THE UNIVERSITY *of* EDINBURGH

Unearthing the genome of the earthworm *Lumbricus rubellus*

Benjamin Elsworth

University of Edinburgh

Doctor of Philosophy

2012

Contents

List of Figures

List of Tables

Declaration

I declare that this thesis has been composed by myself and, except where otherwise stated, is entirely my own work.

Benjamin Elsworth

Abstract

The earthworm has long been of interest to biologists, most notably Charles Darwin, who was the first to reveal their true role as eco-engineers of the soil. However, to fully understand an animal one needs to combine observational data with the fundamental building blocks of life, DNA. For many years, sequencing a genome was an incredibly costly and time-consuming process. Recent advances in sequencing technology have led to high quality, high throughput data being available at low cost. Although this provides large amounts of sequence data, the bioinformatics knowledge required to assemble and annotate these new data are still in their infancy. This bottleneck is slowly opening up, and with it come the first glimpses into the new and exciting biology of many new species.

This thesis provides the first high quality draft genome assembly and annotation of an earthworm, *Lumbricus rubellus*. The assembly process and resulting data highlight the complexity of assembling a eukaryotic genome using short read data. To improve assembly, a novel approach was created utilising transcripts to scaffold the genome (https://github.com/elswob/SCUBAT). The annotation of the assembly provides the draft of the complete proteome, which is also supported by the first RNA-Seq generated transcriptome. These annotations have enabled detailed analysis of the protein coding genes including comparative analysis with two other annelids (a leech and a polychaete worm) and a symbiont (*Verminephrobacter*). This analysis identified four key areas which appear to be either highly enhanced or unique to *L. rubellus*. Three of these may be related to the unique environment from which the sequenced worms originated and add to the mounting evidence for the use of earthworms as bioindicators of soil quality.

All data is stored in relational databases and available to search and browse via a website at www.earthworms.org. It is hoped that this genome will provide a springboard for many future investigations into the earthworm and continue research into this wonderful animal.

Acknowledgements

"It may be doubted whether there are many other animals which have played so important a part in the history of the world," Darwin (1881).

I would like to thank my supervisor Mark Blaxter for giving me the opportunity to take on such an interesting and challenging project, and for his continuing support and encouragement. Thanks also to the members of the GenePool and Blaxter lab for generating masses of data and helping me understand and use it in a manageable and meaningful way.

Many thanks to all members of the earthworm consortium, in particular the Morgan and Kille groups, who have been fundamental in the development and success of the project.

Most important thanks go to my family, in particular the newest members who make everything worthwhile.

Chapter 1

Introduction

1.1 The project

Earthworms have long held a special place in the history of biology, from their first detailed studies during Darwin's life-long fascination [33] [35] to the present day where they are of interest more than ever (Figure 1.1). Prior to Darwin they were seen as a soil pest, until he revealed them to be the eco-engineers of the soil. It would seem fitting with the recent 200th anniversary of Darwin's birth and 150th anniversary of his most famous book [34] in 2009 that his favourite creature were to become a genetic model organism for environmental soil science. By constructing the first draft of an earthworm genome, it is hoped that research continues to flourish on this remarkable animal and many more of its wonderful secrets are revealed. Here I present the stages of assembling and annotating a genome: from the *de novo* assembly of a 420 Mb genome from over 400 million reads and 20 billion bases to the identification of the features within and the first glimpses of some exciting new biology.

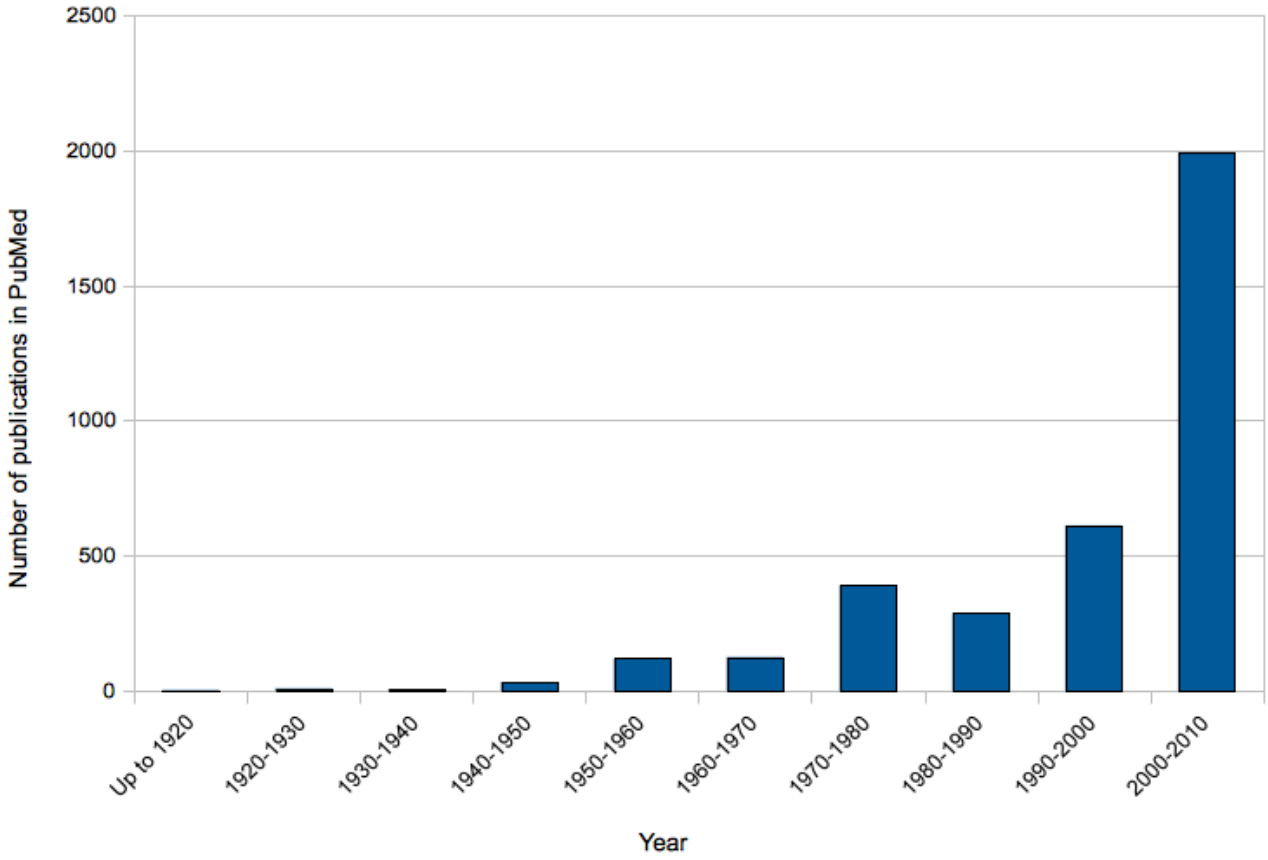

Figure 1.1: Number of earthworm publications in PubMed

1.2 Annelids

The Annelids (ringed worms) are a bilateral invertebrate phylum which, based on recent molecular phylogenetic analyses [153], form a monophyletic clade with respect to the sipunculans within the Lophotrochozoa (Figure 1.2). They are characterised by a cylindrical segmented body and a true coelom (fluid filled cavity), with most having a pair of coelomic cavities in each segment.

The phylum contains over 15,000 described species; representatives of which have recently been used to investigate annelid evolution through a detailed multi-gene phylogenetic analysis [160]. Traditionally, the annelids were split into two distinct groups, Clitellata (earthworms and leeches) and polychaetes (bristle worms). However, this latest study posits a three clade system (Figure 1.3). The majority of species studied fall into two clades based on lifestyle, the Sedentaria (sedentary) and Errantia (more mobile and active), with a third clade which includes phyla previously excluded from the annelids such as Chaetopteridae, Myzostomida and Sipuncula.

1.3 Earthworms

All known species of earthworm fall into a monophyletic clade in the class Clitellata known as the Oligochaeta with terrestrial species forming the order Opisthophora. The regional distribution of the 10 recognised major families from this order can be seen in Table 1.1. Recently there has been much interest in cryptic speciation within what were thought to be single species, including within the lumbricids [85][122][6][78][66], so the phylogenetic relationships and numbers of species are expected to change in the near future.

Perhaps the estimated 3,000 described species [148] and high level of divergence are a reflection of earthworms' amazing abilities to adapt and thrive in almost any environment. Combined with a world-wide distribution it hardly seems surprising that earthworms were recently voted the most successful species of all time [99].

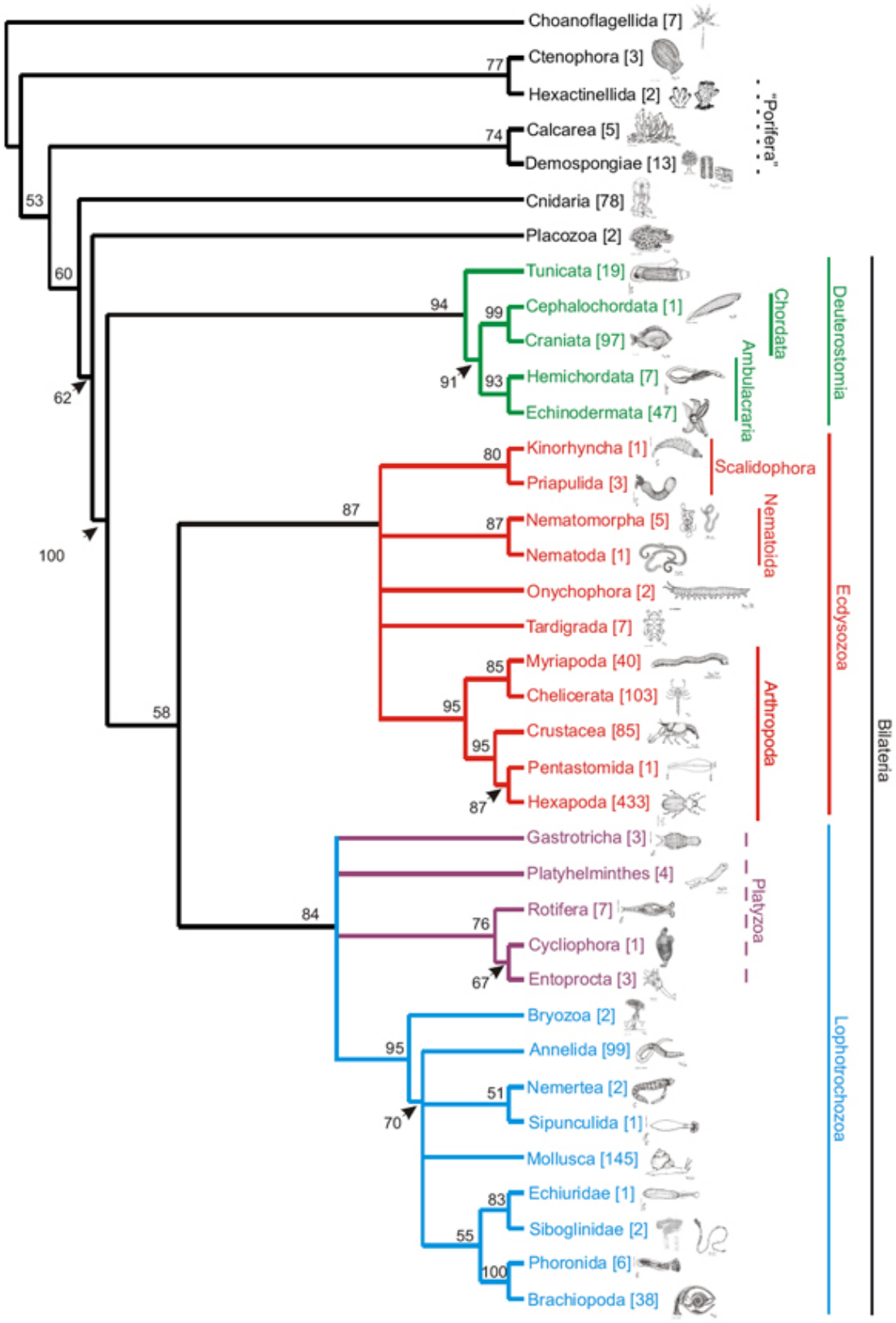

Figure 1.2: Phylogenetic tree of metazoan taxa (taken from [57])

Sequences of the 18S rRNA gene from 1,269 species were used to represent 36 phyla. Numbers in brackets represent the number of species in a subtree. The Annelida are located within an unresolved section of the Lophotrochozoa.

