimosarum, M. martensii, I. scapularis, H. dujardini, C. elegans, B. malayi, T. spiralis, M. musculus, H. sapiens*, and *B. floridae*. For orthologous group analyses, gene families were identified using OrthoFinder [59]. The canonical proteome was used as a query in BlastP against proteomes from 16 species to generate a distance matrix for OrthoFinder to normalize and then cluster with MCL. Detailed methods are described in Supplemental Data 10. For the comparative BLAST analysis, five additional transcriptome datasets were used from the following crustacean species: *Litopenaeus vannamei, Echinogammarus veneris, Eucyclops serrulatus, Calanus finmarchicus, Speleonectes tulumensis*.

Fluorescence in situ hybridization detection of Hox genes

Embryo fixation and in-situ hybridization was performed according to [40]. To enhance the nascent nuclear signal over mature cytoplasmic transcript, we used either early germband embryos (Stages 11 – 15) in which expression of *lab*, *Dfd*, and *Scr* are just starting [18], or probes that contain almost exclusively intron sequence (*Ubx, abd-A, Abd-B, and en1). Lab, Dfd*, and *Scr* probes are described in [18]. Template for the intron-spanning probes were amplified using the following primers: en1-Intron1, AAGACACGACGAGCATCCTG and CTGTGTATGGCTACCCGTCC; Ubx-Intron1, GGTATGACAGCCGTCCAACA and AGAGTGCCAAGGATACCCGA; abd-A, CGATATACCCAGTCCGGTGC and TCATCAGCGAGGGCACAATT; Abd-B, GCTGCAGGATATCCACACGA and TGCAGTTGCCGCCATAGTAA.

A T7-adapter was appended to the 5' end of each reverse primer to enable direct transcription from PCR product. Probes were labeled with either Digoxigenin (DIG) or Dinitrophenol (DNP) conjugated UTPs, and visualized using sheep α-DIG (Roche) and donkey α-Sheep AlexaFluor 555 (Thermo Fischer Scientific), or Rabbit α-DNP (Thermo Fischer Scientific) and Donkey α -Rabbit AlexaFluor 488 (Jackson ImmunoResearch), respectively. Preparations were imaged on an LSM 780 scanning laser confocal (Zeiss), and processed using Volocity software (Perkin-Elmer).

Cross species identification of GH family genes and immune-related genes

The identification of GH family genes was done by obtaining Pfam annotations [96] for the *Parhyale* canonical proteome. Pfam domains were classified into different GH families based on the CAZy database [95]. For immune-related genes, best-reciprocal blast was performed with ImmunoDB genes [98].

Phylogenetic tree construction

Multiple sequence alignments of protein sequences for gene families of *FGF, FGFR, CERS, GH7, GH9, PGRP*, Toll-like receptors, *DICER*, Piwi and Argonaute were performed using MUSCLE [174]. Phylogenetic tree construction was performed with RAxML [175] using the WAG+G model from MUSCLE multiple alignments.

Bisulfite sequencing

Libraries for DNA methylation analysis by bisulfite sequencing were constructed from 100ng of genomic DNA extracted from one *Parhyale* male individual, using the Illumina Truseq DNA methylation kit according to manufacturers instructions. Alignments to the *Parhyale* genome were generated using the core Bismark module from the program Bismark [176], having first artificially joined the *Parhyale* contigs to generate 10 pseudo-contigs as the program is limited as to the number of separate contigs it can analyse. We then generated genome-wide cytosine coverage maps using the bismark_methylation_extraction module with the parameter –CX specified to generate annotations of CG, CHH and CHG sites. In order to analyse genome-wide methylation patterns only cytosines with more than a 10 read depth of coverage were selected. Overall methylation levels at CG, CHH and CHG sites were generated using a custom Perl script. To analyse which regions were methylated we mapped back from the joined contigs to the original contigs and assigned these to functional regions based on RepeatMasker [172] and transcript annotations of repeats and genes respectively. To generate overall plots of methylation levels in different features we averaged over all sites mapping to particular features, focusing on CG methylation and measuring the %methylation at each site as the number of reads showing methylation divided by the total number of reads covering the site. Meta gene plots over particular features were generated similarly except that sites mapping within a series of 100bp wide bins from 1000bp upstream of the feature start site and onward were collated.

Identification and cloning of Dscam alternative spliced variants

For the identification of *Dscam* in the *Parhyale*, we used the Dscam protein sequence from crustaceans *D. pulex* [114] and *L. vannamei* [177] as queries to probe the assembled genome using tBlastN. A 300kb region on scaffold phaw_30.0003392 was found corresponding to the *Parhyale Dscam* extending from IG1 to FN6 exons. This sequence was annotated using transcriptome data together with manual searches for open reading frames to identify IG, FN exons and exon-intron boundaries (Figure 13 supplemental figure 1). Hypervariable regions of IG2, IG3 and IG7 were also annotated accordingly on the scaffold (Figure 13 supplemental figure 1). This region represents a bona fide *Dscam* paralog as it matches the canonical extracellular *Dscam* domain structure of nine IGs – four FNs – one IG and two FNs. *Parhyale* mRNA extractions were performed using the Zymo Research Direct-zol RNA MiniPrep kit according to manufacturer's instructions. Total RNA extract was used for cDNA synthesis using the Qiagen QuantiTect Reverse Transcription Kit according to manufacturer's instructions. To identify and confirm potential hypervariable regions from the *Parhyale* (Ph-Dscam) transcript, three regions of Ph-Dscam corresponding to IG2, IG3 and IG7 exons respectively were amplified using the following primer pairs. IG2 region:

DF1 = 5'-CCCTCGTGTTCCCGCCCTTCAAC-3'

DR1 = 5'-GCGATGTGCAGCTCTCCAGAGGG-3'

IG3 region:

DF2 = 5'-TCTGGAGAGCTGCACATCGCTAAT-3'

DR2 = 5'-GTGGTCATTGCGTACGAAGCACTG-3'

IG7 region:

DF3 = 5'-CGGATACCCCATCGACTCCATCG-3'

DR3 = 5'-GAAGCCGTCAGCCTTGCATTCAA-3'

PCR of each region was performed using Phusion High-fidelity polymerase from Thermo Fisher Scientific and thermal cycling was done as the following: 98°C 30s, followed by 30 cycles of 98°C 10s, 67°C 30s, 72°C 1m30s, and then 72°C 5m. PCR products were cloned into pGEMT-Easy vector and a total of 81 clones were selected and Sanger sequenced and in silico translated in the correct reading frame using Geneious (R7; [173] for multiple sequence alignment.

Identification of non-protein-coding RNAs

Parhyale non-protein-coding RNAs were identified using two independent approaches. Infernal 1.1.1 [178] was used with the RFAM 12.0 database [130] to scan the genome to identify potential non-coding RNAs. Additionally, MiRPara [129] was used to scan the genome for potential miRNA precursors. These potential precursors were further filtered using small RNA read mapping and miRBase mapping [179]. Putative lncRNAs were identified from the transcriptome by applying filtering criteria including removal of known and predicted coding RNAs. Detailed methods are available in Supplementary Data 11.

CRISPR/Cas genome editing

To genotype our wild-type population, extraction of total RNA and preparation of cDNA from embryos were carried out as previously described [25]. The PhDll-e cDNA was amplified with primers PhDlle_2For (5'-TTTGTCAGGGATCTGCCATT-3') and PhDlle_1852Rev (5'-TAGCGGCTGACGGTTGTTAC-3'), purified with the DNA Clean and Concentrator kit (Zymo Research), cloned with the Zero Blunt TOPO PCR Cloning Kit (Thermo Fisher Scientific) and sequenced with primers M13 forward (5'- GTAAAACGACGGCCAG-3') and M13 reverse (5'-CAGGAAACAGCTATGAC-3').

Each template for sgRNA synthesis was prepared by annealing and PCR amplification of the sgRNA-specific forward primer Dll1: (18 nt PhDll-e-targeted sequence underlined)
5'-GAAATTAATACGACTCACTATA
AGAGTTGTTACCAAAGAAGTTTTAGAGCTAGAAATAGC-3'
or Dll2: (20 nt PhDll-e-targeted sequence underlined)
5'-GAAATTAATACGACTCACTAT
AGGCTTCCCCGCCGCCATGTAGTTTTAGAGCTAGAAATAGC-3'
together with the universal reverse primer:
5'-AAAAGCACCGACTCGGTGCCACTTTTTCAAGTTGATAA
CGGACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAAC-3'
using the Phusion DNA polymerase (New England Biolabs).

Each PCR product was gel-purified with the Zymoclean DNA recovery kit (Zymo Research) and 150 ng of DNA were used as template in an in vitro transcription reaction with the Megashortscript T7 kit (Thermo Fisher Scientific). A 4-hour incubation at 37°C was followed by DNAse digestion, phenol/chloroform extraction, ethanol precipitation and storage in ethanol at -20°C according to the manufacturer's instructions. Before microinjection, a small aliquot of the sgRNA was centrifuged, the pellet was washed with 70% ethanol, resuspended in nuclease-free water and quantified on a Nanodrop spectrophotometer (Thermo Scientific). The Cas9 was provided either as in vitro synthesized caped mRNA or as recombinant protein. Cas9 mRNA synthesis was carried out as previously described [45] using plasmid T7-Cas9 (a gift from David Stern and Justin Crocker) linearized with EcoRI digestion. The lyophilized Cas9 protein (PNA Bio Inc) was resuspended in nuclease-free water at a concentration of 1.25 µg/µl and small aliquots were stored at -80°C. For microinjections, we mixed 400 ng/µl of Cas9 protein with 40-200 ng/µl sgRNA, incubated at 37°C for 5 min, transferred on ice, added the inert dye phenol red (5x from Sigma-Aldrich) and, for knock-in experiments, the tagging plasmid at a concentration of 10 ng/µl. The injection mix was centrifuged for 20 min at 4°C and the cleared solution was microinjected into 1-cell-stage embryos as previously described [45].

In the knock-out experiments, embryos were scored for phenotypes under a bright-field stereomicroscope 7-8 days after injection (stage S25-S27) when organogenesis is almost complete and the limbs are clearly visible through the transparent egg shell. To image the cuticle, anaesthetized hatchlings were fixed in 2%

paraformaldehyde in 1xPBS for 24 hours at room temperature. The samples were then washed in PTx (1xPBS containing 1% TritonX-100) and stained with 1 mg/ml Congo Red (Sigma-Aldrich) in PTx at room temperature with agitation for 24 hours. Stained samples were washed in PTx and mounted in 70% glycerol for imaging. Serial optical sections were obtained at 2 μm intervals with the 562 nm laser line on a Zeiss 710 confocal microscope using the Plan-Apochromat 10x/0.45 NA objective. Images were processed with Fiji (http://fiji.sc) and Photoshop (Adobe Systems Inc).

This methodology enabled us to also extract genomic DNA for genotyping from the same imaged specimen. Each specimen was disrupted with a disposable pestle in a 1.5 ml microtube (Kimble Kontes) in 50 μl of Squishing buffer (10 mM Tris-HCl pH 8, 1 mM EDTA, 25 mM NaCl, 200 μg/ml Proteinase K). The lysate was incubated at 37°C for a minimum of 2 hours, followed by heat inactivation of the Proteinase K for 5 min at 95°C, centrifugation at full speed for 5 min and transferring of the cleared lysate to a new tube. To recover the sequences in the PhDll-e locus targeted by the Dll1 and Dll2 sgRNAs, 5 μl of the lysate were used as template in a 50 μl PCR reaction with the Phusion DNA polymerase (New England Biolabs) and primers 313For (5'-TGGTTTTAGCAACAGTGAAGTGA-3') and 557Rev (5'-GACTGGGAGCGTGAGGGTA-3'). The amplified products were purified with the DNA Clean and Concentrator kit (Zymo Research), cloned with the Zero Blunt TOPO PCR Cloning Kit (Thermo Fisher Scientific) and sequenced with the M13 forward primer.

For the knock-in experiments, we constructed the tagging plasmid pCRISPR-NHEJ-KI-Dll-T2A-H2B-Ruby2 that contained the PhDll-e coding sequence fused in-frame with the T2A self-cleaving peptide, the *Parhyale histone* H2B and the Ruby 2 monomeric red fluorescent protein, followed by the PhDll-e 3'UTR and the pGEM-T Easy vector backbone (Promega). This tagging plasmid has a modular design with unique restriction sites for easy exchange of any desired part. More details are available upon request. Embryos co-injected with the Cas9 protein, the Dll2 sgRNA and the pCRISPR-NHEJ-KI-Dll-T2A-H2B-Ruby2 tagging plasmid were screened for nuclear fluorescence in the developing appendages under an Olympus MVX10 epi-fluorescence stereomicroscope. To image expression, live embryos at stage S22 were mounted in 0.5% SeaPlaque low-melting agarose (Lonza) in glass bottom microwell dishes (MatTek Corporation) and scanned as described above acquiring both the fluorescence and transmitted light on an inverted Zeiss 880 confocal microscope. To recover the chromosome-plasmid junctions, genomic DNA was extracted from transgenic siblings with fluorescent limbs and used as template in PCR reaction as described above with primer pair 313For and H2BRev (5'-TTACTTAGAAGAAGTGTACTTTG-3') for the left junction and primer pair M13 forward and 557Rev for the right junction. Amplified products were purified and cloned as described above and sequenced with the M13 forward and M13 reverse primers.

LIST OF FIGURES AND TABLES

Primary Figures

Primary Tables

Secondary Data

Figure 4, Supplemental Table 1 Classification of repeat elements

Figure 4, Supplemental Figure 2 CEGMA statistics of transcriptome and genome

Figure 5, Supplemental Figure 1 Expansion of gene families

Figure 6, Supplemental Figure 1 Variation in selected developmental genes

Figure 8, Supplemental Figure 1 *FGF* genes among arthropods and vertebrates

Figure 8, Supplemental Table 2 Classification of putative transcription factor DNA binding domains

Figure 8, Supplemental Figure 3 Ceramide Synthase (CERS) genes in metazoa

Figure 12, Supplemental Figure 1 Alignment of GH genes

Figure 13 Supplemental Figure 1 Alignment of DSCAM alternative exons

Figure 14, Supplemental Figure 1 Phylogenetic relationship of *DICER* and *PIWI* genes

Figure 14, Supplemental Figure 2 Clustering o miRNA in the *Parhyale* genome

Figure 16, Supplemental Figure 1 Molecular constructs used for genome editing

Supplemental Data

Data 1 List of external data sources and software used

Data 2 Annotation of *Parhyale* genes used

Data 3 Variation found in *Parhyale* developmental genes

Data 4 KEGG annotation of *Parhyale* genes

Data 5 RFAM classification of putative RNA elements

Data 6 IPython notebook detailing the assembly process

Data 7 IPython notebook detailing the repeat masking process and results

Data 8 IPython notebook detailing the annotation process

Data 9 IPython notebook detailing variant analysis

Data 10 IPython notebook detailing the orthology analysis

Data 11 IPython notebook detailing the RNA classification analysis

Data 12 txt file listing gene ids of amphipod specific genes

Data 13 txt file indexing KEGG ids to gene ids

Data 14 txt file listing gene ids of malacostraca specific genes

Data 15 txt file listing orthologous groups as outputted by OrthoFinder

Data 16 txt file indexing PFAM ids to gene IDS

Data 17 zip file of txt files listing gene ids of various clade specific gene ids

FIGURES AND TABLES

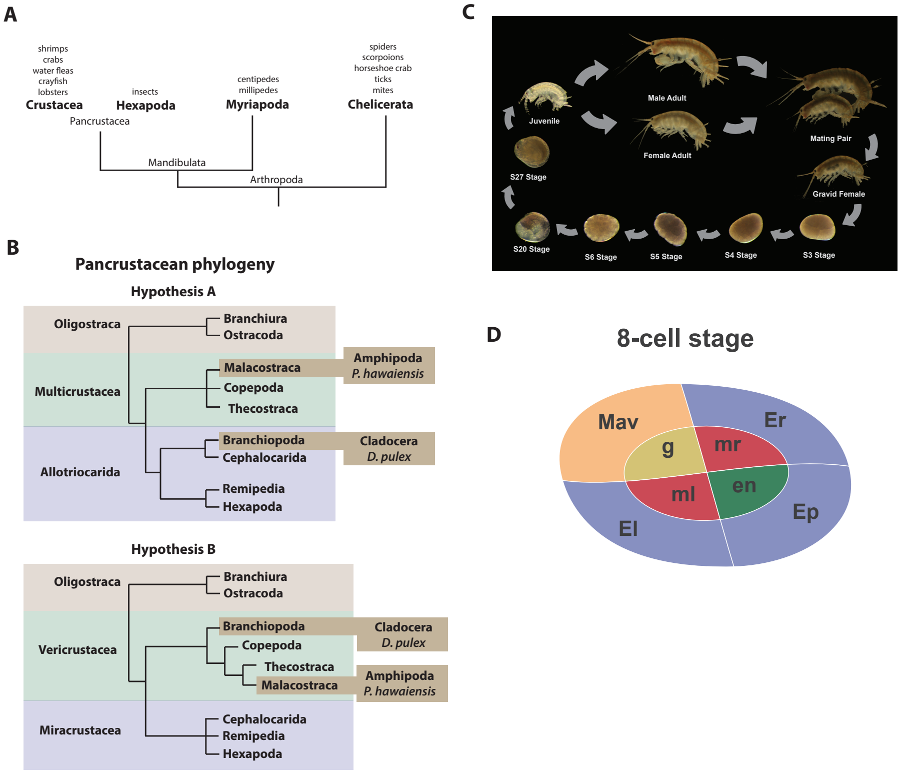


Figure 1. **Introduction.** **(A)** Phylogenetic relationship of Arthropods showing the Chelicerata as an outgroup to Mandibulata and the Pancrustacea clade which includes crustaceans and insects. Species listed for each clade have ongoing or complete genomes. Species include Crustacea: *Parhyale hawaiensis*, *D. pulex*; Hexapoda: *Drosophila melanogaster*, *Apis mellifera*, *Bombyx mori*, *Aedis aegypti*, *Tribolium castaneum*; Myriapoda: *Strigamia maritima*, *Trigoniulus corallines*; Chelicerata: *Ixodes scapularis*, *Tetranychus urticae*, *Mesobuthus martensii*, *Stegodyphus mimosarum*. **(B)** One of the unresolved issues concerns the placement of the Branchiopoda either together with the Cephalocarida, Remipedia and Hexapoda (Allotriocarida hypothesis A) or with the Copepoda, Thecostraca and Malacostraca (Vericrustacea hypothesis B). **(C)** Life cycle of *Parhyale* that takes about two months at 26°C. *Parhyale* is a direct developer and a sexually dimorphic species. The fertilized egg undergoes stereotyped total cleavages and each blastomere becomes committed to a particular germ layer already at the 8-cell stage depicted in **(D)**. The three macromeres Er, El, and Ep give rise to the anterior right, anterior left, and posterior ectoderm, respectively, while the fourth macromere Mav gives rise to the visceral mesoderm and anterior head somatic mesoderm. Among the 4 micromeres, the mr and ml micromeres give rise to the right and left somatic trunk mesoderm, en gives rise to the endoderm, and g gives rise to the germline.

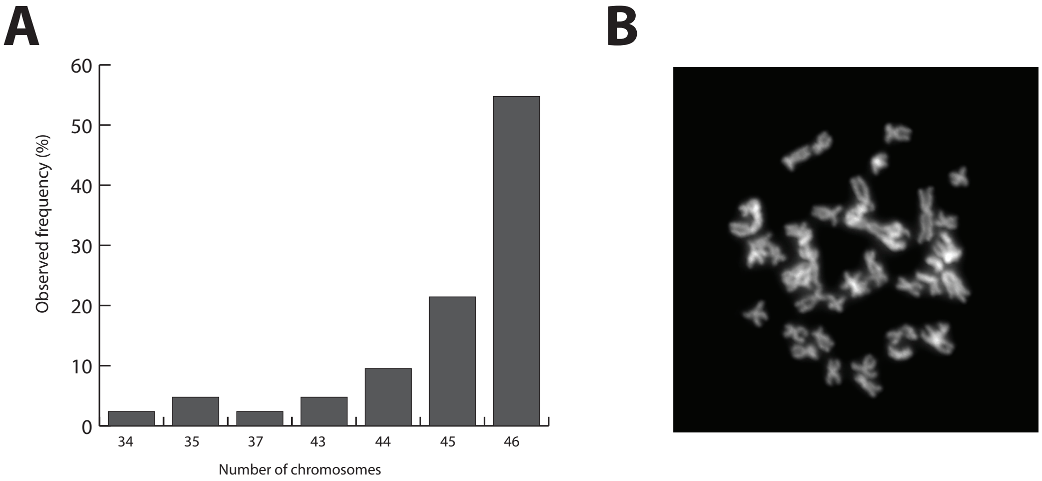


Figure 2. ***Parhyale* karyotype.** **(A)** Frequency of the number of chromosomes observed in 42 mitotic spreads. Forty-six chromosomes were observed in more than half of all preparations. **(B)** Representative image of Hoechst-stained chromosomes.

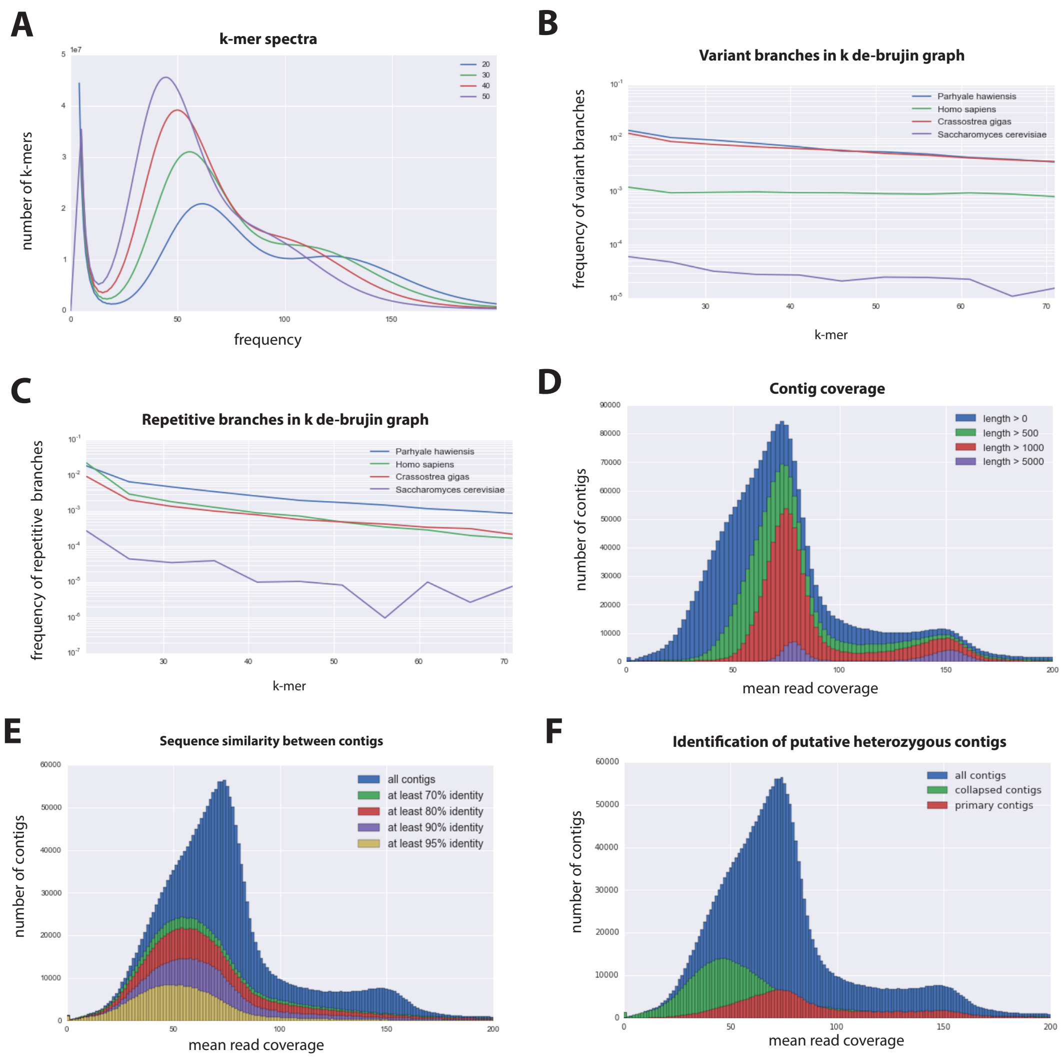


Figure 3. ***Parhyale*** **genome assembly metrics.** **(A)** K-mer frequency spectra of all reads for k-lengths ranging from 20 to 50. **(B)** K-mer branching analysis showing the frequency of k-mer branches classified as variants compared to *Homo sapiens* (human), *Crassostrea gigas* (oyster), and *Saccharomyces cerevisiae* (yeast). **(C)** K-mer branching analysis showing the frequency of k-mer branches classified as repetitive compared to *H. sapiens, C. gigas and S. cerevisiae*. **(D)** Histogram of read coverages of assembled contigs. **(E)** The number of contigs with an identity ranging from 70-95% to another contig in the set of assembled contigs. **(F)** Collapsed contigs (green) are contigs with at least 95% identity with a longer primary contig (red). These contigs were removed prior to scaffolding and added back as potential heterozygous contigs after scaffolding.

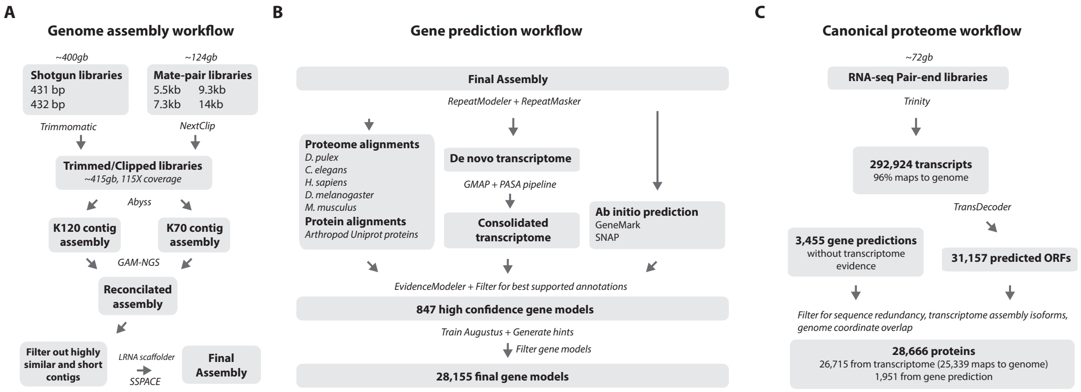


Figure 4. Workflows of assembly, annotation, and proteome generation. **(A)** Flowchart of the genome assembly. Two shotgun libraries and four mate-pair libraries with the indicated average sizes were prepared from a single male animal and sequenced to a predicted depth of 115x coverage after read filtering, based on a predicted size of 3.6 Gbp. Contigs were assembled at two different k-lengths with Abyss and the two assemblies were merged with GAM-NGS. Filtered contigs were scaffolded with SSPACE. **(B)** The final scaffolded assembly was annotated with a combination of Evidence Modeler to generate 847 high quality gene models and Augustus for the final set of 28,155 predictions. These protein-coding gene models were generated based on a *Parhyale* transcriptome consolidated from multiple developmental stages and conditions, their homology to the species indicated, and *ab initio* predictions with GeneMark and SNAP. **(C)** The *Parhyale* proteome contains 28,666 entries based on the consolidated transcriptome and gene predictions. The transcriptome contains 292,924 coding and non-coding RNAs, 96% of which could be mapped to the assembled genome.

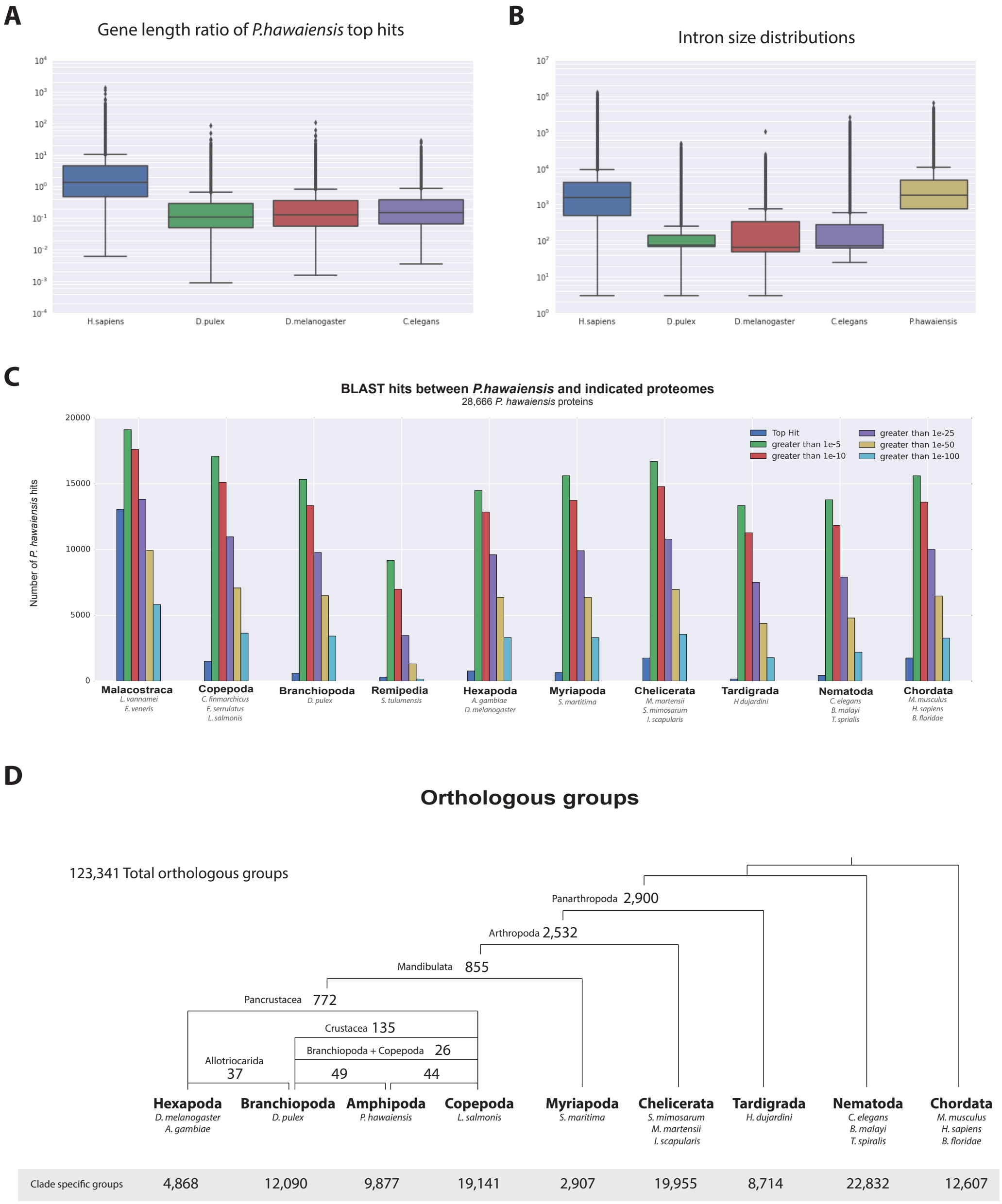


Figure 5. ***Parhyale* genome comparisons. (A)** Box plots comparing gene sizes between *Parhyale* and humans (*H. sapiens*), water fleas (*D. pulex*), flies (*D. melanogaster*) and nematodes (*C. elegans*). Ratios were calculated by dividing the size of the top blast hit in each species with the corresponding *Parhyale* gene size. **(B)** Box plots showing the distribution of intron sizes in the same species used in A. **(C)** Comparison between *Parhyale* and representative proteomes from the indicated animal taxa. Colored bars indicate the number of blast hits recovered across various thresholds of E-values. The top hit value represents the number of proteins with a top hit corresponding to the respective species. **(D)** Cladogram showing the number of shared orthologous protein groups at various taxonomic levels, as well as the number of clade-specific groups. A total of 123,341 orthogroups were identified with Orthofinder across the 16 genomes used in this analysis. Within Pancrustacea, 37 orthogroups were shared between Branchiopoda and Hexapoda (supporting the Allotriocarida hypothesis) and 49 orthogroups were shared between Branchiopoda and Amphipoda (supporting the Vericrustacea hypothesis).