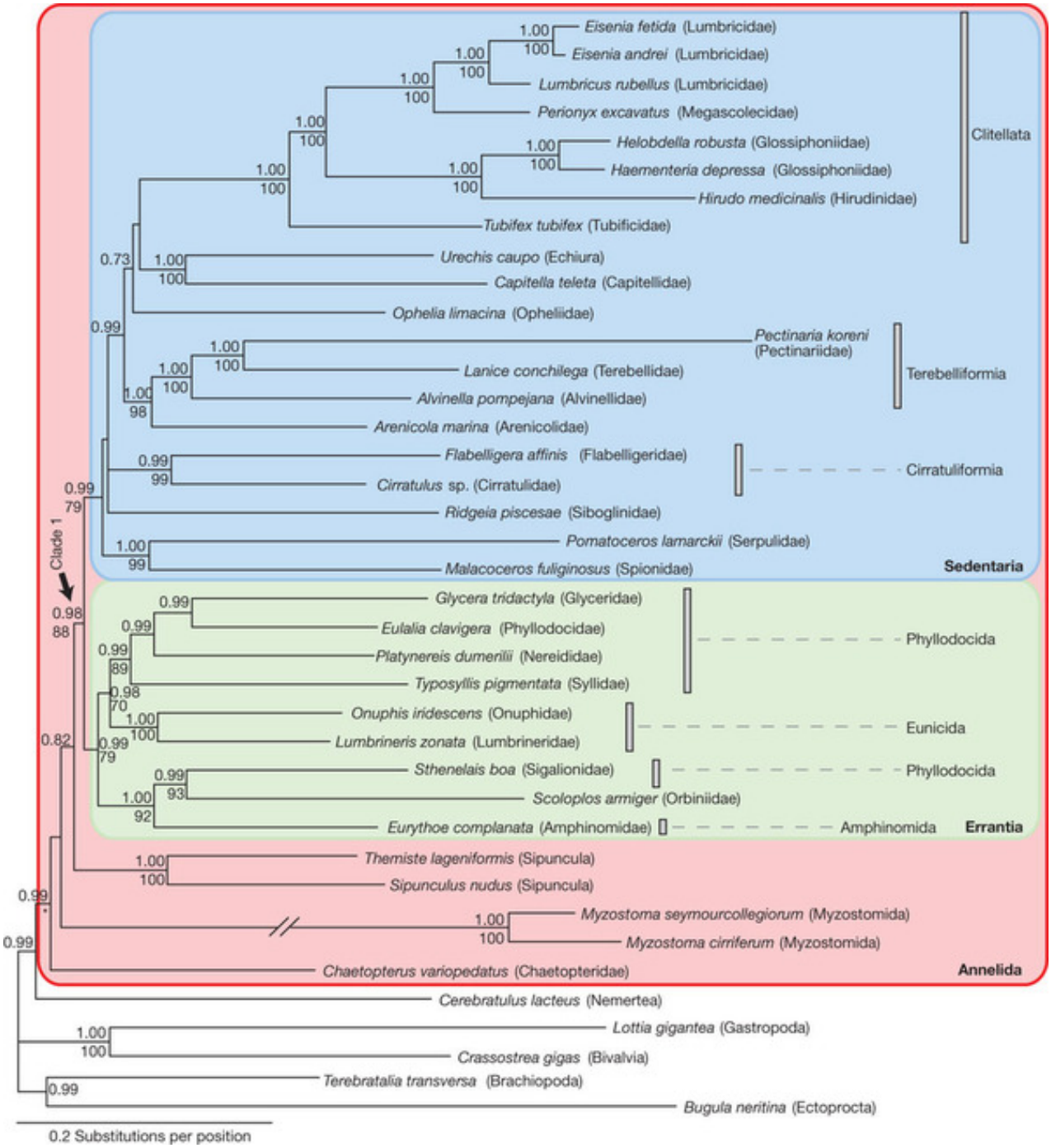

Figure 1.3: The phylogeny of the Annelida (taken from [160])

Majority rule consensus trees of the Bayesian inference analysis using the site-heterogeneous CAT model of the data set with 39 taxa and 47,953 amino acid positions. Only posterior probability (top of branch or alone) and bootstrap (bottom) values $\geq$0.70 or 70, respectively, are shown. The uppermost clade is a monophyletic group traditionally named the Clitellata which contains the earthworms, leeches and tubificids.

Table 1.1: The regional distributions of the 10 recognised major families of terrestrial earthworms

Family	Geographical region of origin
Ailoscolecidae	Europe
Eudrilidae	Africa
Glossoscolecidae	Central America, South America
Hormogastridae	Mediterranean
Komarekionidae	North America
Kynotidae	Madagascar
Lumbricidae	Europe, North America
Megascolecidae	Africa, Central America, North America, South America, Asia, Madagascar, Oceania
Microchaetidae	Africa
Ocnerodrilidae	Africa, Central America, South America, Asia, Madagascar

Taken from [161] .

1.3.1 Structure, ecology and physiology

Earthworms have a closed vascular system with dorsal and ventral trunks and a ventral nerve cord. The nerve cord has an anterior enlargement (brain) which controls the muscles and connects to various sense organs. They are externally segmented (with corresponding internal segments) and have a digestive tract that consists of an anterior-posterior tube with excretion occurring through the anus or nephridia [47].

Despite being ubiquitous, they share key physiological traits. Respiration is mainly cuticular and they can survive in water if the level of dissolved oxygen is sufficiently high. Their body temperature is dependent on the external environment which therefore creates a positive correlation between respiration rate and temperature. They feed on organic matter such as plants, microfauna, bacteria, decaying animal matter and fungi [47]. Earthworms have been shown to be sensitive to light but rely more on vibration and touch. They are also responsive to acidity and humidity.

All earthworms are hermaphroditic and may reproduce biparentally or uniparentally depending on the species. After fertilisation cocoons are produced, each containing one or two worms. Young emerge as small but fully formed earthworms except for a lack of sexual structures.

Species generally range in size from a few millimetres to 2 metres, from 10 mg to almost 1 kg and can be up to 40 mm in diameter [46]. One of the largest is the giant Gippsland earthworm *Megascolides australis* with an average size of 750 x 20 mm and weight up to 381 g. Populations range from only a few individuals per square metre to more than 1000 [47], depending on multiple factors including soil type, pH, rainfall and temperature. The largest populations are often lumbricids as they appear to be best at surviving adverse soil and litter conditions. Estimates of the amount of soil moved and digested by earthworms vary widely from 2 to 250 tons per hectare per annum. However, no doubt can be cast on the importance earthworms have on the health of the soil in terms of nutrient cycling, drainage, aeration and many other factors.

1.3.2 Symbionts

The only well documented symbiont of earthworms are bacteria of the genus *Verminephrobacter*. They are found in the nephridia: long coiled tube-like organs which are found in pairs in each segment and which exit the body wall via an exterior pore. The nephridia play an important role in the excretion of metabolic waste similar in function to mammalian kidneys (Figure 1.4). The *Verminephrobacter* are confined to the second loop (ampulla) where they form dense populations lining the lumen wall [100]. They are most closely related to the betaproteobacterial genus *Acidovorax* [141] and form a monophyletic clade within the genus.

Lund *et al* [100] detected evidence of the genus in 19 out of 23 investigated earthworm species from the Lumbricidae. It is proposed that the *Vermineprobacter* and earthworms have co-diversified based on the vertical transmission of the symbiont and their species specificity. Although prevalent within earthworms, the only proposed function of the *Verminephrobacter* is to enhance nitrogen retention [101], a function linked to their location within the earthworm.

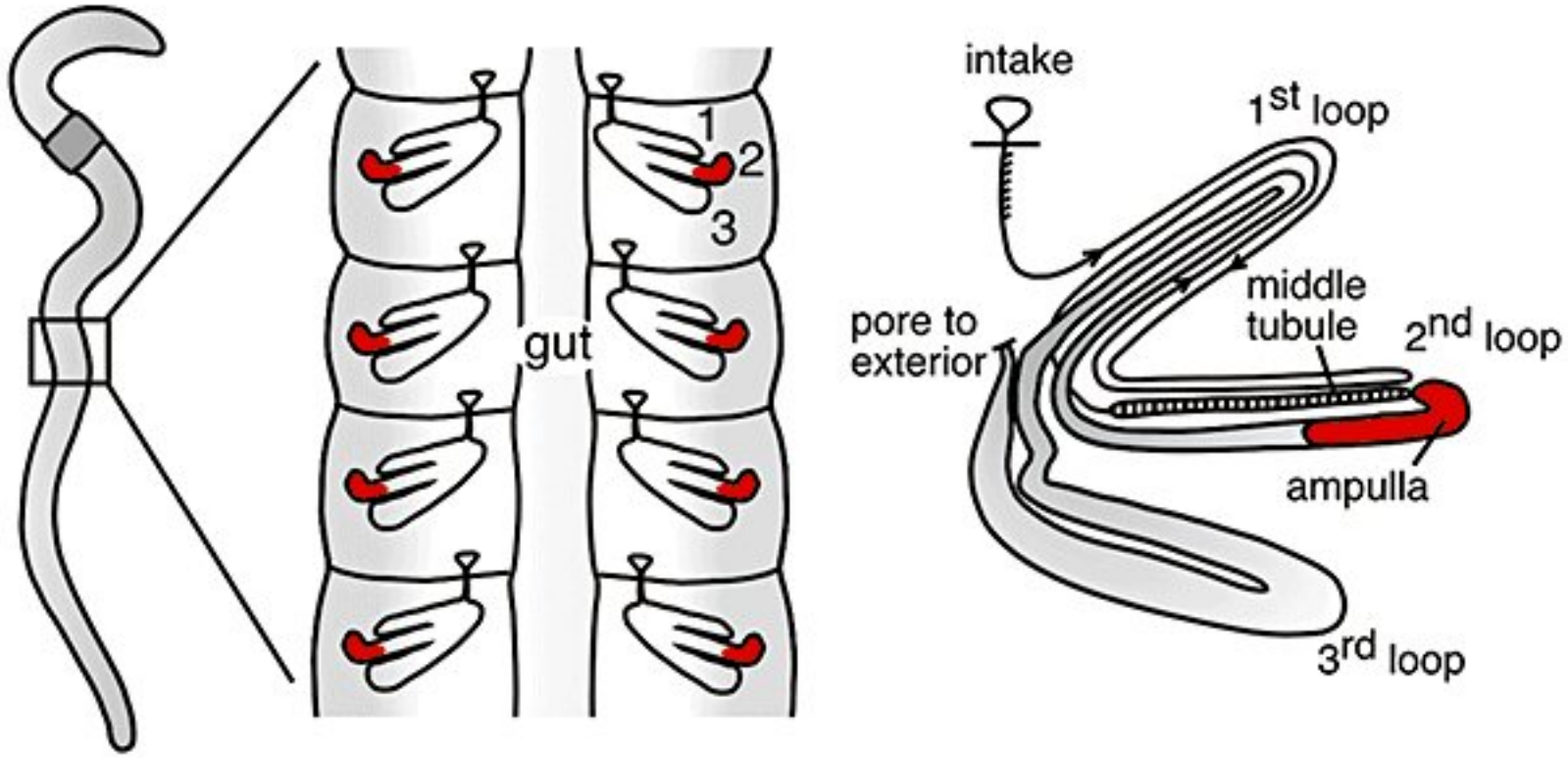

Figure 1.4: Schematic of the nephridia in an earthworm. Taken from [100]

Symbionts are restricted to the nephridia which are highlighted in red.