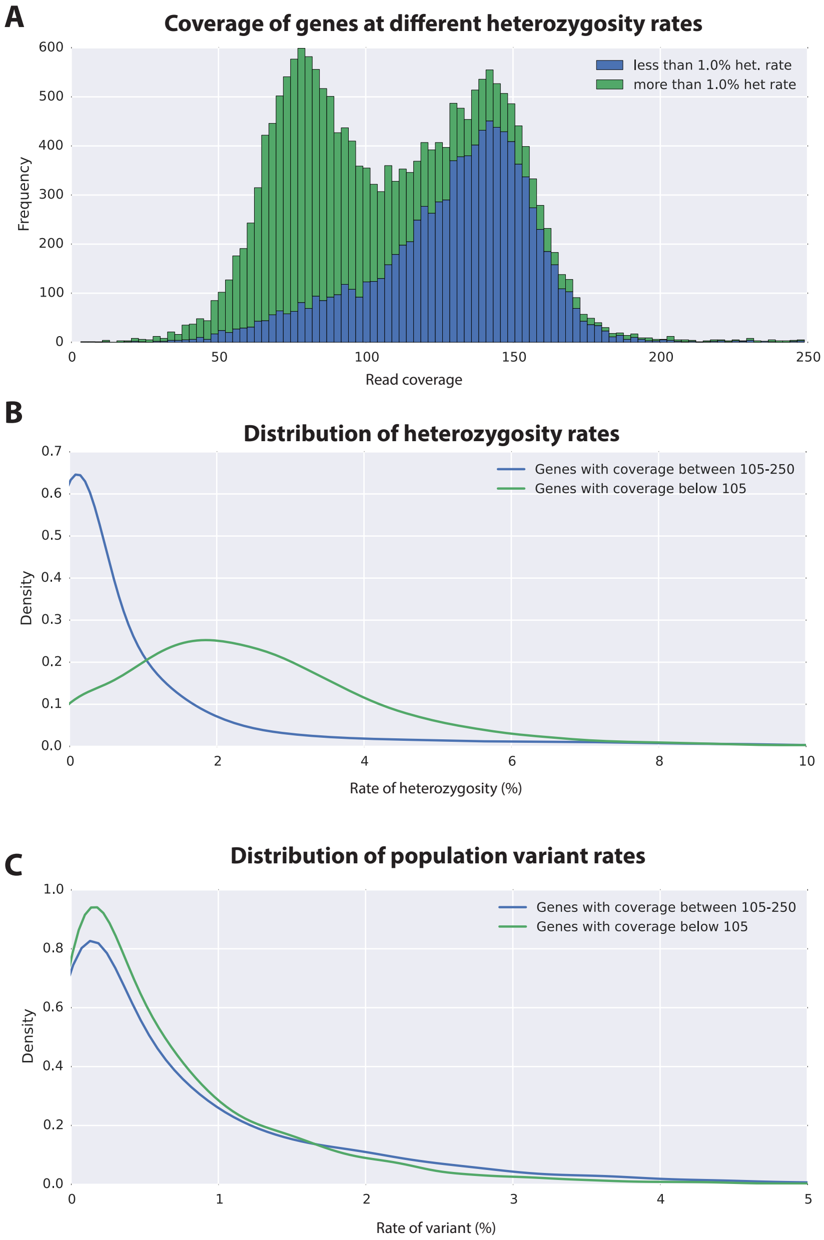
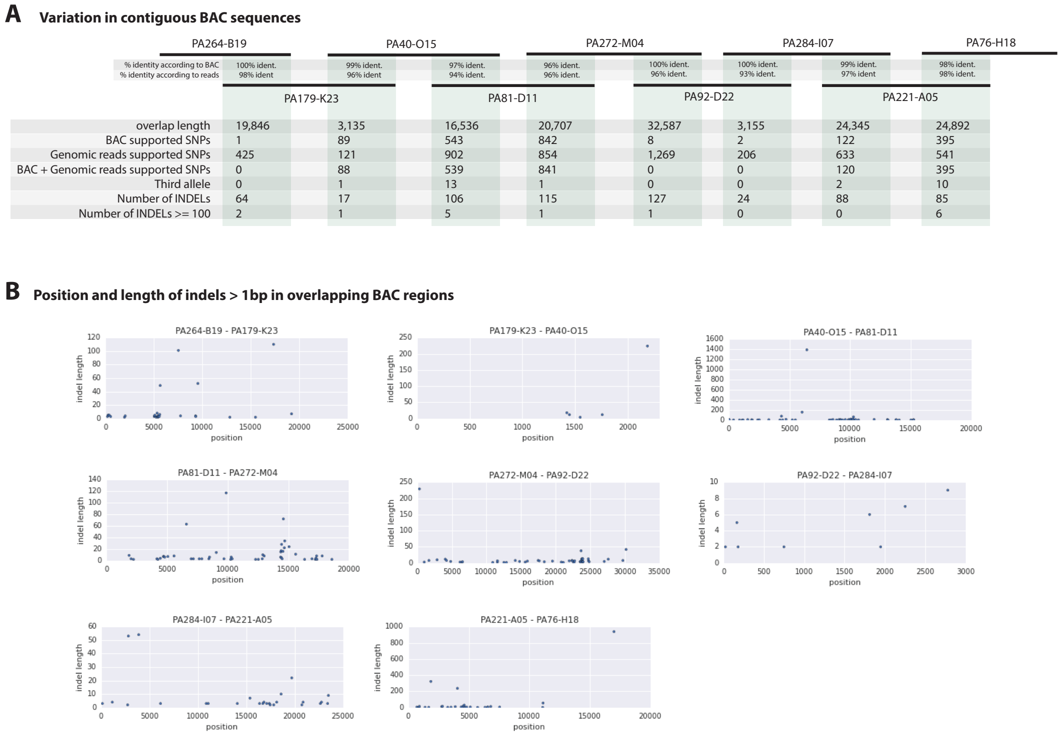


Figure 6. **Variation analyses of predicted genes.** **(A)** A read coverage histogram of predicted genes. Reads were first mapped to the genome, then coverage was calculated for transcribed regions of each defined locus. **(B)** A coverage distribution plot showing that genes in the lower coverage region (<105x coverage, peak at 75x) have a higher level of heterozygosity than genes in the higher coverage region (>105 coverage and <250, peak at approximately 150x coverage). **(C)** Distribution plot indicating that mean level of population variance is similar for genes in the higher and lower coverage regions.



% identity according to BAC	100% ident.	99% ident.	97% ident.	96% ident.	100% ident.	100% ident.	99% ident.	98% ident.
% identity according to reads	98% ident	96% ident	94% ident.	96% ident.	96% ident.	93% ident.	97% ident	98% ident.
overlap length	19,846	3,135	16,536	20,707	32,587	3,155	24,345	24,892
BAC supported SNPs	1	89	543	842	8	2	122	395
Genomic reads supported SNPs	425	121	902	854	1,269	206	633	541
BAC + Genomic reads supported SNPs	0	88	539	841	0	0	120	395
Third allele	0	1	13	1	0	0	2	10
Number of INDELs	64	17	106	115	127	24	88	85
Number of INDELs >= 100	2	1	5	1	1	0	0	6

Figure 7. Variation observed in contiguous BAC sequences. (A) Schematic diagram of the contiguous BAC clones tiling across the HOX cluster and their % sequence identities. "Overlap length" refers to the lengths (bp) of the overlapping regions between two BAC clones. "BAC supported single nucleotide polymorphisms (SNPs)" refer to the number of SNPs found in the overlapping regions by pairwise alignment. "Genomic reads supported SNPs" refer to the number of SNPs identified in the overlapping regions by mapping all reads to the BAC clones and performing variant calling with GATK. "BAC + Genomic reads supported SNPs" refer to the number of SNPs identified from the overlapping regions by pairwise alignment that are supported by reads. "Third allele" refers to presence of an additional polymorphism not detected by genomic reads. "Number of INDELs" refer to the number of all insertion or deletions found in the contiguous region. "Number of INDELs >100" are insertion or deletions greater than or equal to 100. **(B)** Position versus indel lengths across each overlapping BAC region.

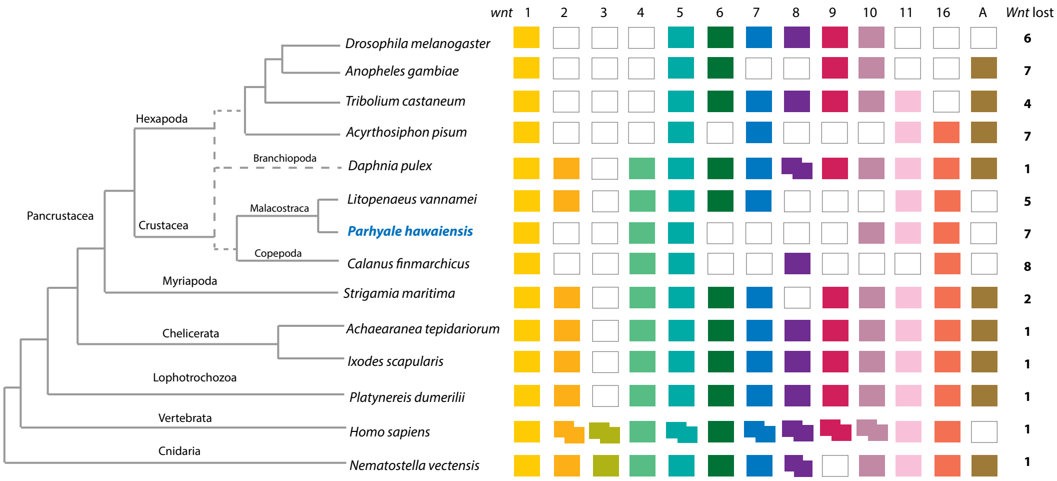


Figure 8. Comparison of Wnt family members across Metazoa. Comparison of Wnt genes across Metazoa. Tree on the left illustrates the phylogenetic relationships of species used. Dotted lines in the phylogenetic tree illustrate the alternative hypothesis of Branchiopoda + Hexapoda versus Branchiopoda + Multicrustacea. Colour boxes indicate the presence of certain Wnt subfamily members (wnt1 to wnt11, wnt16 and wntA) in each species. Empty boxes indicate the loss of particular Wnt genes. Two overlapping colour boxes represent duplicated Wnt genes.

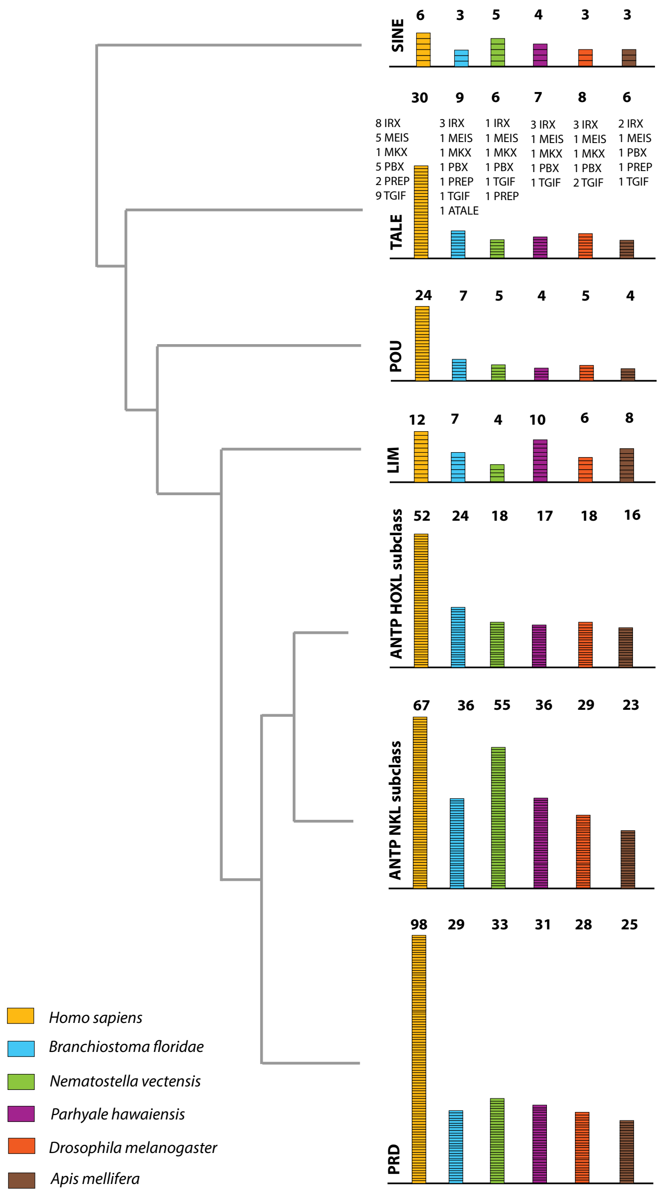


Figure 9. **Homeodomain protein family tree.** The overview of homeodomain radiation and phylogenetic relationships among homeodomain proteins from Arthropoda (*P. hawaiensis, D. melanogaster and A. mellifera*), Chordata (*H. sapiens and B. floridae*), and Cnidaria (*N. vectensis*). Six major homeodomain classes are illustrated (SINE, TALE, POU, LIM, ANTP and PRD) with histograms indicating the number of genes in each species belonging to a given class.

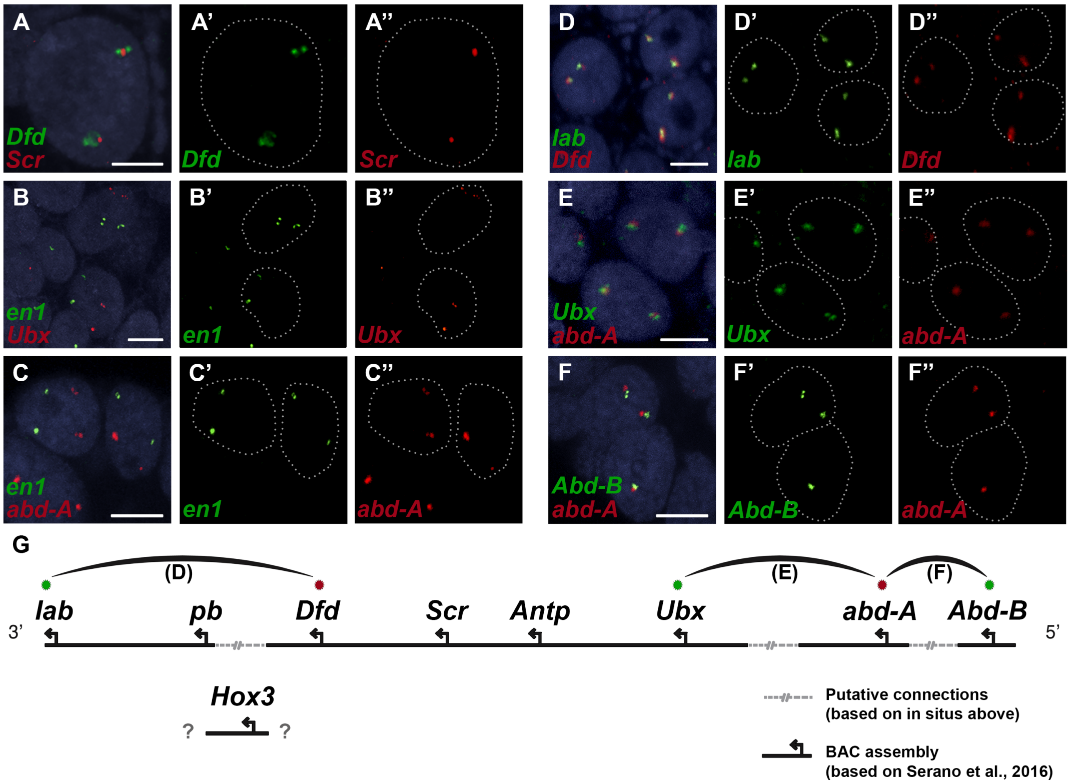


Figure 10. **Evidence for an intact Hox cluster in *Parhyale*. (A-F'')** Double fluorescent *in situ* hybridizations (FISH) for nascent transcripts of genes. **(A-A'')** *Deformed* (*Dfd*) and *Sex combs reduced* (*Scr*), **(B-B'')** *engrailed 1* (*en1*) and *Ultrabithorax* (*Ubx*), **(C-C'')** *en1* and *abdominal-A* (*abd-A*), **(D-D'')** *labial* (*lab*) and *Dfd*, **(E-E'')** *Ubx* and *abd-A*, and **(F-F'')** *Abdominal-B* (*Abd-B*) and *abd-A*. Cell nuclei are stained with DAPI (blue) in panels A-F and outlined with white dotted lines in panels A'-F' and A''-F''. Co-localization of nascent transcript dots in A, D, E and F suggest the proximity of the corresponding Hox genes in the genomic DNA. As negative controls, the *en1* nascent transcripts in B and C do not co-localize with those of Hox genes *Ubx* or *abd-A*. **(G)** Schematic representation of the predicted configuration of the Hox cluster in *Parhyale*. Previously identified genomic linkages are indicated with solid black lines, whereas linkages established by FISH are shown with dotted gray lines. The arcs connecting the green and red dots represent the linkages identified in D, E and F, respectively. The position of the Hox3 gene is still uncertain. Scale bars are 5μm.

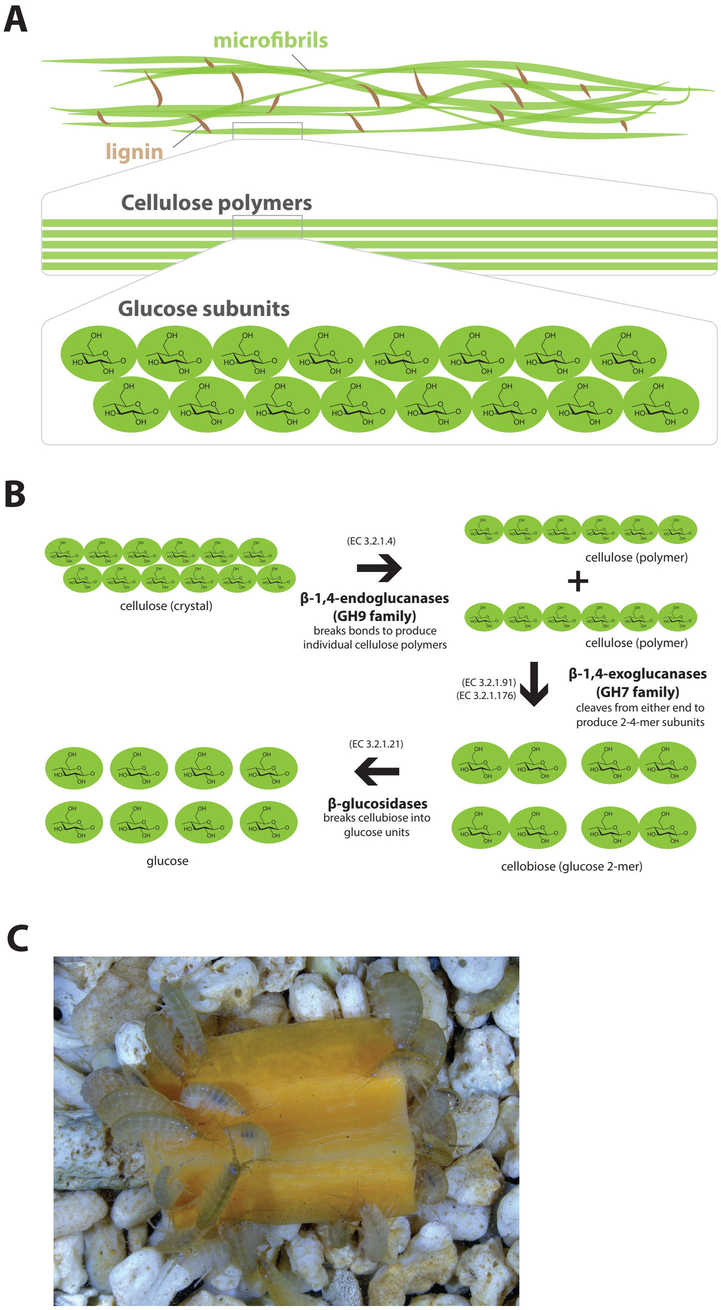


Figure 11. Lignocellulose digestion overview. (A) Simplified drawing of lignocellulose structure. The main component of lignocellulose is cellulose, which is a β-1,4-linked chain of glucose monosaccharides. Cellulose and lignin are organized in structures called microfibrils, which in turn form macrofibrils. **(B)** Summary of cellulolytic enzymes and reactions involved in the breakdown of cellulose into glucose. β-1,4-endoclucanases of the GH9 family catalyze the hydrolysis of crystalline cellulose into cellulose chains. β-1,4-exoclucanases of the GH7 family break down cellulose chains into cellobiose (glucose disaccharide) that can be converted to glucose by β-glucosidases. **(C)** Adult *Parhyale* feeding on a slice of carrot.