1.3.3 Research areas

A search for 'earthworm' in PubMed retrieves 3781 separate texts (June 2012). These cover wide ranging areas of science, such as:

Vermicomposting

Vermicomposting is an area of earthworm biology that is growing in popularity not only for garden compost but also for recycling of a range of waste substances throughout the world from industries such as textile [56], olive oil [110], pineapple [104] and sugar [139]. Understanding the key proteins involved in this process may help to continue research in this excellent field.

Organic farming

Comparisons of farming techniques identified a higher biomass and abundance of earthworms in organic farming plots compared to conventional farming [102][69][77]. The increase in number is likely the result of an increase in the use of farmyard manure in organic farming, and may reflect an enhanced soil fertility and higher biodiversity. Sustainable and environmentally friendly food production is a hugely important issue, and undoubtedly the role earthworms play in this will be crucial.

Origin of nervous system centralisation in the bilateria

Although dorsal in chordates and ventral in other bilateria, a central nervous system appears to be ancestral to all bilaterian animals [108][7]. Evidence from developmental studies and patterning genes have identified similar expression patterns of homeobox genes in the CNS of annelids and chordate species [39]. Our understanding of these and other similar evolutionary events can only be improved by sequencing more representatives from the animal kingdom. Therefore, earthworm genome will be a major addition to this investigation.

Soil pollutants

Much work has focused on soil pollutants; the ways in which earthworms have evolved to adapt to toxic components of soils, and the direct and indirect positive effects they exert on soil quality and fertility especially in areas of industrial or mining disturbance. Examples include the detoxification of heavy metals such as cadmium [162][154][123] and the response to other xenobiotics [126][20][149]. The complete genome will be used to finalise many of these investigations which until now have been relying on partial transcript data.

Regeneration

The ability to regenerate lost body parts varies dramatically across the animal kingdom but is thought to be ancestral to the annelids [14] with posterior and anterior regeneration seemingly widespread and well studied [47] [26]. Recent work has increased interest in this area as a new protein associated with CNS repair was identified in the leech *Hirudo medicinalis* [166]. Perhaps the new wave of genomic data from HTS will help unravel the true ancestry of this fascinating biological feature.

Medicine

Traditional Asian medicine is a growing industry that has long used earthworms to treat various ailments. The validity of the treatments are perhaps a little exaggerated. Earthworms express proteins with commercial promise, such as antibacterials [173] and distinct protease and fibrinolytic activities for modulation of blood clotting [169] [164]. Perhaps to add validity to the medicinal claims there is a large amount of interest in the molecular biology of earthworms and it is anticipated that the genome will be of great interest to these groups.

It is hoped that the completion and release of the earthworm genome will produce an increase in research in these and many other new areas, similar to the way genome sequence has underpinned research on other model organisms.

1.4 *Lumbricus rubellus*

L. rubellus, commonly named the red worm due to its colouring, is used as a model species by researchers investigating many aspects of biology, especially the effects of pollutants and toxins on soil. It is a common earthworm found in many temperate ecosystems, and ranges in length from 50-150 millimetres. This earthworm was selected for genome sequencing based on previous studies in which the Blaxter lab had collaborated. The EcoWorm consortium (http://xyala.cap.ed.ac.uk/Lumbribase/ecoworm/index.shtml) was established in 2002 with the aim of using *L. rubellus* alongside the model nematode *Caenorhabditis elegans* in a wide ranging program of investigations into the responses of 'soil organisms' to heavy metal and organic pollutants. The underpinning data for this project were expressed sequence tags (ESTs), short single read sequences obtained from cDNA that provide cost effective information on mRNA expression. Over the course of 6 years, over 20,000 ESTs were generated from nine different libraries and submitted to GenBank [15], culminating in the publication of a partial transcriptome in 2008 [126]. The raw sequences were processed and analysed using PartiGene [128] creating 8,129 EST clusters (UniGenes) each of which represents a putative gene. Therefore, approximately 40% of the estimated total gene set of 20,000 genes was identified. These data are stored in a web accessible database called Lumbribase [126].

This set of EST clusters was used to create a custom cDNA microarray which in turn was used to generate a set of dose-response transcript profiles for three xenobiotics: cadmium (an inorganic metal and recognised carcinogen), fluoranthene (an organic hydrocarbon that represents a class of persistent unnatural pollutants) and atrazine (a widely used herbicide associated with numerous health issues) [126]. Although the majority of clusters with statistically significant relationships between xenobiotic exposure and transcript levels were not informatively annotated, many were.

The most significant expression response to cadmium exposure was over expression of transcripts encoding the small metal-binding protein metallothionein, as well as a reduction in expression of cytochrome c oxidase and NADH-ubiquinone oxidoreductase II.

These were expected from previous work. However, novel insights into regulation of the process of DNA repair by cadmium were proposed as genes linked to double stranded DNA breaks and excision repair were found to be down-regulated. Fluoranthene exposure produced responses previously observed in mammals. However, the responses of cyctochrome P450 enzymes suggested a new mechanism for biotransformation of organic xenobiotics may be at work in *L. rubellus*. Finally, the most significant response to atrazine was from a group of genes associated with protein synthesis and catabolism suggesting a degradation and re-synthesis of proteins.

With the advent of high-throughput sequencing (HTS), the Ecoworm consortium decided to move to whole genome sequencing, something previously only available for major research programmes with significant funding and resources. Therefore, in 2008 it was decided that an attempt at the first earthworm genome should be made. With this new resource, it was hoped that key genes and pathways identified in the EST analysis as responsive to a range of toxicants, especially those with no annotation, could be mapped to the genome and regions upstream could be examined for conserved binding sites for known regulators. Detailed work on the pathways involved in these systems could also be investigated if all genes were available.

1.5 The genome

A genome is the complete set of genetic material contained within a set of haploid chromosomes. The genome of *L. rubellus* is estimated to be 420 megabases (Mb) and distributed over 18 chromosome pairs [168]. Table 1.2 puts the *L. rubellus* genome into context by comparing it to model organisms, as well as the two other annelid genomes which have been sequenced and assembled by the DOE Joint Genome Institute (http://www.jgi.doe.gov): the polychaete worm *Capitella telata* (http://genome.jgi-psf.org/Capca1/Capca1.home.html) and the leech *Helobella robusta* (http://genome.jgi-psf.org/Helro1/Helro1.home.html). Figure 1.3 shows the

relationships between the three annelids, with *H. robusta* more closely related to *L. rubellus* than *C. telata*, but all three within the Sedentaria clade. Both of the JGI annelids were sequenced using traditional Sanger sequencing over many years at great cost and are as yet unpublished but freely available. Section 1.6 discusses the costs associated with genome sequencing projects and Table 1.3 highlights the huge reduction in both price and time per base, while Figure 1.5 shows the cost per base over the last decade. The dramatic drop in 2008 coincides with HTS and the start of this project.

Table 1.2: Genome size comparisons

Species	Common Name	Genome size (Mb)	No. chromosomes
Escherichia coli [175]	*E. coli*	4.6	1
Caenorhabditis elegans [1]	Nematode worm	100	6
Drosophila melanogaster [3]	Fruit fly	180	4
Capitella telata [145]	Polychaete worm	324	10
Helobdella robusta [145]	Leech	228	18
Lumbricus rubellus [168]	Red earthworm	420	18
Homo sapiens [171]	Human	3,200	23

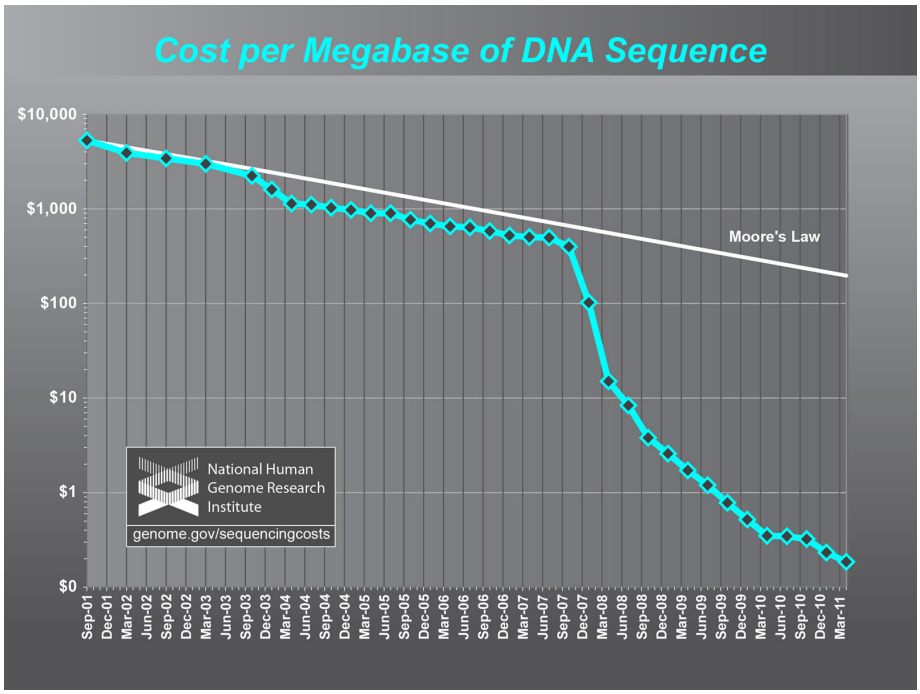

Figure 1.5: Cost per megabase of DNA Sequence (taken from http://www.genome.gov/sequencingcosts)

The alternative to sequencing the genome of *L. rubellus* was to focus on the coding regions and sequence the transcriptome. However, to maximise the chances of sequencing all genes, numerous individuals at various life stages and under varying conditions would need to be sequenced and then assembled; a labour intensive and complex process. In addition the cost and time associated with this approach were predicted to be comparable to genomic sequencing costs at the time, and with the understanding that sequencing would get better and cheaper it was decided to sequence the entire genome. The hoped-for benefits of this were many:

1. All genes would be sequenced, including all alternate transcripts. Additionally, a genomic gene prediction provides extra evidence as to a gene's validity, in terms of presence of introns, UTRs, coverage, GC and so on. A transcript provides no proof of origin, as contamination in the sequenced data could produce misleading transcripts.

2. Non-coding regions would be identified, e.g. ncRNAs, transposons, promoter regions.

3. This would be the first oligochaete genome and a huge addition to the genomic world.

4. Proof of principle that a small lab can sequence and assemble a large genome *de novo*.

From the outset, the sequencing data available and estimates of repeat content meant the project was never going to produce a fully scaffolded genome (see Section 2.2). However, it was assumed that even a fragmented assembly would still contain all the objects of interest, albeit with some fragmentation. The aim therefore was to produce a 'high quality draft' genome, as defined by Chain et al [24], i.e. with 90% of the genome represented with efforts having been made to remove contaminating sequences, thus making the genome appropriate for general assessment of gene content.

Towards the end of this project, Illumina RNA-Seq transcriptome data became available as part of a separate project and this was used to help both assemble and annotate the genome.