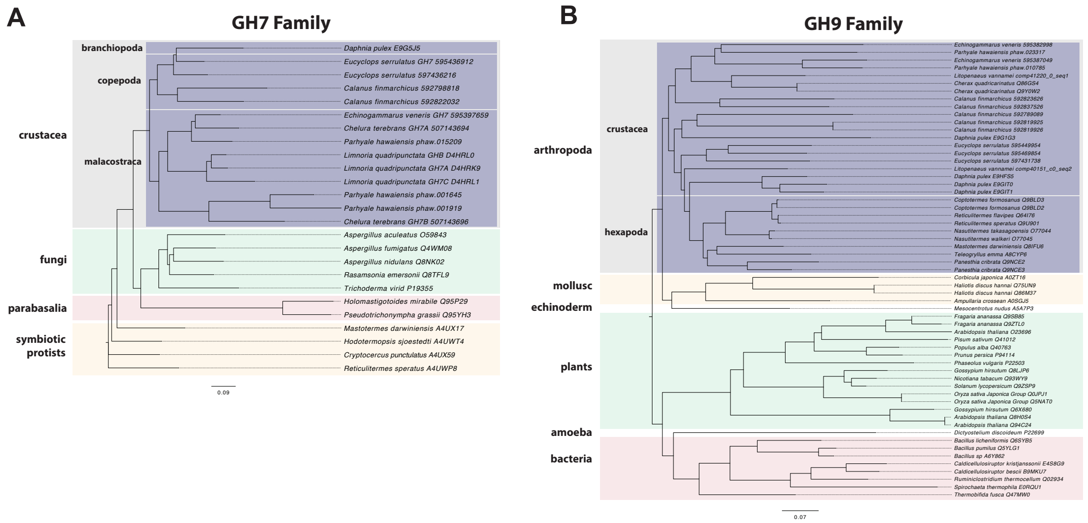

A
GH7 Family
branchiopoda
copepoda
crustacea
malacostraca
fungi
parabasalia
symbiotic protists
Daphnia pulex E9G5J5
Eucyclops serrulatus GH7 595436912
Eucyclops serrulatus 597436216
Calanus finmarchicus 592798818
Calanus finmarchicus 592822032
Echinogammarus veneris GH7 595397659
Chelura terebrans GH7A 507143694
Parhyale hawaiensis phaw.015209
Limnoria quadripunctata GHB D4HRL0
Limnoria quadripunctata GH7A D4HRK9
Limnoria quadripunctata GH7C D4HRL1
Parhyale hawaiensis phaw.001645
Parhyale hawaiensis phaw.001919
Chelura terebrans GH7B 507143696
Aspergillus aculeatus O59843
Aspergillus fumigatus Q4WM08
Aspergillus nidulans Q8NK02
Rasamsonia emersonii Q8TFL9
Trichoderma virid P19355
Holomastigotoides mirabile Q95P29
Pseudotrichonympha grassii Q95YH3
Mastotermes darwiniensis A4UX17
Hodotermopsis sjoestedti A4UWT4
Cryptocercus punctulatus A4UX59
Reticulitermes speratus A4UWP8
0.09
B
GH9 Family
arthropoda
crustacea
hexapoda
mollusc
echinoderm
plants
amoeba
bacteria
Echinogammarus veneris 595382998
Parhyale hawaiensis phaw.023317
Echinogammarus veneris 595387049
Parhyale hawaiensis phaw.010785
Litopenaeus vannamei comp41220_0_seq1
Cherax quadricarinatus Q86GS4
Cherax quadricarinatus Q9Y0W2
Calanus finmarchicus 592823626
Calanus finmarchicus 592837526
Calanus finmarchicus 592789089
Calanus finmarchicus 592819925
Calanus finmarchicus 592819926
Daphnia pulex E9G1G3
Eucyclops serrulatus 595449954
Eucyclops serrulatus 595469854
Eucyclops serrulatus 597431738
Litopenaeus vannamei comp40151_c0_seq2
Daphnia pulex E9HFS5
Daphnia pulex E9GIT0
Daphnia pulex E9GIT1
Coptotermes formosanus Q9BLD3
Coptotermes formosanus Q9BLD2
Reticulitermes flavipes Q64I76
Reticulitermes speratus Q9U901
Nasutitermes takasagoensis O77044
Nasutitermes walkeri O77045
Mastotermes darwiniensis Q8IFU6
Teleogryllus emma A8CYP6
Panesthia cribrata Q9NCE2
Panesthia cribrata Q9NCE3
Corbicula japonica A0ZT16
Haliotis discus hannai Q75UN9
Haliotis discus hannai Q86M37
Ampullaria crossean A0SGJ5
Mesocentrotus nudus A5A7P3
Fragaria ananassa Q9SB85
Fragaria ananassa Q9ZTL0
Arabidopsis thaliana O23696
Pisum sativum Q41012
Populus alba Q40763
Prunus persica P94114
Phaseolus vulgaris P22503
Gossypium hirsutum Q8LJP6
Nicotiana tabacum Q93WY9
Solanum lycopersicum Q9ZSP9
Oryza sativa Japonica Group Q0JPJ1
Oryza sativa Japonica Group Q5NAT0
Gossypium hirsutum Q6X680
Arabidopsis thaliana Q8H0S4
Arabidopsis thaliana Q94C24
Dictyostelium discoideum P22699
Bacillus licheniformis Q6SYB5
Bacillus pumilus Q5YLG1
Bacillus sp A6Y862
Caldicellulosiruptor kristjanssonii E4S8G9
Caldicellulosiruptor bescii B9MKU7
Ruminiclostridium thermocellum Q02934
Spirochaeta thermophila E0RQU1
Thermobifida fusca Q47MW0
0.07

Figure 12. Phylogenetic analysis of GH7 and GH9 family proteins. **(A)** Phylogenetic tree showing the relationship between GH7 family proteins of *Parhyale*, other crustaceans (Malacostraca, Branchiopoda, Copepoda), fungi and symbiotic protists (root). UniProt and GenBank accessions are listed next to the species names. **(B)** Phylogenetic tree showing the relationship between GH9 family proteins of *Parhyale*, crustaceans, insects, molluscs, echinoderms, amoeba, bacteria and plants (root). UniProt and GenBank accessions are listed next to the species names. Both trees were constructed with RAxML using the WAG+G model from multiple alignments of protein sequences created with MUSCLE.

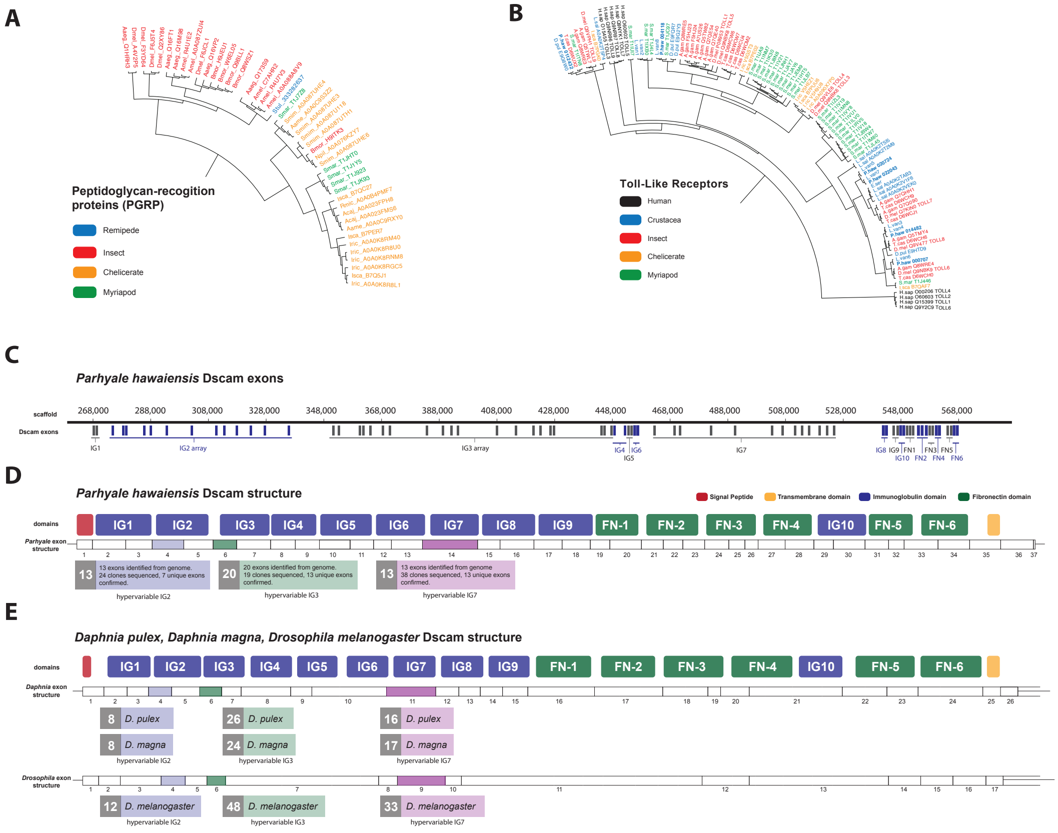


Figure 13. Comparison of innate immunity genes. (A) Phylogenetic tree of peptidoglycan recognition proteins (PGRPs). With the exception of Remipedes, PGRPs were not found in Crustaceans. PGRPs have been found in Arthropods, including insects, Myriapods and Chelicerates. **(B)** Phylogenetic tree of Toll-like receptors (TLRs) generated from five Crustaceans, three Hexapods, two Chelicerates, one Myriapod and one vertebrate species. **(C)** Genomic organization of the *Parhyale* Dscam locus showing the individual exons and exon arrays encoding the immunoglobulin (IG) and fibronectin (FN) domains of the protein. **(D)** Structure of the *Parhyale* Dscam locus and comparison with the **(E)** Dscam loci from *Daphnia pulex, Daphnia magna* and *Drosophila melanogaster*. The white boxes represent the number of predicted exons in each species encoding the signal peptide (red), the IGs (blue), the FNs and transmembrane (yellow) domains of the protein. The number of alternatively spliced exons in the arrays encoding the hypervariable regions IG2 (exon 4 in all species), IG3 (exon 6 in all species) and IG7 (exon 14 in *Parhyale*, 11 in *D. pulex* and 9 in *Drosophila*) are indicated under each species schematic in the purple, green and magenta boxes, respectively. Abbreviations of species used: *Parhyale hawaiensis* (Phaw), *Bombyx mori* (Bmor), *Aedes aegypti* (Aaeg), *Drosophila melanogaster* (Dmel), *Apis mellifera* (Amel), *Speleonectes tulumensis* (Stul), *Strigamia maritima* (Smar), *Stegodyphus mimosarum* (Smim), *Ixodes scapularis* (Isca), *Amblyomma americanum* (Aame), *Nephila pilipes* (Npil), *Rhipicephalus microplus* (Rmic), *Ixodes ricinus* (Iric), *Amblyomma cajennense* (Acaj), *Anopheles gambiae* (Agam), *Daphnia pulex* (Apul), *Tribolium castaneum* (Tcas), *Litopenaeus vannamei* (Lvan), *Lepeophtheirus salmonis* (Lsal), *Eucyclops serrulatus* (Eser), *Homo sapiens* (H.sap). Both trees were constructed with RAxML using the WAG+G model from multiple alignments of protein sequences created with MUSCLE.

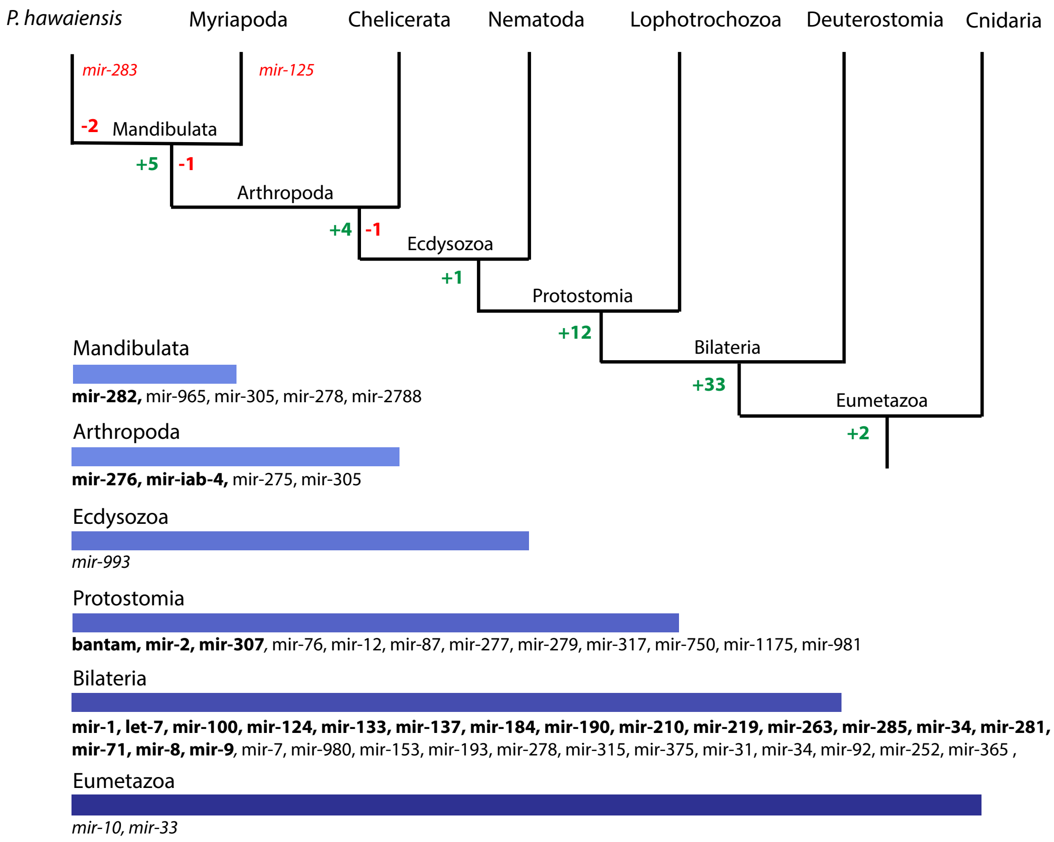

P. hawaiensis
Myriapoda
Chelicerata
Nematoda
Lophotrochozoa
Deuterostomia
Cnidaria
mir-283
mir-125
-2
Mandibulata
+5
-1
Arthropoda
+4
-1
Ecdysozoa
+1
Protostomia
+12
Bilateria
+33
Eumetazoa
+2
Mandibulata
mir-282, mir-965, mir-305, mir-278, mir-2788
Arthropoda
mir-276, mir-iab-4, mir-275, mir-305
Ecdysozoa
mir-993
Protostomia
bantam, mir-2, mir-307, mir-76, mir-12, mir-87, mir-277, mir-279, mir-317, mir-750, mir-1175, mir-981
Bilateria
mir-1, let-7, mir-100, mir-124, mir-133, mir-137, mir-184, mir-190, mir-210, mir-219, mir-263, mir-285, mir-34, mir-281, mir-71, mir-8, mir-9, mir-7, mir-980, mir-153, mir-193, mir-278, mir-315, mir-375, mir-31, mir-34, mir-92, mir-252, mir-365 ,
Eumetazoa
mir-10, mir-33

Figure 14. **Evolution of miRNA families in Eumetazoans.** Phylogenetic tree showing the gains (in green) and losses (in red) of miRNA families at various taxonomic levels of the Eumetazoan tree leading to *Parhyale*. miRNAs marked with plain characters were identified by MirPara with small RNA sequencing read support. miRNAs marked with bold characters were identified by Rfam and MirPara with small RNA sequencing read support.