1.6 Genome sequencing

Traditionally, sequencing technology has been based on three stages: initial DNA fragmentation, fragment amplification, and then sequencing of the amplified fragments. With a few exceptions (3rd generation sequencing technologies), this process remains unchanged, as to sequence a genome the extracted DNA needs to be fragmented (as whole chromosome sequencing is not yet available) and these fragments need to be amplified to produce enough DNA for the sequencing reactions. The recent developments associated with HTS have changed the second and third stages, making them much cheaper and quicker, and ultimately increasing the throughput of data.

Sanger sequencing

Sanger sequencing has been the major sequencing technology since the late 1970's, and was used exclusively for the most famous of sequencing projects, the human genome [171]. However it was, and still is, limited by time and cost. For example, the human genome project used over 27 million Sanger reads totalling almost 15 Gb, costing around $3 billion and taking 13 years to complete [171]. More recently, the *Schistosoma japonica* genome [187] which is around 400 Mb was assembled from over 3.74 million Sanger reads, and with an average cost of $1 a read.

One of the defining and rate limiting steps during Sanger sequencing is in vivo cloning and amplification of DNA. This step is performed in vitro in the HTS technologies.

Roche 454 pyrosequencing (www.454.com)

This technology uses emulsion PCR whereby droplets act as individual amplification reactors, each producing around one thousand clonal copies per isolated bead. Hundreds of thousands of unique amplified fragments can be analysed in parallel. Each bead is loaded into an individual well on a picotiter plate. Sequence determination occurs by pyrosequencing reactions, (where nucleotide addition is detected by measuring the release of inorganic pyrophospate) through a coupled chemiluminescent reaction. The intensity of the light emitted reflects the number of bases added. This can lead to errors in base calling in regions of homopolymeric sequence [114].

Illumina Solexa sequencing (www.illumina.com)

Here, single-stranded DNA fragments are attached to a solid surface and amplification occurs via solid-phase bridge amplification. This creates local clusters each containing approximately 1000 clonal copies of a unique sequence. Again a sequencing-by-synthesis approach is used where chain-terminating nucleotides with cleavable fluorescent tags are used to add each base sequentially. Each cycle of addition is scanned using a laser, and the incorporated base determined based on the colour detected. Illumina sequencing has no particular issues with homopolymeric regions, but it is limited to producing shorter reads.

The current price (Dec 2011) for Illumina data is 7 pence per Mb compared to almost 30 times that amount 2 years earlier. This means that at these prices, the same level of sequencing required for this project would cost just £3000. This doubling of Moore's law is expected to continue as is the increase in read quality and length. Table 1.3 gives a brief comparison of these three technologies.

Table 1.3: Comparison of sequencing technologies available in the GenePool (http://genepool.bio.ed.ac.uk/) in mid 2009.

Sequencing technology	Max read length (bases)	Read data per run (Mb)	Time per run (hours)	Cost per Mb raw data (£)	Error rate per base (%)*
ABI3730 (Sanger)	Up to 900	0.5	4	5,000	0.1-1
Roche Titanium (454)	Up to 500	400	12	20	1
Illumina GAII (Solexa)	Up to 150	5000	72-144	2	∼0.1

* error rates taken from [59]

1.6.1 History of genome assembly

The very first genome of a DNA-based organism to be sequenced was the bacteriophage phi X174. This was achieved by a team led by the founder of Sanger sequencing, Fred Sanger, in 1977 [138]. Although this genome was just 5,386 base pairs (bp), this was a major achievement at the time and it was another 18 years before the first bacterial genome, that of *Haemophilus influenza*, was completed [53]. The first eukaryotic genome, the yeast *Saccharomyces cerevisiae*, was completed in 1996 [61] and just two years later the first animal genome was sequenced, that of the nematode *Caenorhabditis elegans* at 100 Mb. Three years later the first human genome was published [171]. Since then numerous genomes have been completed and released, with ever more in the pipeline (see Table 3.1).

The very first genome assemblies used relatively low numbers of long reads enabling a greedy all-against-all overlap-layout-consensus approach whereby those sequences that were most similar were joined together under the assumption that shared sequence implied a shared origin. This approach fails when applied to complex genomes that are large and contain repetitive sequence data longer than the individual reads. Whole Genome Assembly (WGA) begins with an all-against-all read comparison in the overlap step (1.5 x 10^{15} for the 27 million reads in the human genome project), but then reads are arranged according to their pattern of overlap in the layout phase. Multiple alignments are then produced from the overlapping regions and a consensus sequence is derived. Paired read data and other methods such as end sequences from bacterial artificial chromosomes (BACs) can then be used to scaffold and finish a genome. This remained the method *de rigueur* for many years as use of assemblers such as Phrap (`www.phrap.org`), CAP3 [72], ARACHNE [12], Celera [119] and Phusion [117] became widespread. HTS assembly methods still follow the same shotgun sequencing approach. However, the huge increase in data has required new assembly algorithms which are discussed in Section 3.1.1.

1.6.2 Coverage

The ultimate aim of *de novo* genome sequencing is to generate a representation of each base for a set of haploid chromosomes. Theoretically this could be achieved by producing sequence data that covers each base once, but this would require a single sequencing run producing genome or chromosome-length stretches of data with zero error. Current sequencing technologies have the following attributes:

Short read lengths

Table 1.3 lists the features of the sequencing technologies available as of September 2009. 454 and Illumina reads are much shorter than Sanger reads.

Non-uniform read distribution

Each sequencing platform has its own unique pattern of biased sequence coverage [68]. However, as the technologies progress these issues are being addressed. Additionally there are other factors that continue to affect the distribution of sequencing data. One issue is PCR bias during the amplification stage of library preparation [4]. Another, the stochastic distribution of the sequenced reads, causes peaks and troughs of coverage. Finally there are DNA sequence attributes which can cause a loss of data, for example, hairpin regions which snap-back and restrict the sequencing of these sections.

Sequencing errors

Table 1.3 lists the error rates for the sequencing technologies available as of September 2009. Although the error rates are all low and can be corrected by high coverage, they will produce many incorrect bases in the raw data.

To compensate for these issues multiple sequencing coverage is required, often denoted as X. This is the average number of reads representing each nucleotide in the assembled sequence and can be calculated using the number of reads N, the average read length L and the genome size G (1.1).

$$X = N \times \frac{L}{G} \tag{1.1}$$

Equation (1.1) can be used to calculate the number of reads required to produce a desired coverage, for example, achieving a coverage of 10X using 50 base reads for the *L.rubellus* genome would require 80 million reads (1.2).

$$10 = N \times \frac{50}{420000000} \qquad N = \frac{420000000 \times 10}{50} \qquad N = 80000000 \tag{1.2}$$

As sequencing costs continue to decrease the optimum sequencing depth is still being assessed. High coverage can complicate an assembly with too much information inducing unnecessary errors (this can be avoided with a strict filtering step). Low coverage may produce areas with very little or no coverage and a failure to assemble at those regions. Figure 1.6 is a screenshot of a display generated by Tablet [111] showing an example of read coverage and error rates from a typical short read assembly from 2009. The upper window displays a view across the entire contig whilst the main window focuses on a 150 bp region and highlights the read errors which are identified as white bases within reads. The figure highlights the variation in read coverage across the contig. The figure also demonstrates how frequent base discrepancies/errors require a high coverage to produce a reliable consensus assembly. These factors were taken into account by the consortium when deciding on a sequence coverage strategy for the *L. rubellus* project.

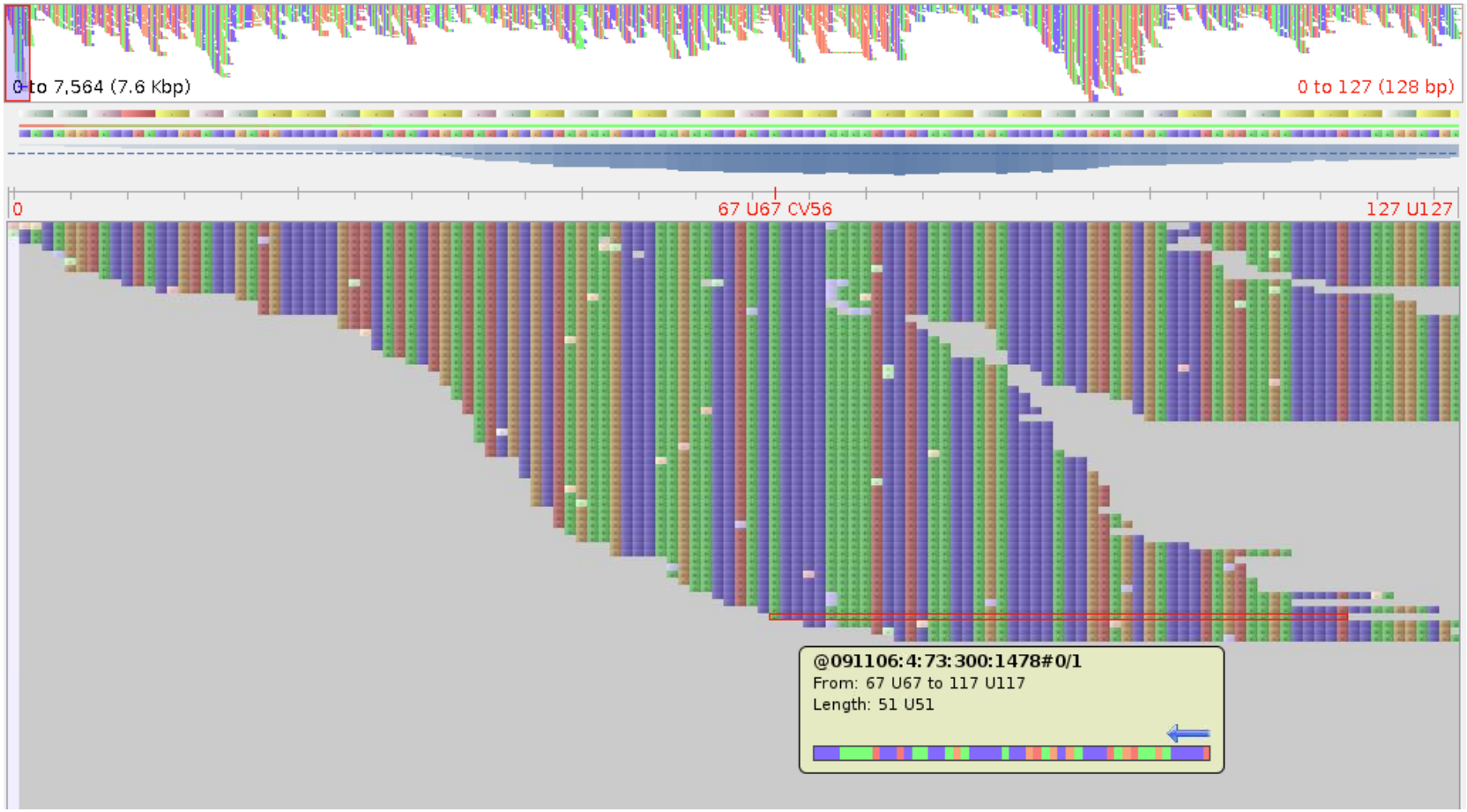

Figure 1.6: Tablet display of a typical short read assembly and mapped reads

Data is from a 7.6 Kbp contig from preliminary *L. rubellus* assembly. The top panel displays an overview of the aligned reads on the entire contig with the section in the main window highlighted. Immediately below this is the consensus sequence and coverage of the section in the main window. The main window itself displays detailed read information and demonstrates the ability to highlight base discrepancies by varying the contrast of low frequency bases. The bottom right inset displays information for an individual read, e.g. read name, position, length and direction.

Sanger based genome projects use relatively long reads (800-1000 bp) but at high expense (Table 1.3). For these reasons, sequencing coverage for these genome projects is often 5-10X, a compromise between cost and read length. Conversely, HTS reads are relatively short (up to 400 bp, more often much less) and have a much lower cost per base. Therefore, coverage for HTS projects is much higher. This can have an added benefit as a higher coverage value provides more scope for using this attribute during the assembly (see Chapter 3).