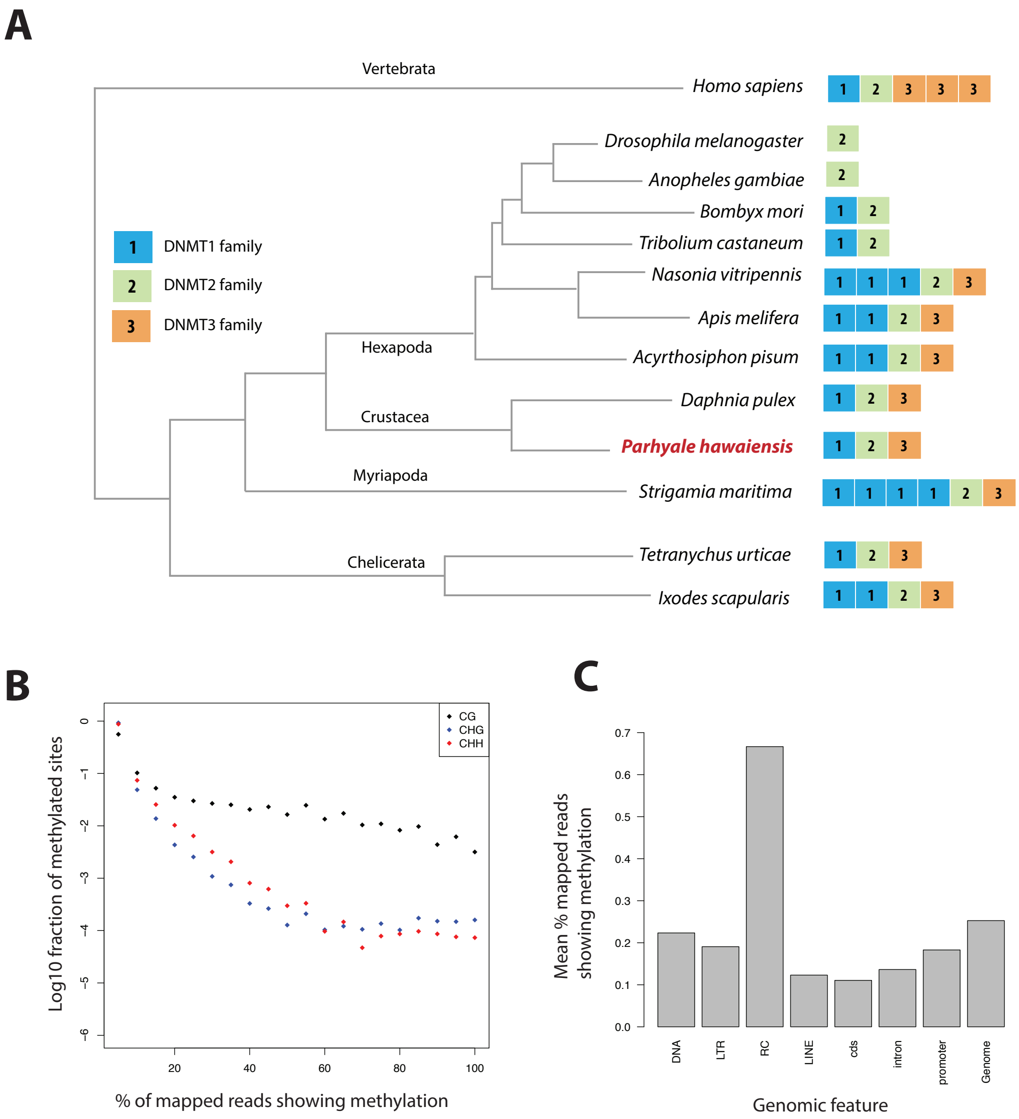


Figure 15. Analysis of *Parhyale* genome methylation. **(A)** Phylogenetic tree showing the families and numbers of DNA methyltransferases (DNMTs) present in the genomes of indicated species. *Parhyale* has one copy from each DNMT family. **(B)** Amounts of methylation detected in the *Parhyale* genome. Amount of methylation is presented as percentage of reads showing methylation in bisulfite sequencing data. DNA methylation was analyzed in all sequence contexts (CG shown in dark, CHG in blue and CHH in red) and was detected preferentially in CpG sites. **(C)** Histograms showing mean percentages of methylation in different fractions of the genome: DNA transposons (DNA), long terminal repeat transposable elements (LTR), rolling circle transposable elements (RC), long interspersed elements (LINE), coding sequences (cds), introns, promoters, and the rest of the genome.

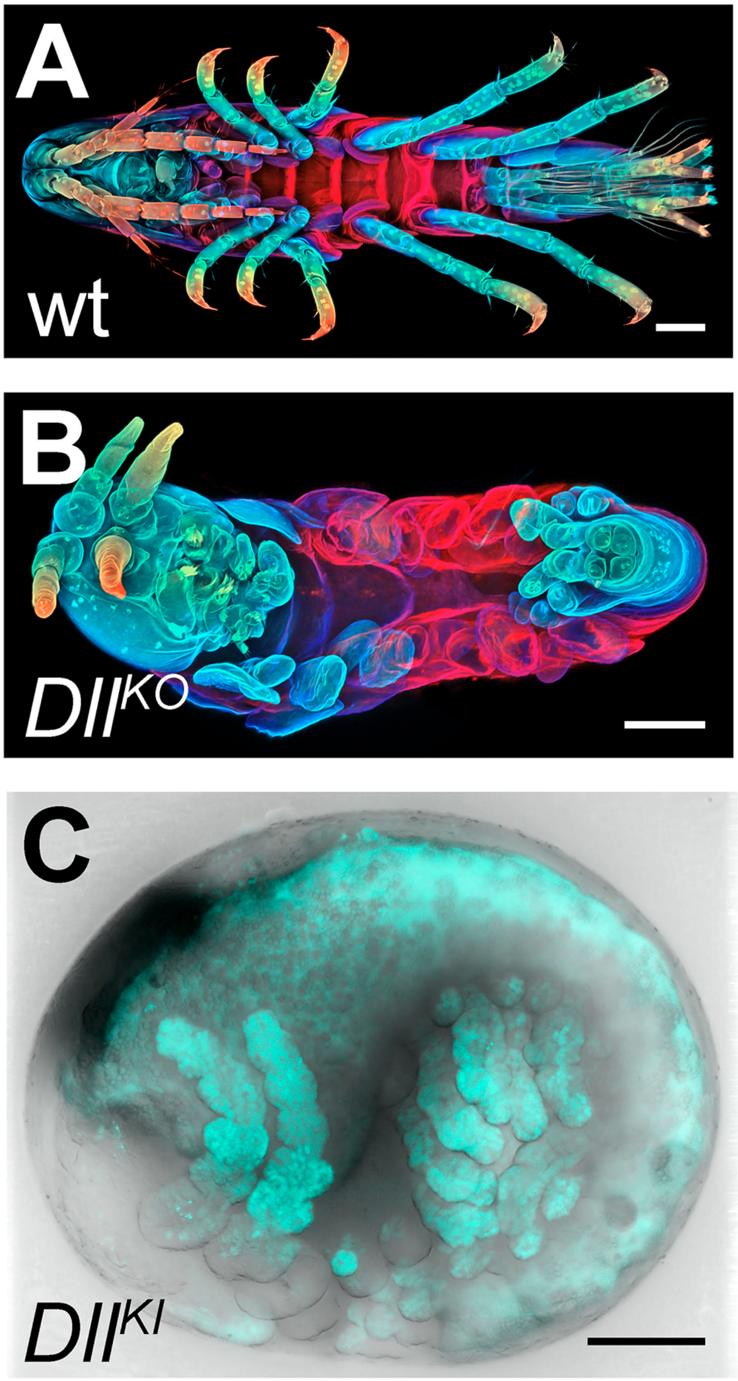


Figure 16. CRISPR/Cas9-based genome editing in *Parhyale*. (A) Wild-type morphology. **(B)** Mutant *Parhyale* with truncated limbs after CRISPR-mediated knock-out (DllKO) of the limb patterning gene *Distal-less* (*PhDll-e*). Panels show ventral views of juveniles stained for cuticle and color-coded by depth with anterior to the left. **(C)** Fluorescent tagging of *PhDll-e* expressed in most limbs (shown in cyan) by CRISPR-mediated knock-in (DllKI) using the non-homologous-end-joining repair mechanism. Panel shows a lateral view with anterior to the left and dorsal to the top of a live embryo (stage S22) with merged bright-field and fluorescence channels. Yolk autofluorescence produces a dorsal crescent of fluorescence in the gut. Scale bars are 100 µm.

Table 1. Experimental resources. Available experimental resources in *Parhyale* and corresponding references.

Experimental Resources	References
Embryological manipulations Cell microinjection, isolation, ablation	[36–38, 41–46]
Gene expression studies In situ hybridization, antibody staining	[39, 40]
Gene knock-down RNA interference, morpholinos	[24, 50]
Transgenesis Transposon-based, integrase-based	[45, 48, 49]
Gene trapping Exon/enhancer trapping, iTRAC (trap conversion)	[49]
Gene misexpression Heat-inducible	[25]
Gene knock-out CRISPR/Cas	[19]
Gene knock-in CRISPR/Cas homology-dependent or homology-independent	[18]
Live imaging Bright-field, confocal, light-sheet microscopy	[43, 44, 47]

Table 2. Assembly statistics. Length metrics of assembled scaffolds and contigs.

	# sequences	N90	N50	N10	Sum Length	Max Length	# Ns
scaffolds	133,035	14,799	81,190	289,705	3.63GB	1,285,385	1.10GB
unplaced contigs	259,343	304	627	1,779	146MB	40,222	23,431
hetero. contigs	584,392	265	402	1,038	240MB	24,461	627
genic scaffolds	15,160	52,952	161,819	433,836	1.49GB	1,285,385	323MB

Table 3. **BAC variant statistics.** Level of heterozygosity of each BAC sequence determined by mapping genomic reads to each BAC individually. Population variance rate represent additional alleles found (more than 2 alleles) from genomic reads.

BAC ID	Length	Heterozygosity	Pop.Variance
PA81-D11	140,264	1.654	0.568
PA40-O15	129,957	2.446	0.647
PA76-H18	141,844	1.824	0.199
PA120-H17	126,766	2.673	1.120
PA222-D11	128,542	1.344	1.404
PA31-H15	140,143	2.793	0.051
PA284-I07	141,390	2.046	0.450
PA221-A05	148,703	1.862	1.427
PA93-L04	139,955	2.177	0.742
PA272-M04	134,744	1.925	0.982
PA179-K23	137,239	2.671	0.990
PA92-D22	126,848	2.650	0.802
PA268-E13	135,334	1.678	1.322
PA264-B19	108,571	1.575	0.157
PA24-C06	141,446	1.946	1.488

Table 4. Small RNA processing pathway members. The *Parhyale* orthologs of small RNA processing pathway members.

Gene	Counts	Gene ID
Armitage	2	phaw_30_tra_m.006391 phaw_30_tra_m.007425
Spindle_E	3	phaw_30_tra_m.000091 phaw_30_tra_m.020806 phaw_30_tra_m.018110
rm62	7	phaw_30_tra_m.014329 phaw_30_tra_m.012297 phaw_30_tra_m.004444 phaw_30_tra_m.012605 phaw_30_tra_m.001849 phaw_30_tra_m.006468 phaw_30_tra_m.023485
Piwi/aubergine	2	phaw_30_tra_m.011247 phaw_30_tra_m.016012
Dicer 1	1	phaw_30_tra_m.001257
Dicer 2	1	phaw_30_tra_m.021619
argonaute 1	1	phaw_30_tra_m.006642
arogonaute 2	3	phaw_30_tra_m.021514 phaw_30_tra_m.018276 phaw_30_tra_m.012367
Loquacious	2	phaw_30_tra_m.006389 phaw_30_tra_m.000074
Drosha	1	phaw_30_tra_m.015433

REFERENCES

[1] M Akam. Arthropods: Developmental diversity within a (super) phylum. *Proceedings of the National Academy of Sciences of the United States of America*, 97(9):1–4, April 2000.

[2] Graham E Budd and Maximilian J Telford. The origin and evolution of arthropods. *Nature*, 457(7231):812–817, February 2009.

[3] Andrew D Peel, Ariel D Chipman, and Michael Akam. Arthropod Segmentation: beyond the Drosophila paradigm. *Nature reviews. Genetics*, 6(12):905–916, November 2005.

[4] G Scholtz and C Wolff. Arthropod embryology: cleavage and germ band development. *Arthropod Biology and Evolution*, 2013.

[5] Jon M Mallatt, James R Garey, and Jeffrey W Shultz. Ecdysozoan phylogeny and Bayesian inference: first use of nearly complete 28S and 18S rRNA gene sequences to classify the arthropods and their kin. *Molecular Phylogenetics and Evolution*, 31(1):178–191, April 2004.

[6] C E Cook, Q Yue, and M Akam. Mitochondrial genomes suggest that hexapods and crustaceans are mutually paraphyletic. *Proceedings. Biological sciences / The Royal Society*, 272(1569):1295–1304, June 2005.

[7] Jerome C Regier, Jeffrey W Shultz, and Robert E Kambic. Pancrustacean phylogeny: hexapods are terrestrial crustaceans and maxillopods are not monophyletic. *Proceedings. Biological sciences / The Royal Society*, 272(1561):395–401, February 2005.

[8] B Ertas, B M von Reumont, J W Wagele, B Misof, and T Burmester. Hemocyanin Suggests a Close Relationship of Remipedia and Hexapoda. *Molecular biology and evolution*, 26(12):2711–2718, November 2009.

[9] S Richter. The Tetraconata concept: hexapod-crustacean relationships and the phylogeny of Crustacea. *Organisms Diversity & Evolution*, 2(3):217–237, 2002.

[10] John K Colbourne, Michael E Pfrender, Donald Gilbert, W Kelley Thomas, Abraham Tucker, Todd H Oakley, Shinichi Tokishita, Andrea Aerts, Georg J

Arnold, Malay Kumar Basu, Darren J Bauer, Carla E Caceres, Liran Carmel, Claudio Casola, Jeong-Hyeon Choi, John C Detter, Qunfeng Dong, Serge Dusheyko, Brian D Eads, Thomas Froehlich, Kerry A Geiler-Samerotte, Daniel Gerlach, Phil Hatcher, Sanjuro Jogdeo, Jeroen Krijgsveld, Evgenia V Kriventseva, Dietmar Kueltz, Christian Laforsch, Erika Lindquist, Jacqueline Lopez, J Robert Manak, Jean Muller, Jasmyn Pangilinan, Rupali P Patwardhan, Samuel Pitluck, Ellen J Pritham, Andreas Rechtsteiner, Mina Rho, Igor B Rogozin, Onur Sakarya, Asaf Salamov, Sarah Schaack, Harris Shapiro, Yasuhiro Shiga, Courtney Skalitzky, Zachary Smith, Alexander Souvorov, Way Sung, Zuojian Tang, Dai Tsuchiya, Hank Tu, Harmjan Vos, Mei Wang, Yuri I Wolf, Hideo Yamagata, Takuji Yamada, Yuzhen Ye, Joseph R Shaw, Justen Andrews, Teresa J Crease, Haixu Tang, Susan M Lucas, Hugh M Robertson, Peer Bork, Eugene V Koonin, Evgeny M Zdobnov, Igor V Grigoriev, Michael Lynch, and Jeffrey L Boore. The Ecoresponsive Genome of Daphnia pulex. *Science*, 331(6017):555–561, 2011.

[11] K Meusemann, B M von Reumont, S Simon, F Roeding, S Strauss, P Kuck, I Ebersberger, M Walzl, G Pass, S Breuers, V Achter, A von Haeseler, T Burmester, H Hadrys, J W Wagele, and B Misof. A Phylogenomic Approach to Resolve the Arthropod Tree of Life. *Molecular biology and evolution*, 27(11):2451–2464, October 2010.

[12] Jerome C Regier, Jeffrey W Shultz, Andreas Zwick, April Hussey, Bernard Ball, Regina Wetzer, Joel W Martin, and Clifford W Cunningham. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature*, 463(7284):1079–1083, February 2010.

[13] T H Oakley, J M Wolfe, A R Lindgren, and A K Zaharoff. Phylotranscriptomics to Bring the Understudied into the Fold: Monophyletic Ostracoda, Fossil Placement, and Pancrustacean Phylogeny. *Molecular biology and evolution*, 30(1):215–233, December 2012.

[14] Bjoern M von Reumont, Ronald A Jenner, Matthew A Wills, Emiliano Dell'ampio, Günther Pass, Ingo Ebersberger, Benjamin Meyer, Stefan Koenemann, Thomas M Iliffe, Alexandros Stamatakis, Oliver Niehuis, Karen Meusemann, and Bernhard Misof. Pancrustacean phylogeny in the light of new phylogenomic data: support for Remipedia as the possible sister group of Hexapoda. *Molecular biology and evolution*, 29(3):1031–1045, March 2012.

[15] Lorena Rivarola-Duarte, Christian Otto, Frank Jühling, Stephan Schreiber, Daria Bedulina, Lena Jakob, Anton Gurkov, Denis Axenov-Gribanov, Abdullah H. Sahyoun, Magnus Lucassen, Jörg Hackermüller, Steve Hoffmann, Franz Sartoris, Hans-Otto Pörtner, Maxim Timofeyev, Till Luckenbach, and Peter F. Stadler. A first glimpse at the genome of the baikalian amphipodEulimnogammarus verrucosus. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 322(3):177–189, feb 2014.

[16] Nathan Kenny, Yung Sin, Xin Shen, Qu Zhe, Wei Wang, Ting Chan, Stephen Tobe, Sebastian Shimeld, Ka Chu, and Jerome Hui. Genomic sequence and experimental tractability of a new decapod shrimp model, neocaridina denticulata. *Marine Drugs*, 12(3):1419–1437, mar 2014.

[17] Le Cong, F Ann Ran, David Cox, Shuailiang Lin, Robert Barretto, Naomi Habib, Patrick D Hsu, Xuebing Wu, Wenyan Jiang, Luciano A Marraffini, and Feng Zhang. Multiplex genome engineering using CRISPR/Cas systems. *Science*, 339(6121):819–823, February 2013.

[18] Julia M Serano, Arnaud Martin, Danielle M Liubicich, Erin Jarvis, Heather S Bruce, Konnor La, William E Browne, Jane Grimwood, and Nipam H Patel. Comprehensive analysis of Hox gene expression in the amphipod crustacean Parhyale hawaiensis. *Developmental Biology*, pages 1–13, November 2015.

[19] Arnaud Martin, Julia M Serano, Erin Jarvis, Heather S Bruce, Jennifer Wang, Shagnik Ray, Carryn A Barker, Liam C O'Connell, and Nipam H Patel. CRISPR/Cas9 Mutagenesis Reveals Versatile Roles of Hox Genes in Crustacean Limb Specification and Evolution. *Current biology : CB*, December 2015.

[20] Prashant Mali, Luhan Yang, Kevin M Esvelt, John Aach, Marc Guell, James E DiCarlo, Julie E Norville, and George M Church. RNA-guided human genome engineering via Cas9. *Science*, 339(6121):823–826, February 2013.

[21] Martin Jinek, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A Doudna, and Emmanuelle Charpentier. A programmable dual-RNA-guided DNA

endonuclease in adaptive bacterial immunity. *Science*, 337(6096):816–821, August 2012.

[22] Anna F Gilles and Michalis Averof. Functional genetics for all: engineered nucleases, CRISPR and the gene editing revolution. *EvoDevo*, 5(1):43–13, 2014.

[23] M Averof and N H Patel. Crustacean appendage evolution associated with changes in Hox gene expression. *Nature*, 388(6643):682–686, 1997.

[24] Danielle M Liubicich, Julia M Serano, Anastasios Pavlopoulos, Zacharias Kontarakis, Meredith E Protas, Elaine Kwan, Sandip Chatterjee, Khoa D Tran, Michalis Averof, and Nipam H Patel. Knockdown of Parhyale Ultrabithorax recapitulates evolutionary changes in crustacean appendage morphology. *Proceedings of the National Academy of Sciences of the United States of America*, 106(33):13892–13896, August 2009.

[25] Anastasios Pavlopoulos, Zacharias Kontarakis, Danielle M Liubicich, Julia M Serano, Michael Akam, Nipam H Patel, and Michalis Averof. Probing the evolution of appendage specialization by Hox gene misexpression in an emerging model crustacean. *Proceedings of the National Academy of Sciences of the United States of America*, 106(33):13897–13902, August 2009.

[26] Nikolaos Konstantinides and Michalis Averof. A common cellular basis for muscle regeneration in arthropods and vertebrates. *Science*, 343(6172):788–791, February 2014.

[27] Jeanne L. Benton, Rachel Kery, Jingjing Li, Chadanat Noonin, Irene Söderhäll, and Barbara S. Beltz. Cells from the Immune System Generate Adult-Born Neurons in Crayfish. *Developmental Cell*, 30(3):322–333, August 2014.

[28] L Vazquez, J Alpuche, G Maldonado, C Agundis, A Pereyra-Morales, and E Zenteno. Review: Immunity mechanisms in crustaceans. *Innate Immunity*, 15(3):179–188, May 2009.

[29] Chris Hauton. The scope of the crustacean immune system for disease control. *Journal of Invertebrate Pathology*, 110(2):251–260, June 2012.

[30] Andrew J King, Simon M Cragg, Yi Li, Jo Dymond, Matthew J Guille, Dianna J Bowles, Neil C Bruce, Ian A Graham, and Simon J McQueen-Mason. Molecular insight into lignocellulose digestion by a marine isopod in the absence of gut microbes. *Proceedings of the National Academy of Sciences*, 107(12):5345–5350, March 2010.

[31] Marcelo Kern, John E. McGeehan, Simon D. Streeter, Richard N. A. Martin, Katrin Besser, Luisa Elias, Will Eborall, Graham P. Malyon, Christina M. Payne, Michael E. Himmel, Kirk Schnorr, Gregg T. Beckham, Simon M. Cragg, Neil C. Bruce, and Simon J. McQueen-Mason. Structural characterization of a unique marine animal family 7 cellobiohydrolase suggests a mechanism of cellulase salt tolerance. *Proceedings of the National Academy of Sciences of the United States of America*, 110(25):10189–10194, June 2013.

[32] P J Boyle and R Mitchell. Absence of Microorganisms in Crustacean Digestive Tracts. *Science*, 200(4346):1157–1159, 1978.

[33] M Zimmer, J Danko, S Pennings, A Danford, and T Carefoot. Cellulose digestion and phenol oxidation in coastal isopods (Crustacea: Isopoda). *Marine Biology*, 2002.

[34] Carsten Wolff and Matthias Gerberding. “Crustacea”: Comparative Aspects of Early Development. In *Evolutionary Developmental Biology of Invertebrates 4*, pages 39–61. Springer Vienna, Vienna, 2015.

[35] William E Browne, Alivia L Price, Matthias Gerberding, and Nipam H Patel. Stages of embryonic development in the amphipod crustacean, Parhyale hawaiensis. *Genesis (New York, N.Y. : 2000)*, 42(3):124–149, July 2005.

[36] Matthias Gerberding, William E Browne, and Nipam H Patel. Cell lineage analysis of the amphipod crustacean Parhyale hawaiensis reveals an early restriction of cell fates. *Development*, 129(24):5789–5801, December 2002.

[37] Cassandra G Extavour. The fate of isolated blastomeres with respect to germ cell formation in the amphipod crustacean Parhyale hawaiensis. *Developmental Biology*, 277(2):387–402, January 2005.

[38] E J Rehm, R L Hannibal, R C Chaw, M A Vargas-Vila, and N H Patel. Fixation and Dissection of Parhyale hawaiensis Embryos. *Cold Spring Harbor Protocols*, 2009(1):pdb.prot5127–pdb.prot5127, January 2009.

[39] E J Rehm, R L Hannibal, R C Chaw, M A Vargas-Vila, and N H Patel. Antibody Staining of Parhyale hawaiensis Embryos. *Cold Spring Harbor Protocols*, 2009(1):pdb.prot5129–pdb.prot5129, January 2009.

[40] E J Rehm, R L Hannibal, R C Chaw, M A Vargas-Vila, and N H Patel. In Situ Hybridization of Labeled RNA Probes to Fixed Parhyale hawaiensis Embryos. *Cold Spring Harbor Protocols*, 2009(1):pdb.prot5130–pdb.prot5130, January 2009.

[41] E Jay Rehm, Roberta L Hannibal, R Crystal Chaw, Mario A Vargas-Vila, and Nipam H Patel. Injection of Parhyale hawaiensis blastomeres with fluorescently labeled tracers. *Cold Spring Harbor Protocols*, 2009(1):pdb.prot5128–pdb.prot5128, January 2009.

[42] Alivia L Price, Melinda S Modrell, Roberta L Hannibal, and Nipam H Patel. Mesoderm and ectoderm lineages in the crustacean Parhyale hawaiensis display intra-germ layer compensation. *Developmental Biology*, 341(1):256–266, May 2010.

[43] Frederike Alwes, Billy Hinchen, and Cassandra G Extavour. Patterns of cell lineage, movement, and migration from germ layer specification to gastrulation in the amphipod crustacean Parhyale hawaiensis. *Developmental Biology*, 359(1):110–123, November 2011.

[44] Roberta L Hannibal, Alivia L Price, and Nipam H Patel. The functional relationship between ectodermal and mesodermal segmentation in the crustacean, Parhyale hawaiensis. *Developmental Biology*, 361(2):427–438, January 2012.

[45] Zacharias Kontarakis and Anastasios Pavlopoulos. Transgenesis in Non-model Organisms: The Case of Parhyale. In *Molecular Methods for Evolutionary Genetics*, pages 145–181. Springer New York, New York, NY, July 2014.

[46] Anastasia R Nast and Cassandra G Extavour. Ablation of a Single Cell From Eight-cell Embryos of the Amphipod Crustacean Parhyale hawaiensis. *Journal of visualized experiments : JoVE*, (85), 2014.

[47] R Crystal Chaw and Nipam H Patel. Independent migration of cell populations in the early gastrulation of the amphipod crustacean Parhyale hawaiensis. *Developmental Biology*, 371(1):94–109, November 2012.

[48] Anastasios Pavlopoulos and Michalis Averof. Establishing genetic transformation for comparative developmental studies in the crustacean Parhyale hawaiensis. *Proceedings of the National Academy of Sciences of the United States of America*, 102(22):7888–7893, May 2005.

[49] Zacharias Kontarakis, Anastasios Pavlopoulos, Alexandros Kiupakis, Nikolaos Konstantinides, Vassilis Douris, and Michalis Averof. A versatile strategy for gene trapping and trap conversion in emerging model organisms. *Development*, 138(12):2625–2630, June 2011.

[50] Günes Özhan-Kizil, Johanna Havemann, and Matthias Gerberding. Germ cells in the crustacean Parhyale hawaiensis depend on Vasa protein for their maintenance but not for their formation. *Developmental Biology*, 327(1):230–239, March 2009.

[51] Ronald J Parchem, Francis Poulin, Andrew B Stuart, Chris T Amemiya, and Nipam H Patel. BAC library for the amphipod crustacean, Parhyale hawaiensis. *Genomics*, 95(5):261–267, May 2010.

[52] Xianhui Wang, Xiaodong Fang, Pengcheng Yang, Xuanting Jiang, Feng Jiang, Dejian Zhao, Bolei Li, Feng Cui, Jianing Wei, Chuan Ma, Yundan Wang, Jing He, Yuan Luo, Zhifeng Wang, Xiaojiao Guo, Wei Guo, Xuesong Wang, Yi Zhang, Meiling Yang, Shuguang Hao, Bing Chen, Zongyuan Ma, Dan Yu, Zhiqiang Xiong, Yabing Zhu, Dingding Fan, Lijuan Han, Bo Wang, Yuanxin Chen, Junwen Wang, Lan Yang, Wei Zhao, Yue Feng, Guanxing Chen, Jinmin Lian, Qiye Li, Zhiyong Huang, Xiaoming Yao, Na Lv, Guojie Zhang, Yingrui Li, Jian Wang, Jun Wang, Baoli Zhu, and Le Kang. The locust genome provides insight into swarm formation and long-distance flight. *Nature communications*, 5:2957–2959, 2014.

[53] J. T. Simpson and R. Durbin. Efficient de novo assembly of large genomes using compressed data structures. *Genome Research*, 22(3):549–556, dec 2011.

[54] Jared T Simpson. Exploring genome characteristics and sequence quality without a reference. *Bioinformatics*, 30(9):1228–1235, May 2014.

[55] Guofan Zhang, Xiaodong Fang, Ximing Guo, Li Li, Ruibang Luo, Fei Xu, Pengcheng Yang, Linlin Zhang, Xiaotong Wang, Haigang Qi, Zhiqiang Xiong, Huayong Que, Yinlong Xie, Peter W H Holland, Jordi Paps, Yabing Zhu, Fucun Wu, Yuanxin Chen, Jiafeng Wang, Chunfang Peng, Jie Meng, Lan Yang, Jun Liu, Bo Wen, Na Zhang, Zhiyong Huang, Qihui Zhu, Yue Feng, Andrew Mount, Dennis Hedgecock, Zhe Xu, Yunjie Liu, Tomislav Domazet-Lošo, Yishuai Du, Xiaoqing Sun, Shoudu Zhang, Binghang Liu, Peizhou Cheng, Xuanting Jiang, Juan Li, Dingding Fan, Wei Wang, Wenjing Fu, Tong Wang, Bo Wang, Jibiao Zhang, Zhiyu Peng, Yingxiang Li, Na Li, Jinpeng Wang, Maoshan Chen, Yan He, Fengji Tan, Xiaorui Song, Qiumei Zheng, Ronglian Huang, Hailong Yang, Xuedi Du, Li Chen, Mei Yang, Patrick M Gaffney, Shan Wang, Longhai Luo, Zhicai She, Yao Ming, Wen Huang, Shu Zhang, Baoyu Huang, Yong Zhang, Tao Qu, Peixiang Ni, Guoying Miao, Junyi Wang, Qiang Wang, Christian E W Steinberg, Haiyan Wang, Ning Li, Lumin Qian, Guojie Zhang, Yingrui Li, Huanming Yang, Xiao Liu, Jian Wang, Ye Yin, and Jun Wang. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature*, 490(7418):49–54, September 2012.

[56] Rei Kajitani, Kouta Toshimoto, Hideki Noguchi, Atsushi Toyoda, Yoshitoshi Ogura, Miki Okuno, Mitsuru Yabana, Masayuki Harada, Eiji Nagayasu, Haruhiko Maruyama, Yuji Kohara, Asao Fujiyama, Tetsuya Hayashi, and Takehiko Itoh. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research*, 24(8):1384–1395, apr 2014.

[57] Brian J Haas, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood, Joshua Bowden, Matthew Brian Couger, David Eccles, Bo Li, Matthias Lieber, Matthew D MacManes, Michael Ott, Joshua Orvis, Nathalie Pochet, Francesco Strozzi, Nathan Weeks, Rick Westerman, Thomas William, Colin N Dewey, Robert Henschel, Richard D LeDuc, Nir Friedman, and Aviv Regev. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform

for reference generation and analysis. *Nature Protocols*, 8(8):1494–1512, July 2013.

[58] G Parra, K Bradnam, and I Korf. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, 23(9):1061–1067, May 2007.

[59] David M Emms and Steven Kelly. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome biology*, 16:157, 2015.

[60] Maura Strigini, Rafael Cantera, Xavier Morin, Michael J Bastiani, Michael Bate, and Domna Karagogeos. The IgLON protein Lachesin is required for the blood-brain barrier in Drosophila. *Molecular and cellular neurosciences*, 32(1-2):91–101, May 2006.

[61] Lindsey S Garver, Zhiyong Xi, and George Dimopoulos. Immunoglobulin superfamily members play an important role in the mosquito immune system. *Developmental & Comparative Immunology*, 32(5):519–531, 2008.

[62] Matthias Siebert, Daniel Banovic, Bernd Goellner, and Hermann Aberle. Drosophila motor axons recognize and follow a Sidestep-labeled substrate pathway to reach their target fields. *Genes & development*, 23(9):1052–1062, May 2009.

[63] C Deraison, I Darboux, L Duportets, T Gorojankina, Y Rahbe, and L Jouanin. Cloning and characterization of a gut-specific cathepsin L from the aphid Aphis gossypii. *Insect Molecular Biology*, 13(2):165–177, April 2004.

[64] B Prud’homme, N Lartillot, G Balavoine, and A Adoutte. Phylogenetic analysis of the Wnt gene family: insights from lophotrochozoan members. *Current Biology*, 12(16):1395–1400, 2002.

[65] Sung-Jin Cho, Yvonne Vallès, Vincent C Giani, Elaine C Seaver, and David A Weisblat. Evolutionary dynamics of the wnt gene family: a lophotrochozoan perspective. *Molecular biology and evolution*, 27(7):1645–1658, July 2010.

[66] Ralf Janssen, Martine Le Gouar, Matthias Pechmann, Francis Poulin, Renata Bolognesi, Evelyn E Schwager, Corinna Hopfen, John K Colbourne, Graham E Budd, Susan J Brown, Nikola-Michael Prpic, Carolin Kosiol, Michel Vervoort,

Wim GM Damen, Guillaume Balavoine, and Alistair P McGregor. Conservation, loss, and redeployment of wnt ligands in protostomes: implications for understanding the evolution of segment formation. *BMC Evol Biol*, 10(1):374, 2010.

[67] Massimo A Hilliard and Cornelia I Bargmann. Wnt Signals and Frizzled Activity Orient Anterior-Posterior Axon Outgrowth in C. elegans. *Developmental Cell*, 10(3):379–390, March 2006.

[68] Renata Bolognesi, Laila Farzana, Tamara D Fischer, and Susan J Brown. Multiple Wnt Genes Are Required for Segmentation in the Short-Germ Embryo of Tribolium castaneum. *Current Biology*, 18(20):1624–1629, October 2008.

[69] Mattias Hogvall, Anna Schönauer, Graham E Budd, Alistair P McGregor, Nico Posnien, and Ralf Janssen. Analysis of the wnt gene repertoire in an onychophoran provides new insights into the evolution of segmentation. *EvoDevo*, 5(1):14, 2014.

[70] Thomas W. Holstein. The evolution of the wnt pathway. *Cold Spring Harbor Perspectives in Biology*, 4(7), 2012.

[71] A K Ryan, B Blumberg, C Rodriguez-Esteban, S Yonei-Tamura, K Tamura, T Tsukui, J de la Pena, W Sabbagh, J Greenwald, S Choe, D P Norris, E J Robertson, R M Evans, M G Rosenfeld, and JCI Belmonte. Pitx2 determines left-right asymmetry of internal organs in vertebrates. *Nature*, 394(6693):545–551, 1998.

[72] Anja C Nagel, Alena Krejci, Gennady Tenin, Alejandro Bravo-Patiño, Sarah Bray, Dieter Maier, and Anette Preiss. Hairless-mediated repression of notch target genes requires the combined activity of Groucho and CtBP corepressors. *Molecular and cellular biology*, 25(23):10433–10441, December 2005.