1.6.3 Repeat regions

Even with sufficient data, the main problem facing any genomic sequence assembly project, especially one using short read data from a complex eukaryotic genome, is that of repeat regions. In the absence of repeats, even a short read would have a unique place within a genome, making the assembly process simple. Repeats fall into two broad categories:

1. Simple.

   Forming the majority of the repetitive elements of many organisms, these repeats are often short simple sequences referred to as micro- and minisatellites. In addition, larger local sequence inversions and duplications can be frequent. These cause major problems as their size is often longer than the mean read length generated from HTS.

2. Complex.

   These are often derived from transposable elements (TEs) and form two major categories, retrotransposons (class I) and DNA transposons (class II). The former are the more prolific as they are able to "copy and paste" themselves within a genome, whereas the latter act via a "cut and paste" mechanism.

The proportion of each genome that is formed from repetitive elements varies dramatically. For example, the TE content of *D. melanogaster* is estimated at 20%, *H. sapiens* 45% and bread wheat 80% [54]. The genome of the blood fluke *Schistosoma mansoni*,

which is of a similar size to *L. rubellus*, has an estimated repeat content of 40% [17] and therefore a similar level might be expected in the earthworm.

During the assembly process, each time a repeat region occurs the assembly algorithm attempts to resolve it. If this is not possible the assembly will be broken at that point and another separate section of contiguous data (contig) is formed. Highly repetitive genomes will therefore be highly fragmented unless special measures are used to tackle this class of problems (see Figure 2.1). There are two options available:

i) Read length

   The longer the reads, the more likely they will be able to bridge the length of a repeat. Sanger read lengths are significantly longer than those of HTS, and therefore this is much more of a problem now for HTS projects.

ii) Distance information between reads.

   Most sequencing platforms offer some kind of paired end read data, i.e. generation of a pair of reads from either end of a single piece of DNA with an estimate of the distance between them. Currently these exist in two formats - standard short insert-size paired-end read (e.g. Illumina paired reads - see Figure 1.7) which are usually under 1000 bases, and the increasingly common long insert-size or mate pair libraries which are usually kilobases long. These large insert libraries work on the same principle, but the insert size between the reads can be significantly larger. Mate pair library preparation is more complicated but the power to scaffold contigs and overcome repeats is much greater. For example, in sequencing and assembling the panda genome, five different insert size libraries were used, 150 bp, 500 bp, 2 Kb, 5 Kb and 10 Kb [95]. Initial assemblies were performed using the two smallest insert size libraries and then scaffolded using the three larger libraries.

Identifying repeat regions is also important for the downstream annotation process as both a way to decrease the search space and minimise the chances of false positive annotations caused by transposable elements. Section 4.1.1 discusses this further.

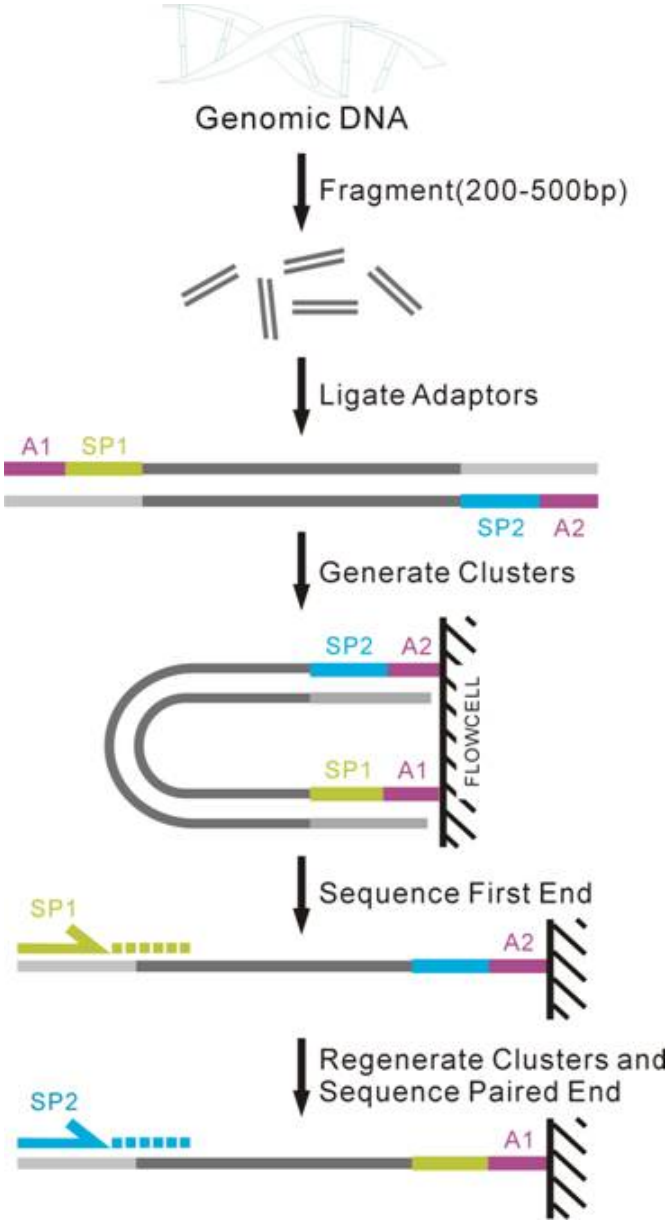

Figure 1.7: Generating Illumina paired-end data (taken from http://www.illumina.com/technology/paired_end_sequencing_assay.ilmn). Adapters containing attachment sequences (A1 and A2) and sequencing primer sites (SP1 and SP2) are ligated onto DNA fragments. The resulting library of single molecules is attached to a flow cell. Each end of every template is read sequentially.

1.7 Project aims

The aims of the project were threefold.

1. Assemble a high quality draft genome for *L. rubellus*.

2. Annotate the assembled genome identifying as many coding and noncoding objects of interest as possible.

3. Use the assembly and annotations to investigate the biology of *L. rubellus*.

Chapter 2

Data generation and QC

The selection of worm, preparation for sequencing, and sequencing itself were all performed by others. The names of these people are cited in the relevant sections. I received both the genomic and transcriptomic data in native raw format and therefore my work began at the data filtering stage described from Section 2.4.

2.1 The chosen worm

L. rubellus in the UK fall into two groups, based on mitochondrial sequence data, called A and B. The B clade is less genetically diverse (less than 2% divergence between haplotypes at cytochrome oxidase 2) than the A clade (up to 8% divergence) [85].

The three individual worms chosen for sequencing were from an abandoned lead mine at Cwmystwyth and selected for sequencing based on background information collated by the Kille and Morgan groups at the University of Cardiff (http://biosi.subsite.cf.ac.uk/biosi/kille-morgan/). For over three decades this location has been the site of much research into earthworm adaptation due to its unique environment and high concentration of heavy metals [116][112][113][106]. A series of individual *L. rubellus* were isolated from a specific location with a high nickel soil contamination, and these were genotyped for both cytochrome oxidase 2 and a series of random amplified fragment length polymorphisms (AFLPs) [6]. A and B group individuals were present at the site,

and there was evidence from discriminant analyses of the presence of hybrid individuals. Using discriminant analyses, 30 B mitochondrial haplotype individuals were compared, and an individual (S17) was identified that had minimal evidence in the AFLP data for shared alleles with A haplotype specimens, and thus most likely to be pure B. S17 (and two other genetically very similar individuals) were shipped to Edinburgh on dry ice. To ensure minimal levels of heterozygosity, only S17 was used for the genomic sequencing, the other two being kept for future sequencing requirements. High molecular weight DNA was prepared from S17 using a 'standard' Maniatis proteinase K/phenol-chloroform extraction by Mark Blaxter. This DNA was then quality tested and passed on to the sequencing team at the GenePool.

2.2 Aim - Data production

Sequencing technology in 2008 / 2009 lent itself to a strategy of generating bulk coverage using the shorter Illumina reads in both single and paired-end format and using a lower coverage of the longer Roche data to help contiguate the data. Therefore, for this project the aim was to produce 50X and 5X for each of the sequencing types respectively. Read lengths were around 50 bp and 200 bp for the Illumina and Roche reads respectively with an insert size of around 180 bp for the paired-end Illumina data. Using this combination of read types it was hoped that the issue of repeats would be reduced (Figure 2.1).

To assist with annotation, the transcriptome of *L. rubellus* was also desired. The existing EST data provided partial coverage. Fortunately, in 2011 Illumina RNAseq data were generated by Pete Kille and colleagues at the University of Cardiff, sequenced at the GenePool and made available for this project.

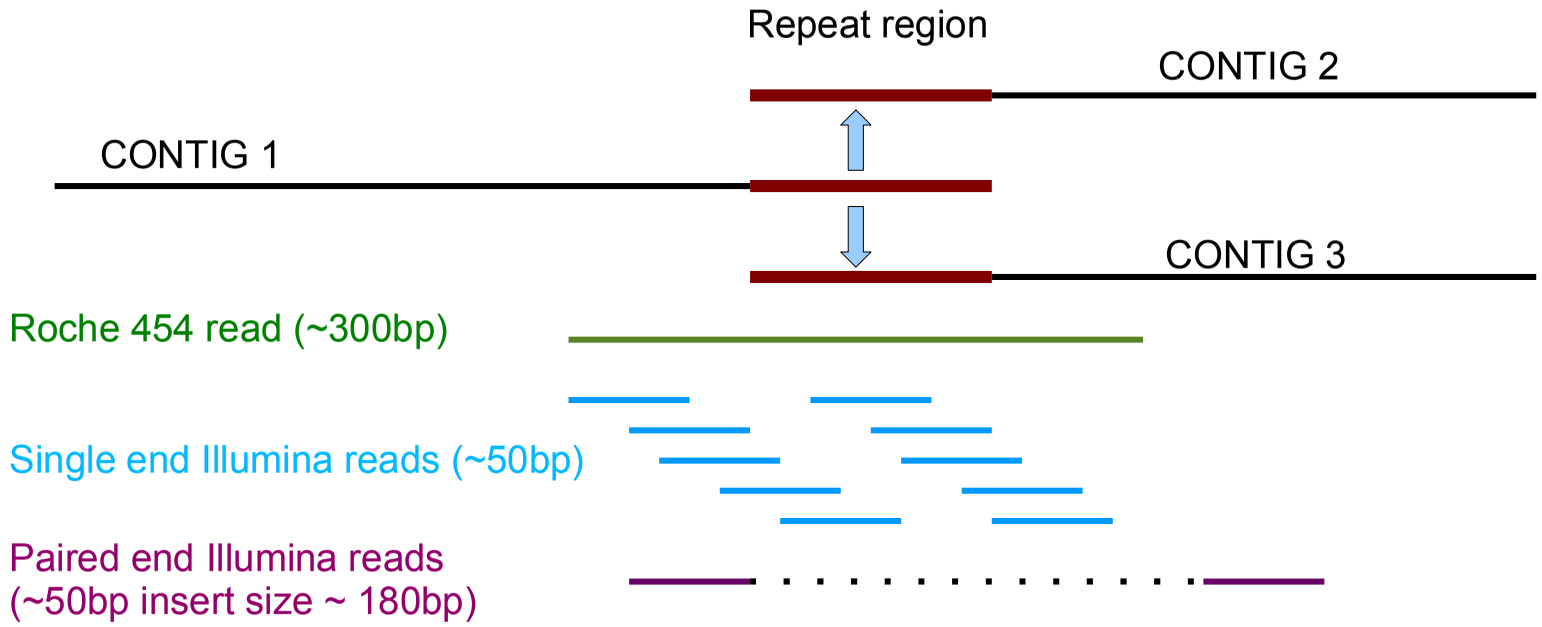

Figure 2.1: Reducing the problem of repeats with multiple read types.

Contig 1 shares sequence with both contig 2 and contig 3, a repeat region. To ascertain which contigs are linked two types of information can be used. A long read (Roche 454) can bridge the repeat region unique or paired end information (Illumina) can links the two correct contigs.

2.3 Raw data generation

2.3.1 Extraction of the DNA and RNA

Genome data

Extraction of genomic DNA from the earthworms was performed by Mark Blaxter, who supplied the following description:

The earthworm specimens were maintained on damp filter paper overnight to allow them to void their gut contents, and surface cleaned, to reduce the potential contamination from commensal or colonising microorganisms. They were then individually snap frozen in liquid nitrogen, and stored at -80^{o}C until extraction. DNA was extracted from the chosen *L. rubellus* specimen S17 by grinding the whole animal in a mortar and pestle under liquid nitrogen. The powdered specimen was then digested with proteinase K in 5 ml SDS buffer (TE [10 mM Tris-HCl, 1 mM EDTA] pH 8.0, with 150 mM NaCl and 1% SDS) at 65^{o}C for 8 hrs, and then placed at room temperature. The digest was extracted with TE-buffered phenol (at pH 8.0) twice, by slow inversion, with the phenol phase being removed from below the aqueous phase after centrifugation at 3000 rpm in a Sorvall benchtop centrifuge. The aqueous phase was then extracted twice by slow inversion with phenol-chloroform-isoamyl alcohol (20:20:1), and then twice with chloroform-isoamyl alcohol (20:1) as before. The final aqueous supernatant (4 ml) was carefully pipetted from the final chloroform extraction using a wide bore pipette, avoiding any of the tight protein phase at the aqueous-organic interface. The DNA was precipitated from this supernatant by underlaying the aqueous phase with 3 ml 100% ethanol (kept at -20^{o}C), and spooling out the high-molecular weight DNA from the ethanol-water interface using a sterilised glass loop. The spooled DNA was air dried briefly, and resuspended in 500 μl TE pH 8.0, supplemented with 1 μg/ml DNase-free RNase, at 4^{o}C overnight. The remainder of the nucleic acid

in the final supernatant was precipitated by slowly mixing the ethanol and water phases, and storing the mixture at -20^{o}C overnight. The precipitate was split between eight 1.5 ml tubes and pelleted by centrifugation in a microcentrifuge at top speed (13000 rpm), washed with 70% ethanol (1 ml per tube). After air drying, the nucleic acids were resuspended in a total of 400 μl (50 μl/tube), pooled and digested overnight with 1 μg/ml DNase-free RNase.

RNase treated DNA was precipitated from both high molecular weight and low molecular weight samples with ethanol, and after brief drying, both were resuspended overnight in TE. DNA was quantitated using a Nanodrop spectrophotometer, and integrity assessed by gel electrophoresis in 0.6% agarose gels. The high molecular weight sample was free of obvious protein or RNA contamination, largely comprised material too large to enter the 0.6% gel, and totalled 27 μg of DNA. This material was used for subsequent library production.

Transcriptome data

RNA was prepared by colleagues in Cardiff University, Pete Kille and Craig Anderson. Craig Anderson provided the following:

Adult Arsenic Exposure

An adult *L. rubellus* exposure was conducted in accordance to that described by Spurgeon et al [155]. Field-collected *Lumbricus rubellus* were purchased from Lasebo (Nijkerkerveen, Holland) and maintained in an uncontaminated culture of artificial soil consisting of a 1:1:1 mix of loam soil: peat: composted bark for two months. The test medium consisted of 1559 g dry, sieved (2 mm) kettering with 3% composted bark. Distilled water or a solution of sodium arsenate (Na_2HAsO_4-$7H_2O$) (Sigma Aldrich, Dorset, UK) was added to provide a moisture content of 60% (dry weight equivalent) and soil concentrations of 0, 3, 12, 36,

and 125 mg As kg-1 (deemed adequate within the literature [52][181][109][89]). Soils were left for 10 days to reach an initial speciation equilibrium. 48 hours prior to exposure, worms were collected from culture and kept under test conditions. Each concentration consisted of 5 replicates, each containing five adult worms that were collectively weighed before being added to the soil. Containers were covered to limit water loss and kept at in constant light for 28 days. Initially, 8 g (dry weight) of dried horse manure was contaminated with corresponding chemical concentrations and rewetted to 80% moisture content before addition across the soil surface in each container. When worms were assessed after 14 days, the remaining manure was removed and replaced by 8 g of fresh food. Following exposure, worms were retrieved from the soil, weighed, and visually inspected for phenotypic characteristics. All individuals were frozen and homogenised in liquid nitrogen using a steel mortar and pestle and stored at -80 ^{o}C. Cocoon production rates (cocoons per worm per day) were determined by sieving the soil at the end of the exposure, and comparing the number of cocoons collected with survival data to calculate cocoon production rates.

Total RNA Purification

Total RNA was purified from each replicate individually according to the Qiagen RNeasy kit protocol for purification of total RNA from animal tissue. 10 mg of tissue was homogenised using a needle and syringe. An on column DNA digest (using the Qiagen RNase-free DNase set) according to the instructions in appendix D, to ensure that there was no DNA contamination. RNA was eluted in 60 μl RNase free water. The samples were then analysed using an Agilent Bioanalyser and were frozen and kept at -20 ^{o}C. Replicates from each exposure concentration were pooled equally and sent to the GenePool for sequencing.

It is important to note that as the *L. rubellus* individuals used for RNA preparation and

transcriptome sequencing were not from the Cwmystwyth mine it was possible that they were a mix of A and B lineages. In addition, the worms did not have their gut faunas removed prior to library generation.

2.3.2 Sequence data

Illumina

Illumina data were generated between 2008 and 2011 using three versions of the first generation of Illumina instrument, the GA, GAII and GAIIx. Libraries were prepared according to manufacturers protocols. Data processing was performed using the Illumina pipeline designed for each machine respectively and the image processing software IPAR versions 1.3 and 1.8 (Tables 2.1, 2.2 and 2.4). Paired-end data were generated from a library with an estimated insert size of 180 bp. In 2011 two transcriptome data sets were produced consisting of 80 M and 130 M paired-end Illumina reads with a mean insert size of 130bp. For this project only the first data set was used due to time constraints (Table 2.4).

Roche

Two sets of Roche data were generated, the first in 2009 using the standard Flx machine, the second in 2010 using the Flx-Titanium upgrade. Both libraries were prepared as per the Roche 454 library preparation protocols for each machine respectively. Sequencing was then performed followed by signal-processing and base-calling using the Roche Shotgun signal-processing software, gsRunProcessor versions 2.0.00.20 and 2.3 (Tables 2.3 and 2.4).

2.3.3 Output format

Illumina

The Illumina instruments pass the raw data to the Illumina Genome Analyser Pipeline. Here the raw data is converted into FASTQ files. These consist of four lines per read, line 1 is the header, line 2 is the base calls, line 3 is a second header and line 4 is an ASCII character string representing the quality call for each base, e.g.

```
@090126:4:1:1134:851/2
ATCGGTCTGTAGTTGTCCTGAATGGATGACAGTAGACTGTTTGAATATTT
+090126:4:1:1134:851/2
hhhhhhhhhhhhhhhhhhhhhhhhhhhhEhVh?hh[hhhhhFhQQhhhhB
```

The first header in line 1 contains information on the origin of the read, each separated by a colon:

090126 = the unique instrument name

4 = flowcell lane

1 = the number within the flowcell lane

1134 = the x coordinate of the cluster within the tile

851 = the y coordinate of the cluster within the tile

/1 or /2 = the member of a pair

Up until version 1.3 of the Illumina Genome Analyzer Pipeline the FASTQ quality scores were calculated in the following way [28] where P_e is the estimated probability of a base calling error calculated by the pipeline:

$$QIllumina = -10 \times log10 \left(\frac{P_e}{1 - P_e} \right) \tag{2.1}$$

This is a variation of the method used to calculate PHRED scores (the quality metric for Sanger sequencing) and the two are easily interchanged. After version 1.3 Illumina PHRED scores were reported with an ASCII offset of 64, thus creating a second variant of the original FASTQ format. The quality scores represent the confidence in the base assignment at each position. For example, a base with an ASCII character that maps to a score of 13 equates to a p value of 0.05. As a rule of thumb a quality score of 30 indicates a 1 in 1000 probability of error. A quality score cut-off of 20 is often used which represents an error rate of 0.01 or 1 in 100.

Roche

The Roche machines produce binary Standard Flowgram Format (SFF) files which also contain quality scores. These can either be used directly by some software or quite easily converted to FASTQ or FASTA.

Table 2.1: Illumina genomic single-end sequencing data

Date	Mean length (bp)	Number (Millions)	Combined Size (Gb)	Coverage
01/12/08	51	7.3	0.4	0.9
08/12/08	41	7.3	0.3	0.7
11/12/08	41	7.2	0.3	0.7
15/12/08	41	7.7	0.3	0.8
18/12/08	51	6.2	0.3	0.8
19/01/09	51	6.7	0.3	0.8
09/02/09	37	7.5	0.3	0.7
12/02/09	37	4.4	0.2	0.4
27/02/09	41	7.3	0.3	0.7
03/04/09	51	8.5	0.4	1.0
16/04/09	51	4.8	0.2	0.6
22/04/09	37	0.6	0.0	0.1
04/05/09	61	8.4	0.5	1.2
18/05/09	31	9.8	0.3	0.7
01/06/09	51	6.4	0.3	0.8
05/06/09	51	10.2	0.5	1.2
08/06/09	27	3.4	0.1	0.2
12/06/09	37	7.8	0.3	0.7
17/06/09	27	5.2	0.1	0.3
30/06/09	51	2.4	0.1	0.3
10/07/09	35	5.5	0.2	0.5
04/08/09	27	15.2	0.4	1.0
12/08/09	51	12.3	0.6	1.5
20/08/09	37	15.3	0.6	1.3
28/10/09	37	6.0	0.2	0.5
30/10/09	37	7.4	0.3	0.7
06/11/09	52	3.5	0.2	0.4
19/11/09	52	8.1	0.4	1.0
24/11/09	51	11.3	0.6	1.4
19/12/09	51	7.7	0.4	0.9
Total	43.7	238	10.4	24.8

Table 2.2: Illumina genomic paired-end sequencing data

Date	Mean length (bp)	Number (Millions)	Combined Size (Gb)	Coverage
26/01/09	51	6.6	0.4	0.8
16/02/09	51	3.0	0.2	0.4
16/02/09	51	3.2	0.2	0.4
16/02/09	51	3.0	0.2	0.4
06/03/09	52	15.0	0.8	1.8
06/03/09	52	15.8	0.8	2.0
12/03/09	52	16.4	0.8	2.0
19/04/09	52	18.8	1.0	2.4
28/04/09	52	10.0	0.6	1.2
08/05/09	52	15.4	0.8	2.0
08/05/09	51	15.4	0.8	1.8
22/06/09	52	22.4	1.2	2.8
01/07/09	49	14.0	0.6	1.6
27/08/09	52	8.0	0.4	1.0
17/09/09	52	8.0	0.4	1.0
26/10/09	38	8.6	0.4	0.8
04/11/09	52	8.4	0.4	1.0
12/11/09	52	7.0	0.4	0.8
02/12/09	52	4.6	0.2	0.6
16/12/09	52	5.8	0.2	0.8
Total	51.1	213.5	10.9	26.0

Table 2.3: Roche genomic sequencing data

Type	Mean length (bp)	Number (Millions)	Combined Size (Gb)	Coverage
FLX	220	1.8	0.4	0.9
Titanium	341	2.2	0.7	1.8
Total	286	3.9	1.1	2.7

Table 2.4: Sequencing data summary

Read Type	**Pre-filtered**				**Post-filtered**			
	Mean length (bp)	Number (Millions)	Combined size (Gb)	Coverage	Mean length (bp)	Number (Millions)	Combined size (Gb)	Coverage
Illumina gemomic	47.2	451.5	21.3	50.7	44.13	338.5	14.9	35.6
Roche genomic	286	3.9	1.1	2.7	301	3.6	1.1	2.6
Illumina RNA-seq	101	80	8.1	n/a	101	48	4.9	n/a

2.4 Filtering

Errors in sequencing data cause problems during assembly (see Section 3.1.1). Therefore, producing high quality data is essential, and where possible producing large amounts of data to allow quality filtering is desirable. It has already been shown in table 1.3 that HTS can produce enough data but it was important to determine the read quality? Early quality checks identified significant error rates associated with both data types.