[73] Victor Zeng, Karina E Villanueva, Ben S Ewen-Campen, Frederike Alwes, William E Browne, and Cassandra G Extavour. De novo assembly and characterization of a maternal and developmental transcriptome for the emerging model crustacean parhyale hawaiensis. *BMC Genomics*, 12(1):581, 2011.

[74] Ho-Ryun Chung, Ulrich Schäfer, Herbert Jäckle, and Siegfried Böhm. Genomic expansion and clustering of ZAD-containing C2H2 zinc-finger genes in Drosophila. *EMBO reports*, 3(12):1158–1162, December 2002.

[75] Hamed S Najafabadi, Sanie Mnaimneh, Frank W Schmitges, Michael Garton, Kathy N Lam, Ally Yang, Mihai Albu, Matthew T Weirauch, Ernest Radovani, Philip M Kim, Jack Greenblatt, Brendan J Frey, and Timothy R Hughes. C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nature Biotechnology*, 33(5):555–562, February 2015.

[76] Ariel D Chipman, David E K Ferrier, Carlo Brena, Jiaxin Qu, Daniel S T Hughes, Reinhard Schröder, Montserrat Torres-Oliva, Nadia Znassi, Huaiyang Jiang, Francisca C Almeida, Claudio R Alonso, Zivkos Apostolou, Peshtewani Aqrawi, Wallace Arthur, Jennifer C J Barna, Kerstin P Blankenburg, Daniela Brites, Salvador Capella-Gutiérrez, Marcus Coyle, Peter K Dearden, Louis Du Pasquier, Elizabeth J Duncan, Dieter Ebert, Cornelius Eibner, Galina Erikson, Peter D Evans, Cassandra G Extavour, Liezl Francisco, Toni Gabaldón, William J Gillis, Elizabeth A Goodwin-Horn, Jack E Green, Sam Griffiths-Jones, Cornelis J P Grimmelikhuijzen, Sai Gubbala, Roderic Guigó, Yi Han, Frank Hauser, Paul Havlak, Luke Hayden, Sophie Helbing, Michael Holder, Jerome H L Hui, Julia P Hunn, Vera S Hunnekuhl, LaRonda Jackson, Mehwish Javaid, Shalini N Jhangiani, Francis M Jiggins, Tamsin E Jones, Tobias S Kaiser, Divya Kalra, Nathan J Kenny, Viktoriya Korchina, Christie L Kovar, F Bernhard Kraus, François Lapraz, Sandra L Lee, Jie Lv, Christigale Mandapat, Gerard Manning, Marco Mariotti, Robert Mata, Tittu Mathew, Tobias Neumann, Irene Newsham, Dinh N Ngo, Maria Ninova, Geoffrey Okwuonu, Fiona Ongeri, William J Palmer, Shobha Patil, Pedro Patraquim, Christopher Pham, Ling-Ling Pu, Nicholas H Putman, Catherine Rabouille, Olivia Mendivil Ramos, Adelaide C Rhodes, Helen E Robertson, Hugh M Robertson, Matthew Ronshaugen, Julio Rozas, Nehad Saada, Alejandro Sánchez-Gracia, Steven E Scherer, Andrew M Schurko, Kenneth W Siggens, DeNard Simmons, Anna Stief, Eckart Stolle, Maximilian J Telford, Kristin Tessmar-Raible, Rebecca Thornton, Maurijn van der Zee, Arndt von Haeseler, James M Williams, Judith H Willis, Yuanqing Wu, Xiaoyan Zou, Daniel Lawson,

Donna M Muzny, Kim C Worley, Richard A Gibbs, Michael Akam, and Stephen Richards. The First Myriapod Genome Sequence Reveals Conservative Arthropod Gene Content and Genome Organisation in the Centipede Strigamia maritima. *PLoS biology*, 12(11):e1002005–24, November 2014.

[77] Y Pewzner-Jung, S Ben-Dor, and A H Futerman. When Do Lasses (Longevity Assurance Genes) Become CerS (Ceramide Synthases)?: INSIGHTS INTO THE REGULATION OF CERAMIDE SYNTHESIS. *Journal of Biological Chemistry*, 281(35):25001–25005, August 2006.

[78] Peter WH Holland, H Anne F Booth, and Elspeth A Bruford. Classification and nomenclature of all human homeobox genes. *BMC biology*, 5(1):47–28, 2007.

[79] Ying-fu Zhong and Peter W H Holland. HomeoDB2: functional expansion of a comparative homeobox gene database for evolutionary developmental biology. *Evolution & Development*, 13(6):567–568, November 2011.

[80] Dave Kosman, Claudia M Mizutani, Derek Lemons, W Gregory Cox, William McGinnis, and Ethan Bier. Multiplex detection of RNA expression in Drosophila embryos. *Science*, 305(5685):846, August 2004.

[81] Matthew Ronshaugen and Mike Levine. Visualization of trans-Homolog Enhancer-Promoter Interactions at the Abd-B Hox Locus in the Drosophila Embryo. *Developmental Cell*, 7(6):925–932, December 2004.

[82] M. Kmita. Organizing axes in time and space; 25 years of colinear tinkering. *Science*, 301(5631):331–333, jul 2003.

[83] N M Brooke, J Garcia-Fernandez, and PWH Holland. The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. *Nature*, 392(6679):920–922, 1998.

[84] S L Pollard and P W Holland. Evidence for 14 homeobox gene clusters in human genome ancestry. *Current Biology*, 10(17):1059–1062, September 2000.

[85] K Jagla, M Bellard, and M Frasch. A cluster of Drosophila homeobox genes involved in mesoderm differentiation programs. *BioEssays*, 23(2):125–133, February 2001.

[86] G N Luke, L F C Castro, K McLay, C Bird, A Coulson, and P W H Holland. Dispersal of NK homeobox gene clusters in amphioxus and humans. *Proceedings of the National Academy of Sciences of the United States of America*, 100(9):1–4, April 2003.

[87] L F C Castro and P W H Holland. Chromosomal mapping of ANTP class homeobox genes in amphioxus: piecing together ancestral genomes. *Evolution & Development*, 5(5):1–7, August 2003.

[88] Michael E Himmel, Shi-You Ding, David K Johnson, William S Adney, Mark R Nimlos, John W Brady, and Thomas D Foust. Biomass recalcitrance: Engineering plants and enzymes for biofuels production. *Science*, 315(5813):804–807, 2007.

[89] David B Wilson. Microbial diversity of cellulose hydrolysis. *Current Opinion in Microbiology*, 14(3):259–263, June 2011.

[90] Simon M Cragg, Gregg T Beckham, Neil C Bruce, Timothy DH Bugg, Daniel L Distel, Paul Dupree, Amaia Green Etxabe, Barry S Goodell, Jody Jellison, John E McGeehan, Simon J McQueen-Mason, Kirk Schnorr, Paul H Walton, Joy EM Watts, and Martin Zimmer. ScienceDirect Lignocellulose degradation mechanisms across the Tree of Life. *Current Opinion in Chemical Biology*, 29(C):108–119, December 2015.

[91] C J Duan, L Xian, G C Zhao, Y Feng, H Pang, X L Bai, J L Tang, Q S Ma, and J X Feng. Isolation and partial characterization of novel genes encoding acidic cellulases from metagenomes of buffalo rumens. *Journal of Applied Microbiology*, 107(1):245–256, July 2009.

[92] Falk Warnecke, Peter Luginbühl, Natalia Ivanova, Majid Ghassemian, Toby H Richardson, Justin T Stege, Michelle Cayouette, Alice C McHardy, Gordana Djordjevic, Nahla Aboushadi, Rotem Sorek, Susannah G Tringe, Mircea Podar, Hector Garcia Martin, Victor Kunin, Daniel Dalevi, Julita Madejska, Edward Kirton, Darren Platt, Ernest Szeto, Asaf Salamov, Kerrie Barry, Natalia Mikhailova, Nikos C Kyrpides, Eric G Matson, Elizabeth A Ottesen, Xinning Zhang, Myriam Hernández, Catalina Murillo, Luis G Acosta, Isidore Rigoutsos, Giselle Tamayo, Brian D Green, Cathy Chang, Edward M Rubin, Eric J Mathur,

Dan E Robertson, Philip Hugenholtz, and Jared R Leadbetter. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature*, 450(7169):560–565, November 2007.

[93] Daniel L Distel, Mehwish Amin, Adam Burgoyne, Eric Linton, Gustaf Mamangkey, Wendy Morrill, John Nove, Nicole Wood, and Joyce Yang. Molecular phylogeny of Pholadoidea Lamarck, 1809 supports a single origin for xylotrophy (wood feeding) and xylotrophic bacterial endosymbiosis in Bivalvia. *Molecular Phylogenetics and Evolution*, 61(2):245–254, November 2011.

[94] Amaia Green Etxabe. *The wood boring amphipod Chelura (terebrans)*. PhD thesis, University of Portsmouth, 2013.

[95] B L Cantarel, P M Coutinho, C Rancurel, T Bernard, V Lombard, and B Henrissat. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Research*, 37(Database):D233–D238, January 2009.

[96] Robert D. Finn, Jaina Mistry, Benjamin Schuster-Böckler, Sam Griffiths-Jones, Volker Hollich, Timo Lassmann, Simon Moxon, Mhairi Marshall, Ajay Khanna, Richard Durbin, Sean R. Eddy, Erik L. L. Sonnhammer, and Alex Bateman. Pfam: clans, web tools and services. *Nucleic Acids Research*, 34(Database issue):D247–251, January 2006.

[97] G D Stentiford, D M Neil, E J Peeler, J D Shields, H J Small, T W Flegel, J M Vlak, B Jones, F Morado, S Moss, J Lotz, L Bartholomay, D C Behringer, C Hauton, and D V Lightner. Disease will limit future food supply from the global crustacean fishery and aquaculture sectors. *Journal of Invertebrate Pathology*, 110(2):141–157, June 2012.

[98] Robert M Waterhouse, Evgenia V Kriventseva, Stephan Meister, Zhiyong Xi, Kanwal S Alvarez, Lyric C Bartholomay, Carolina Barillas-Mury, Guowu Bian, Stephanie Blandin, Bruce M Christensen, Yuemei Dong, Haobo Jiang, Michael R Kanost, Anastasios C Koutsos, Elena A Levashina, Jianyong Li, Petros Ligoxygakis, Robert M Maccallum, George F Mayhew, Antonio Mendes, Kristin Michel, Mike A Osta, Susan Paskewitz, Sang Woon Shin, Dina Vlachou, Lihui

Wang, Weiqi Wei, Liangbiao Zheng, Zhen Zou, David W Severson, Alexander S Raikhel, Fotis C Kafatos, George Dimopoulos, Evgeny M Zdobnov, and George K Christophides. Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. *Science*, 316(5832):1738–1743, June 2007.

[99] Charles A Janeway and Ruslan Medzhitov. Innate immune recognition. *Annual review of immunology*, 20:197–216, 2002.

[100] T Werner, K Borge-Renberg, P Mellroth, H Steiner, and D Hultmark. Functional Diversity of the Drosophila PGRP-LC Gene Cluster in the Response to Lipopolysaccharide and Peptidoglycan. *Journal of Biological Chemistry*, 278(29):26319–26322, July 2003.

[101] C Liu, Z Xu, D Gupta, and R Dziarski. Peptidoglycan Recognition Proteins: A novel family of four human innate immunity pattern recognition molecules. *Journal of Biological Chemistry*, 276(37):34686–34694, September 2001.

[102] Abdur Rehman, Ping Taishi, Jidong Fang, Jeannine A Majde, and James M Krueger. The cloning of a rat peptidoglycan recognition protein (PGRP) and its induction in brain by sleep deprivation. *Cytokine*, 13(1):8–17, January 2001.

[103] Haipeng Liu, Chenglin Wu, Yasuyuki Matsuda, Shun-ichiro Kawabata, Bok Luel Lee, Kenneth Söderhäll, and Irene Söderhäll. Peptidoglycan activation of the proPO-system without a peptidoglycan receptor protein (PGRP)? *Developmental & Comparative Immunology*, 35(1):51–61, January 2011.

[104] Seanna J McTaggart, Claire Conlon, John K Colbourne, Mark L Blaxter, and Tom J Little. The components of the Daphnia pulex immune system as revealed by complete genome sequencing. *BMC Genomics*, 10(1):175–119, 2009.

[105] Catherine Dostert, Emmanuelle Jouanguy, Phil Irving, Laurent Troxler, Delphine Galiana-Arnoux, Charles Hetru, Jules A Hoffmann, and Jean-Luc Imler. The Jak-STAT signaling pathway is required but not sufficient for the antiviral response of drosophila. *Nature Immunology*, 6(9):946–953, August 2005.

[106] T Tanji, X Hu, A N R Weber, and Y T Ip. Toll and IMD Pathways Synergistically

Activate an Innate Immune Response in Drosophila melanogaster. *Molecular and cellular biology*, 27(12):4578–4588, May 2007.

[107] Matthew A. Benton, Matthias Pechmann, Nadine Frey, Dominik Stappert, Kai H. Conrads, Yen-Ta Chen, Evangelia Stamataki, Anastasios Pavlopoulos, and Siegfried Roth. Toll genes have an ancestral role in axis elongation. *Current Biology*, 26(12):1609 – 1615, 2016.

[108] Natalia I Arbouzova and Martin P Zeidler. JAK/STAT signalling in Drosophila: insights into conserved regulatory and cellular functions. *Development*, 133(14):2605–2616, July 2006.

[109] E A Levashina, L F Moita, S Blandin, G Vriend, M Lagueux, and F C Kafatos. Conserved role of a complement-like protein in phagocytosis revealed by dsRNA knockout in cultured cells of the mosquito, Anopheles gambiae. *Cell*, 104(5):709–718, 2001.

[110] H Decker. Recent findings on phenoloxidase activity and antimicrobial activity of hemocyanins. *Developmental & Comparative Immunology*, 28(7-8):673–687, June 2004.

[111] So Young Lee, Bok Luel Lee, and Kenneth Söderhäll. Processing of crayfish hemocyanin subunits into phenoloxidase. *Biochemical and Biophysical Research Communications*, 322(2):490–496, September 2004.

[112] D Schmucker, J C Clemens, H Shu, C A Worby, J Xiao, M Muda, J E Dixon, and S L Zipursky. Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell*, 101(6):671–684, June 2000.

[113] Fiona L Watson, Roland Püttmann-Holgado, Franziska Thomas, David L Lamar, Michael Hughes, Masahiro Kondo, Vivienne I Rebel, and Dietmar Schmucker. Extensive diversity of Ig-superfamily proteins in the immune system of insects. *Science*, 309(5742):1874–1878, September 2005.

[114] Daniela Brites, Seanna McTaggart, Krystalynne Morris, Jobriah Anderson, Kelley Thomas, Isabelle Colson, Thomas Fabbro, Tom J Little, Dieter Ebert, and Louis Du Pasquier. The Dscam homologue of the crustacean Daphnia is diversified by

alternative splicing like in insects. *Molecular biology and evolution*, 25(7):1429–1439, July 2008.

[115] Stephane E Castel and Robert A Martienssen. RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. *Nature reviews. Genetics*, 14(2):100–112, February 2013.

[116] A. A. Aravin, N. M. Naumova, A. V. Tulin, V. V. Vagin, Y. M. Rozovsky, and V. A. Gvozdev. Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the D. melanogaster germline. *Current biology: CB*, 11(13):1017–1027, July 2001.

[117] N J Caplen, S Parrish, F Imani, A Fire, and R A Morgan. Specific inhibition of gene expression by small double-stranded RNAs in invertebrate and vertebrate systems. *Proceedings of the National Academy of Sciences of the United States of America*, 98(17):1–7, August 2001.

[118] Julius Brennecke, Alexei A Aravin, Alexander Stark, Monica Dus, Manolis Kellis, Ravi Sachidanandam, and Gregory J Hannon. Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila. *Cell*, 128(6):1089–1103, March 2007.

[119] Weifeng Gu, Masaki Shirayama, Darryl Conte Jr, Jessica Vasale, Pedro J Batista, Julie M Claycomb, James J Moresco, Elaine M Youngman, Jennifer Keys, Matthew J Stoltz, Chun-Chieh G Chen, Daniel A Chaves, Shenghua Duan, Kristin D Kasschau, Noah Fahlgren, John R Yates III, Shohei Mitani, James C Carrington, and Craig C Mello. Distinct Argonaute-Mediated 22G-RNA Pathways Direct Genome Surveillance in the C. elegans Germline. *Molecular cell*, 36(2):231–244, October 2009.

[120] Heng-Chi Lee, Weifeng Gu, Masaki Shirayama, Elaine Youngman, Darryl Conte, and Craig C Mello. C. elegans piRNAs mediate the genome-wide surveillance of germline transcripts. *Cell*, 150(1):78–87, July 2012.

[121] Lin He and Gregory J Hannon. MicroRNAs: small RNAs with a big role in gene regulation. *Nature reviews. Genetics*, 5(7):522–531, July 2004.

[122] J Michael Thomson, Martin Newman, Joel S Parker, Elizabeth M Morin-Kensicki, Tricia Wright, and Scott M Hammond. Extensive post-transcriptional regulation of microRNAs and its implications for cancer. *Genes & development*, 20(16):2202–2207, August 2006.

[123] Witold Filipowicz, Suvendra N Bhattacharyya, and Nahum Sonenberg. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nature reviews. Genetics*, 2008(2):102–114, February 2008.