2.4.1 Quality

The Illumina data showed a significant decline in quality towards the end of the reads. Figures 2.2 and 2.3 display the mean quality score at each base position calculated using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) for the entire set of pre- and post-filtered Illumina reads. Filtering here used a custom perl script which used the quality score of each base given in the FASTQ file and a sliding window filter whereby a length of sequence at least 30 bp long with quality greater than 20 allowing up to 1 low quality base. The minimum length of retained sequences was chosen following preliminary assembly tests suggesting an optimum k-mer of 27 (Section 3.1.1). Figures 2.4 and 2.5 show the length distributions of the reads that were retained, an artefact of the rapid advances of read length and throughput from the Illumina machines in the early stages of use.

As the transcriptomic data was received in 2011 the data was significantly better, both in terms of length and quality, and by that time new tools were available for filtering. The data used consisted of 80 million 101bp reads, the quality of which can be seen in the FastQC plot Figure 2.6. A filtered data set was produced using tools from the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html). After preliminary analysis the quality filter script was used with q 10 and p 90, increasing the overall quality (Figure 2.7) but reducing the data to 48 M reads (Table 2.4).

Roche 454 data has known issues with homopolymer runs (adjacent identical bases). Due to the method of identifying base incorporation, distinguishing between multiple

bases of the same type is problematic and often incorrect [76]. Identifying and fixing these regions however was not possible, but very short or reads of extreme base composition could still be identified. Therefore, very short reads (less than 100 bases) and very long reads (greater than 600 bases) were removed. In addition reads containing very low base heterogeneity were removed using a custom perl script. These included single base type reads and low complexity repeats. Due to the lower number of Roche reads, filtering was less stringent and resulted in a drop from 3.9 to 3.6 million reads (Figures 2.8 and 2.9).

2.4.2 Additional screening

Initial assemblies produced a complete 15,658 bp mitochondria contig. All reads were mapped to this using Bowtie [90] version 0.11.3 with default settings and all positive mappings were removed from subsequent assemblies resulting in the removal of over 500,000 reads and significant improvements in assemblies.

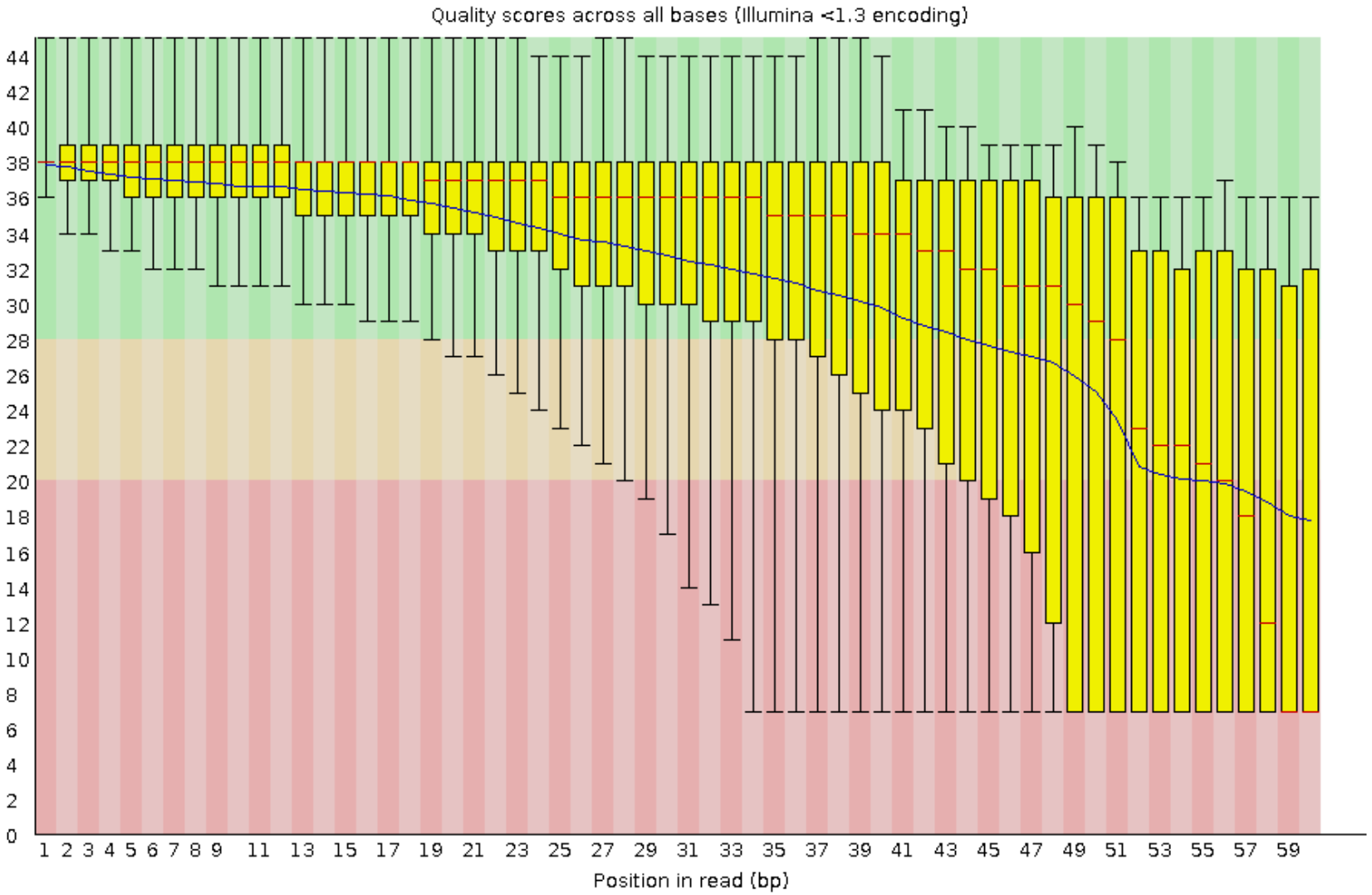

Figure 2.2: FASTQ quality scores of all pre-filtered Illumina genomic data (generated using FastQC)

The y-axis on the graph shows the quality scores. The higher the score the better the base call. The background of the graph divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). Yellow boxes represent the inter-quartile range from the 25th to 75th percentile, red lines the median value black whiskers the 10th and 90th percentile and the blue line the mean value.

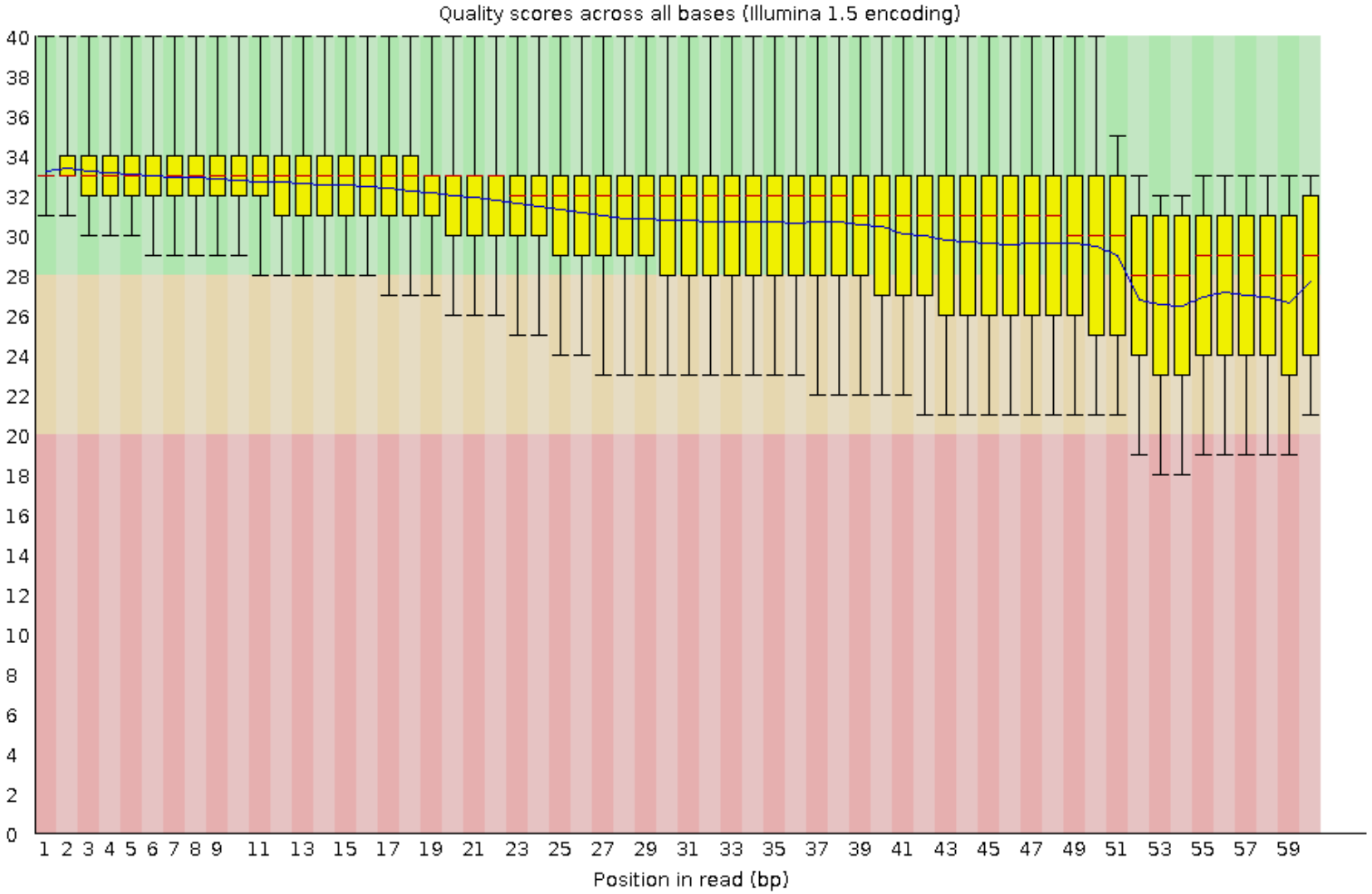

Figure 2.3: FASTQ quality scores of all post-filtered Illumina genomic data (generated using FastQC)

The y-axis on the graph shows the quality scores. The higher the score the better the base call. The background of the graph divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). Yellow boxes represent the inter-quartile range from the 25th to 75th percentile, red lines the median value black whiskers the 10th and 90th percentile and the blue line the mean value.