[124] Peter Sarkies, Murray E Selkirk, John T Jones, Vivian Blok, Thomas Boothby, Bob Goldstein, Ben Hanelt, Alex Ardila-Garcia, Naomi M Fast, Phillip M Schiffer, Christopher Kraus, Mark J Taylor, Georgios Koutsovoulos, Mark L Blaxter, and Eric A Miska. Ancient and Novel Small RNA Pathways Compensate for the Loss of piRNAs in Multiple Independent Nematode Lineages. *PLoS biology*, 13(2):e1002061–20, February 2015.

[125] Ying Dong and Markus Friedrich. Nymphal RNAi: systemic RNAi mediated gene knockdown in juvenile grasshopper. *BMC Biotechnology*, 5:25, 2005.

[126] George M Weinstock, Gene E Robinson, Richard A Gibbs, George M Weinstock, George M Weinstock, Gene E Robinson, Kim C Worley, Hugh M Robertson, Daniel B Weaver, Martin Beye, Peer Bork, Jay D Evans, Klaus Hartfelder, Greg J Hunt, Gene E Robinson, Ryszard Maleszka, George M Weinstock, Klaus Hartfelder, Gro V Amdam, Mrcia M G Bitondi, Anita M Collins, Alexandre S Cristino, H Michael, G Lattorff, Carlos H Lobo, Robin F A Moritz, Francis M F Nunes, Robert E Page, Zil L P Simões, Diana Wheeler, Piero Carninci, Shiro Fukuda, Yoshihide Hayashizaki, Chikatoshi Kai, Jun Kawai, Naoko Sakazume, Daisuke Sasaki, Michihira Tagami, Gro V Amdam, Stefan Albert, Geert Baggerman, Kyle T Beggs, Guy Bloch, Giuseppe Cazzamali, Mira Cohen, Mark David Drapeau, Dorothea Eisenhardt, Christine Emore, Michael A Ewing, Susan E Fahrbach, Sylvain Foret, Cornelis J P Grimmelikhuijzen, Frank Hauser, Amanda B Hummon, Greg J Hunt, Jurgen Huybrechts, Andrew K Jones, Noam Kaplan, Gérard Leboulle, Michal Linial, J Troy Littleton, Alison R Mercer, Robert E Page, Gene E Robinson, Timothy A Richmond, Sandra L RodriguezZas, Elad B Rubin, David B Sattelle, David Schlipalius, Liliane Schoofs, Yair Shemesh,

Jonathan V Sweedler, Rodrigo Velarde, Peter Verleyen, Evy Vierstraete, Michael R Williamson, Martin Beye, Seth A Ament, Susan J Brown, Miguel Corona, Peter K Dearden, W Augustine Dunn, Michelle M Elekonich, Christine G Elsik, Tomoko Fujiyuki, Irene Gattermeier, Tanja Gempe, Martin Hasselmann, Tatsuhiko Kadowaki, Eriko Kage, Azusa Kamikouchi, Takeo Kubo, Robert Kucharski, Takekazu Kunieda, Marcé Lorenzen, Natalia V Milshina, Mizue Morioka, Kazuaki Ohashi, Ross Overbeek, Robert E Page, Gene E Robinson, Christian A Ross, Morten Schioett, Teresa Shippy, Hideaki Takeuchi, Amy L Toth, Judith H Willis, Megan J Wilson, Evgeny M Zdobnov, Karl H J Gordon, Ivica Letunic, Kevin Hackett, Jane Peterson, Adam Felsenfeld, Mark Guyer, Michel Solignac, Richa Agarwala, Jean Marie Cornuet, Christine Emore, Greg J Hunt, Monique Monnerot, Florence Mougel, Justin T Reese, David Schlipalius, Dominique Vautrin, Daniel B Weaver, Joseph J Gillespie, Jamie J Cannone, Robin R Gutell, J Spencer Johnston, Michael B Eisen, Amanda B Hummon, Venky N Iyer, Vivek Iyer, Peter Kosarev, Aaron J Mackey, Timothy A Richmond, Victor Solovyev, Alexandre Souvorov, George M Weinstock, Michael R Williamson, Katherine A Aronstein, Katarina Bilikova, Yan Ping Chen, Andrew G Clark, Laura I Decanini, William M Gelbart, Charles Hetru, Dan Hultmark, Jean-Luc Imler, Haobo Jiang, Michael Kanost, Kiyoshi Kimura, Brian P Lazzaro, Dawn L Lopez, Jozef Simuth, Graham J Thompson, Zhen Zou, Pieter De Jong, Erica Sodergren, Miklós Csűrös, Aleksandar Milosavljevic, J Spencer Johnston, Kazutoyo Osoegawa, Stephen Richards, Chung-Li Shu, George M Weinstock, Laurent Duret, Eran Elhaik, Dan Graur, Daniel B Weaver, Gro V Amdam, Juan M Anzola, Kathryn S Campbell, Kevin L Childs, Derek Collinge, Madeline A Crosby, C Michael Dickens, Karl H J Gordon, L Sian Gramates, Christina M Grozinger, Peter L Jones, Mireia Jorda, Xu Ling, Beverly B Matthews, Jonathan Miller, Natalia V Milshina, Craig Mizzen, Miguel A Peinado, Jeffrey G Reid, Gene E Robinson, Susan M Russo, Andrew J Schroeder, Susan E St Pierre, Ying Wang, Pinglei Zhou, Richa Agarwala, Natalia V Milshina, Daniel B Weaver, Kevin L Childs, C Michael Dickens, William M Gelbart, Huaiyang Jiang, Paul Kitts, Natalia V Milshina, Barbara Ruef, Susan M Russo, Anand Venkatraman, George M Weinstock, Lan Zhang, Pinglei Zhou, J Spencer Johnston, Gildardo Aquino-Perez, Jean Marie Cornuet, Monique

Monnerot, Michel Solignac, Dominique Vautrin, Charles W Whitfield, Susanta K Behura, Stewart H Berlocher, Andrew G Clark, J Spencer Johnston, Walter S Sheppard, Deborah R Smith, Andrew V Suarez, Neil D Tsutsui, and Daniel B and... Weaver. Insights into social insects from the genome of the honeybee Apis mellifera. *Nature*, 443(7114):931–949, October 2006.

[127] Weina Xu and Zhaojun Han. Cloning and phylogenetic analysis of sid-1-like genes from aphids. *Journal of insect science (Online)*, 8(30):1–6, 2008.

[128] J Y Roignant, C Carre, R Mugat, D Szymczak, J A Lepesant, and C Antoniewski. Absence of transitive and systemic pathways allows cell-specific and isoform-specific RNAi in Drosophila. *RNA*, 9(3):299–308, March 2003.

[129] Yonggan Wu, Bo Wei, Haizhou Liu, Tianxian Li, and Simon Rayner. MiRPara: a SVM-based software tool for prediction of most probable microRNA coding regions in genome scale sequences. *BMC bioinformatics*, 12(1):107, 2011.

[130] Eric P Nawrocki, Sarah W Burge, Alex Bateman, Jennifer Daub, Ruth Y Eberhardt, Sean R Eddy, Evan W Floden, Paul P Gardner, Thomas A Jones, John Tate, and Robert D Finn. Rfam 12.0: updates to the RNA families database. *Nucleic Acids Research*, 43(Database issue):D130–7, January 2015.

[131] W Wang, F G Brunet, E Nevo, and M Long. Origin of sphinx, a young chimeric RNA gene in Drosophila melanogaster. *Proceedings of the National Academy of Sciences of the United States of America*, 99(7):4448–4453, 2002.

[132] Martin J. Blythe, Sunir Malla, Richard Everall, Yu-Huan H. Shih, Virginie Lemay, Joanna Moreton, Raymond Wilson, and Aziz A. Aboobaker. High Through-Put sequencing of the parhyale hawaiensis mRNAs and microRNAs to aid comparative developmental studies. *PloS one*, 7(3), 2012.

[133] Benjamin M Wheeler, Alysha M Heimberg, Vanessa N Moy, Erik A Sperling, Thomas W Holstein, Steffen Heber, and Kevin J Peterson. The deep evolution of metazoan microRNAs. *Evolution & Development*, 11(1):50–68, January 2009.

[134] Andrew Grimson, Mansi Srivastava, Bryony Fahey, Ben J Woodcroft, H Rosaria Chiang, Nicole King, Bernard M Degnan, Daniel S Rokhsar, and David P Bartel.

Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature*, 455(7217):1193–1197, October 2008.

[135] Susanta K Behura. Insect microRNAs: Structure, function and evolution. *Insect Biochemistry and Molecular Biology*, 37(1):3–9, January 2007.

[136] Antonio Marco, Katarzyna Hooks, and Sam Griffiths-Jones. Evolution and function of the extended miR-2 microRNA family. *RNA Biology*, 9(3):242–248, November 2014.

[137] Anton J Enright, Bino John, Ulrike Gaul, Thomas Tuschl, Chris Sander, and Debora S Marks. MicroRNA targets in Drosophila. *Genome biology*, 5(1):R1, 2003.

[138] Andrea Tanzer, Chris T Amemiya, Chang-Bae Kim, and Peter F Stadler. Evolution of microRNAs located withinHox gene clusters. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 304B(1):75–85, 2005.

[139] Derek Lemons and William McGinnis. Genomic evolution of Hox gene clusters. *Science*, 313(5795):1918–1922, 2006.

[140] A Stark, N Bushati, C H Jan, P Kheradpour, E Hodges, J Brennecke, D P Bartel, S M Cohen, and M Kellis. A single Hox locus in Drosophila produces functional microRNAs from opposite DNA strands. *Genes & development*, 22(1):8–13, January 2008.

[141] Teresa D Shippy, Matthew Ronshaugen, Jessica Cande, JianPing He, Richard W Beeman, Michael Levine, Susan J Brown, and Robin E Denell. Analysis of the Tribolium homeotic complex: insights into mechanisms constraining insect Hox clusters. *Development Genes and Evolution*, 218(3-4):127–139, April 2008.

[142] Anton J Enright, Bino John, Ulrike Gaul, Thomas Tuschl, Chris Sander, and Debora S Marks. MicroRNA targets in Drosophila. *Genome biology*, 5(1):R1–14, 2003.

[143] S Cumberledge, A Zaratzian, and S Sakonju. Characterization of two RNAs transcribed from the cis-regulatory region of the abd-A domain within the

Drosophila bithorax complex. *Proceedings of the National Academy of Sciences of the United States of America*, 87(9):3259–3263, May 1990.

[144] Assaf Zemach, Ivy E McDaniel, Pedro Silva, and Daniel Zilberman. Genome-Wide Evolutionary Analysis of Eukaryotic DNA Methylation. *Science*, 328(5980):916–919, 2010.

[145] Julie A Law and Steven E Jacobsen. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature reviews. Genetics*, 11(3):204–220, February 2010.

[146] Peter A Jones. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature reviews. Genetics*, 13(7):484–492, May 2012.

[147] Peter A. Jones and Gangning Liang. Rethinking how DNA methylation patterns are maintained. *Nature Reviews Genetics*, 10(11):805–811, September 2009.

[148] Albert Jeltsch, Ann Ehrenhofer-Murray, Tomasz P. Jurkowski, Frank Lyko, Gunter Reuter, Serge Ankri, Wolfgang Nellen, Matthias Schaefer, and Mark Helm. Mechanism and biological role of dnmt2 in nucleic acid methylation. *RNA Biology*, 0(0):1–16, 0. PMID: 27232191.

[149] Mary Grace Goll, Finn Kirpekar, Keith A Maggert, Jeffrey A Yoder, Chih-Lin Hsieh, Xiaoyu Zhang, Kent G Golic, Steven E Jacobsen, and Timothy H Bestor. Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. *Science*, 311(5759):395–398, January 2006.

[150] Farah Jaber-Hijazi, Priscilla J K P Lo, Yuliana Mihaylova, Jeremy M Foster, Jack S Benner, Belen Tejada Romero, Chen Chen, Sunir Malla, Jordi Solana, Alexey Ruzov, and A Aziz Aboobaker. Planarian MBD2/3 is required for adult stem cell pluripotency independently of DNA methylation. *Developmental Biology*, 384(1):141–153, December 2013.

[151] Jamie A Hackett, Roopsha Sengupta, Jan J Zylicz, Kazuhiro Murakami, Caroline Lee, Thomas A Down, and M Azim Surani. Germline DNA Demethylation Dynamics and Imprint Erasure Through 5-Hydroxymethylcytosine. *Science*, 339(6118):448–452, 2013.

[152] Suhua Feng, Shawn J. Cokus, Xiaoyu Zhang, Pao-Yang Chen, Magnolia Bostick, Mary G. Goll, Jonathan Hetzel, Jayati Jain, Steven H. Strauss, Marnie E. Halpern, Chinweike Ukomadu, Kirsten C. Sadler, Sriharsa Pradhan, Matteo Pellegrini, and Steven E. Jacobsen. Conservation and divergence of methylation patterning in plants and animals. *Proceedings of the National Academy of Sciences*, 107(19):8689–8694, 2010.

[153] Albert Jeltsch. Phylogeny of methylomes. *Science*, 328(5980):837–838, 2010.

[154] G Panganiban, S M Irvine, C Lowe, H Roehl, L S Corley, B Sherbon, J K Grenier, J F Fallon, J Kimble, M Walker, G A Wray, B J Swalla, M Q Martindale, and S B Carroll. The origin and evolution of animal appendages. *Proceedings of the National Academy of Sciences of the United States of America*, 94(10):5162–5166, 1997.

[155] Evangelia Stamataki and Anastasios Pavlopoulos. Non-insect crustacean models in developmental genetics including an encomium to Parhyale hawaiensis. *Current Opinion in Genetics & Development*, 39:149–156, August 2016.

[156] Karyn N Johnson, Marielle C W van Hulten, and Andrew C Barnes. “Vaccination” of shrimp against viral pathogens: Phenomenology and underlying mechanisms. *Vaccine*, 26(38):4885–4892, September 2008.

[157] Yanan Lu, Junjun Liu, Liji Jin, Xiaoyu Li, YuHong Zhen, Hongyu Xue, Jiansong You, and Yongping Xu. Passive protection of shrimp against white spot syndrome virus (WSSV) using specific antibody from egg yolk of chickens immunized with inactivated virus or a WSSV-DNA vaccine. *Fish and Shellfish Immunology*, 25(5):604–610, November 2008.

[158] S Rajesh Kumar, V P Ishaq Ahamed, M Sarathi, A Nazeer Basha, and A S Sahul Hameed. Immunological responses of Penaeus monodon to DNA vaccine and its efficacy to protect shrimp against white spot syndrome virus (WSSV). *Fish and Shellfish Immunology*, 24(4):467–478, April 2008.

[159] Andrew F Rowley and Edward C Pope. Vaccines and crustacean aquaculture—A mechanistic exploration. *Aquaculture*, 334-337(C):1–11, March 2012.

[160] J T Simpson, K Wong, S D Jackman, J E Schein, S J M Jones, and I Birol. ABySS: A parallel assembler for short read sequence data. *Genome Research*, 19(6):1117–1123, June 2009.

[161] M Boetzer, C V Henkel, H J Jansen, D Butler, and W Pirovano. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, 27(4):578–579, February 2011.

[162] Guillaume Marçais and Carl Kingsford. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770, March 2011.

[163] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet*, 17(1):10–12, August 2011.

[164] V. Douris, M. J. Telford, and M. Averof. Evidence for multiple independent origins of trans-splicing in metazoa. *Molecular Biology and Evolution*, 27(3):684–693, nov 2009.

[165] Brian J Haas, Steven L Salzberg, Wei Zhu, Mihaela Pertea, Jonathan E Allen, Joshua Orvis, Owen White, C Robin Buell, and Jennifer R Wortman. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome biology*, 9(1):R7, 2008.

[166] M Stanke and S Waack. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, 19(Suppl 2):ii215–ii225, October 2003.

[167] Thomas D Wu and Colin K Watanabe. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21(9):1859–1875, May 2005.

[168] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):516–520, May 2010.

[169] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, January 2013.

[170] Guy St C Slater and Ewan Birney. Automated generation of heuristics for biological sequence comparison. *BMC bioinformatics*, 6:31, 2005.

[171] A V Lukashin and M Borodovsky. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Research*, 26(4):1107–1115, 1998.

[172] A F A Smit, R Hubley, and P Green. *RepeatMasker Open-4.0.*, 2013.

[173] Matthew Kearse, Richard Moir, Amy Wilson, Steven Stones-Havas, Matthew Cheung, Shane Sturrock, Simon Buxton, Alex Cooper, Sidney Markowitz, Chris Duran, Tobias Thierer, Bruce Ashton, Peter Meintjes, and Alexei Drummond. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12):1647–1649, June 2012.

[174] Robert C Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, 2004.

[175] A Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 2014.

[176] Felix Krueger and Simon R Andrews. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, 27(11):1571–1572, June 2011.

[177] Pin-Hsiang Chou, Hao-Shuo Chang, I Tung Chen, Han-You Lin, Yi-Min Chen, Huey-Lang Yang, and K C Han-Ching Wang. The putative invertebrate adaptive immune protein Litopenaeus vannamei Dscam (LvDscam) is the first reported Dscam to lack a transmembrane domain and cytoplasmic tail. *Developmental & Comparative Immunology*, 33(12):1258–1267, December 2009.

[178] E P Nawrocki and S R Eddy. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22):2933–2935, October 2013.

[179] Sam Griffiths-Jones, Harpreet Kaur Saini, Stijn van Dongen, and Anton J Enright. miRBase: tools for microRNA genomics. *Nucleic Acids Research*, 36(Database issue):D154–8, January 2008.