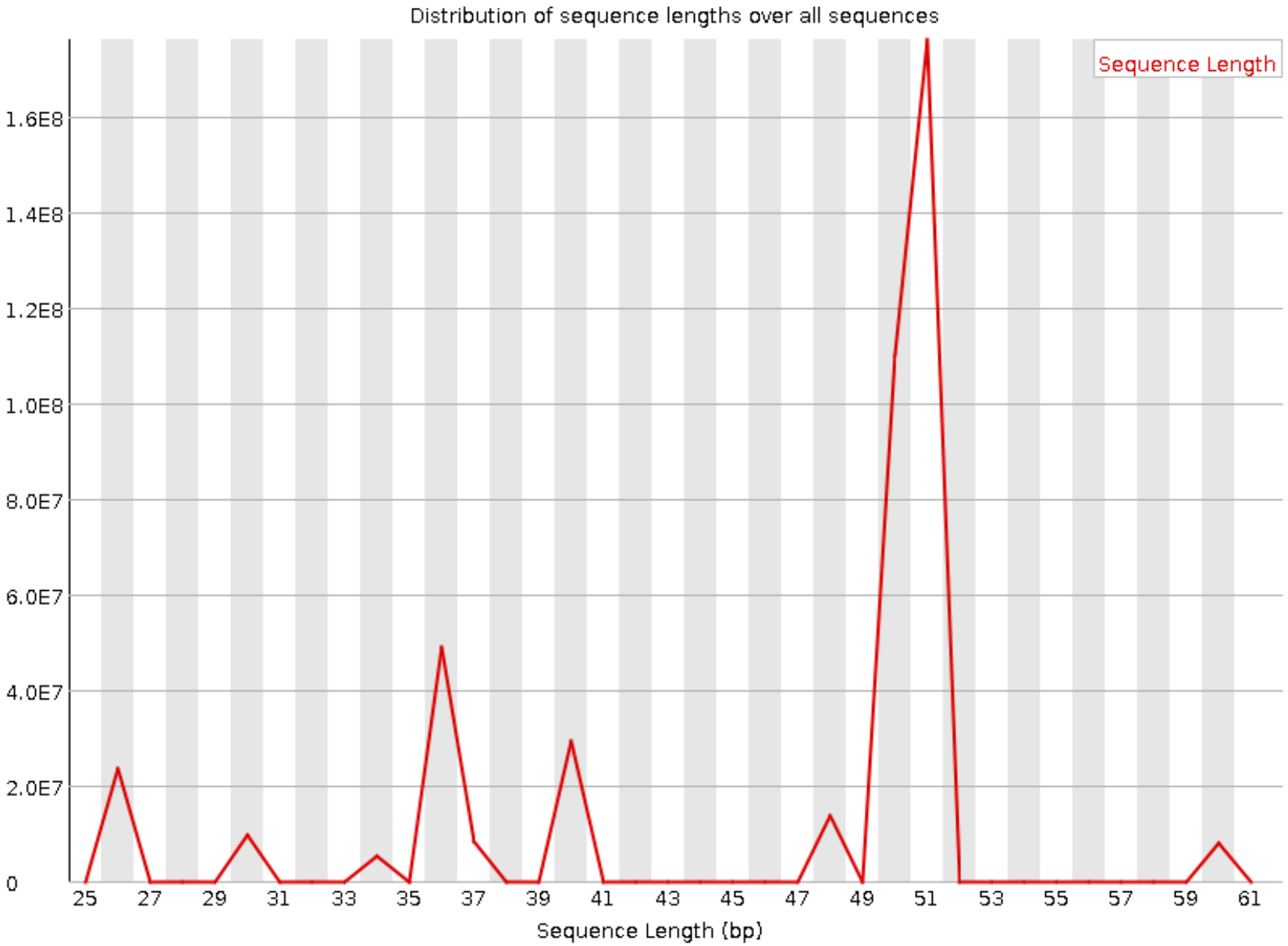

Figure 2.4: Distribution of read lengths of all pre-filtered Illumina genomic data (generated using FastQC)

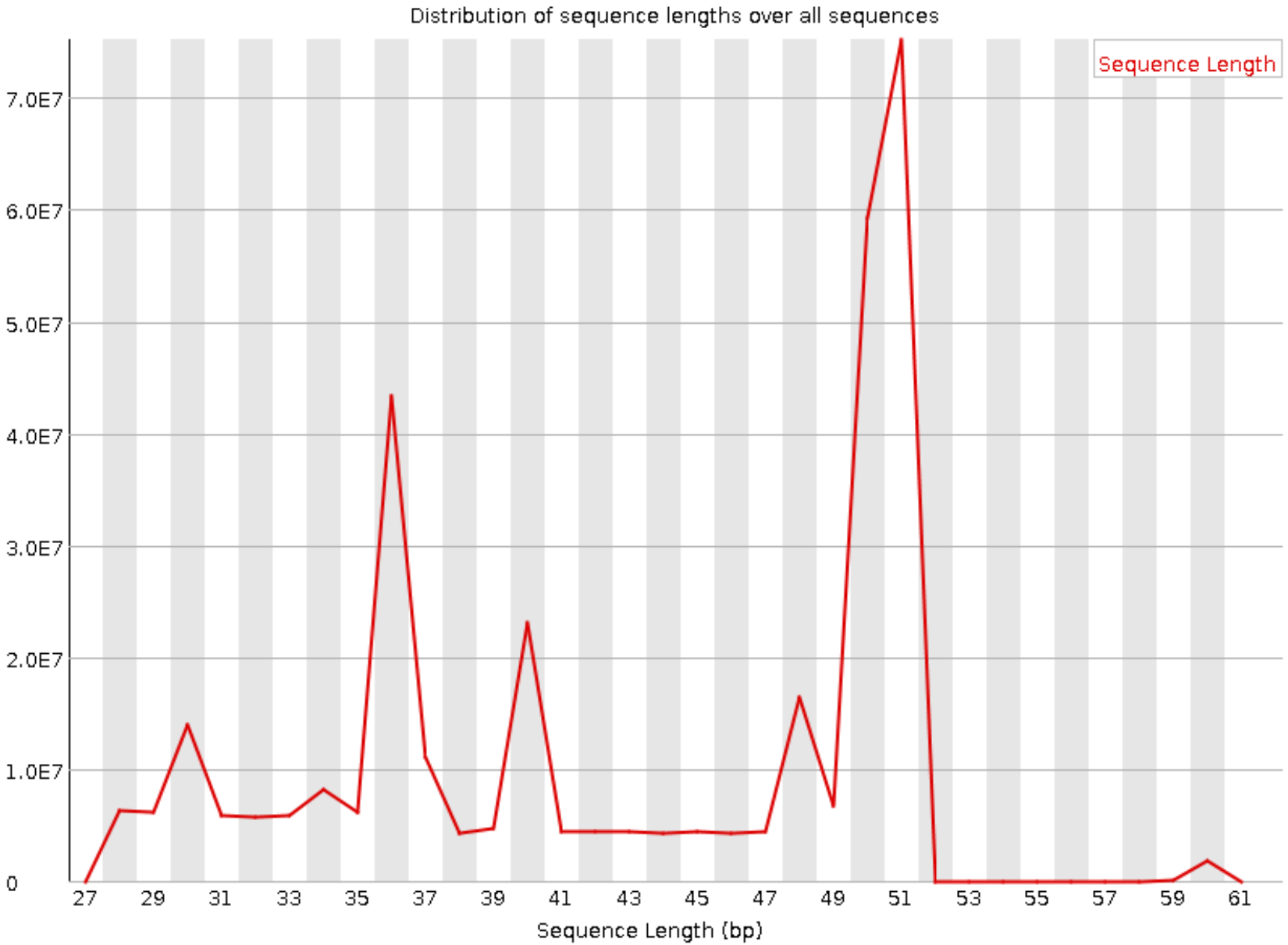

Figure 2.5: Distribution of read lengths of all post-filtered Illumina genomic data (generated using FastQC)

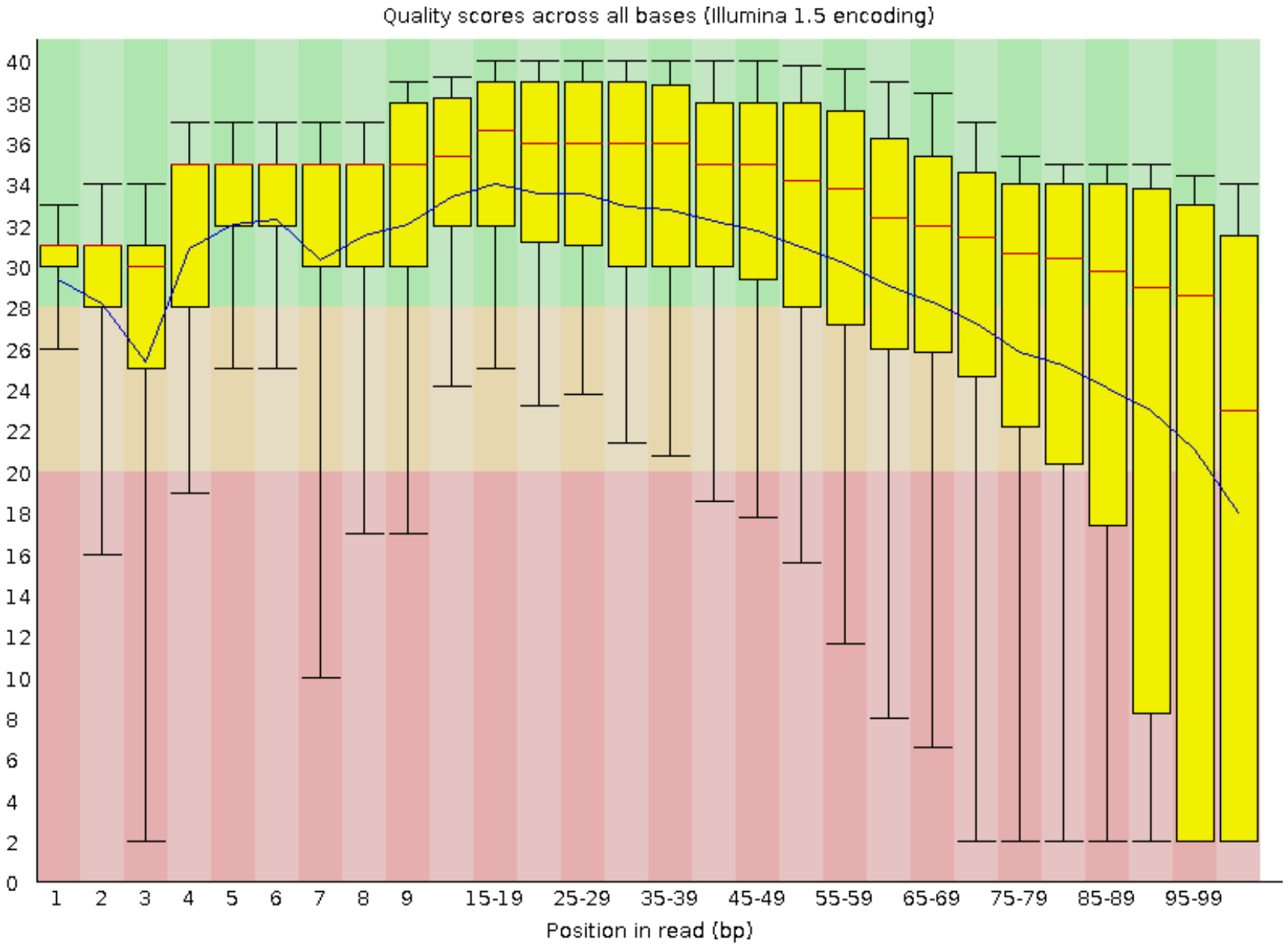

Figure 2.6: FASTQ quality scores of pre-filtered Illumina transcriptomic data (generated using FastQC)

The y-axis on the graph shows the quality scores. The higher the score the better the base call. The background of the graph divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). Yellow boxes represent the inter-quartile range from the 25th to 75th percentile, red lines the median value black whiskers the 10th and 90th percentile and the blue line the mean value.

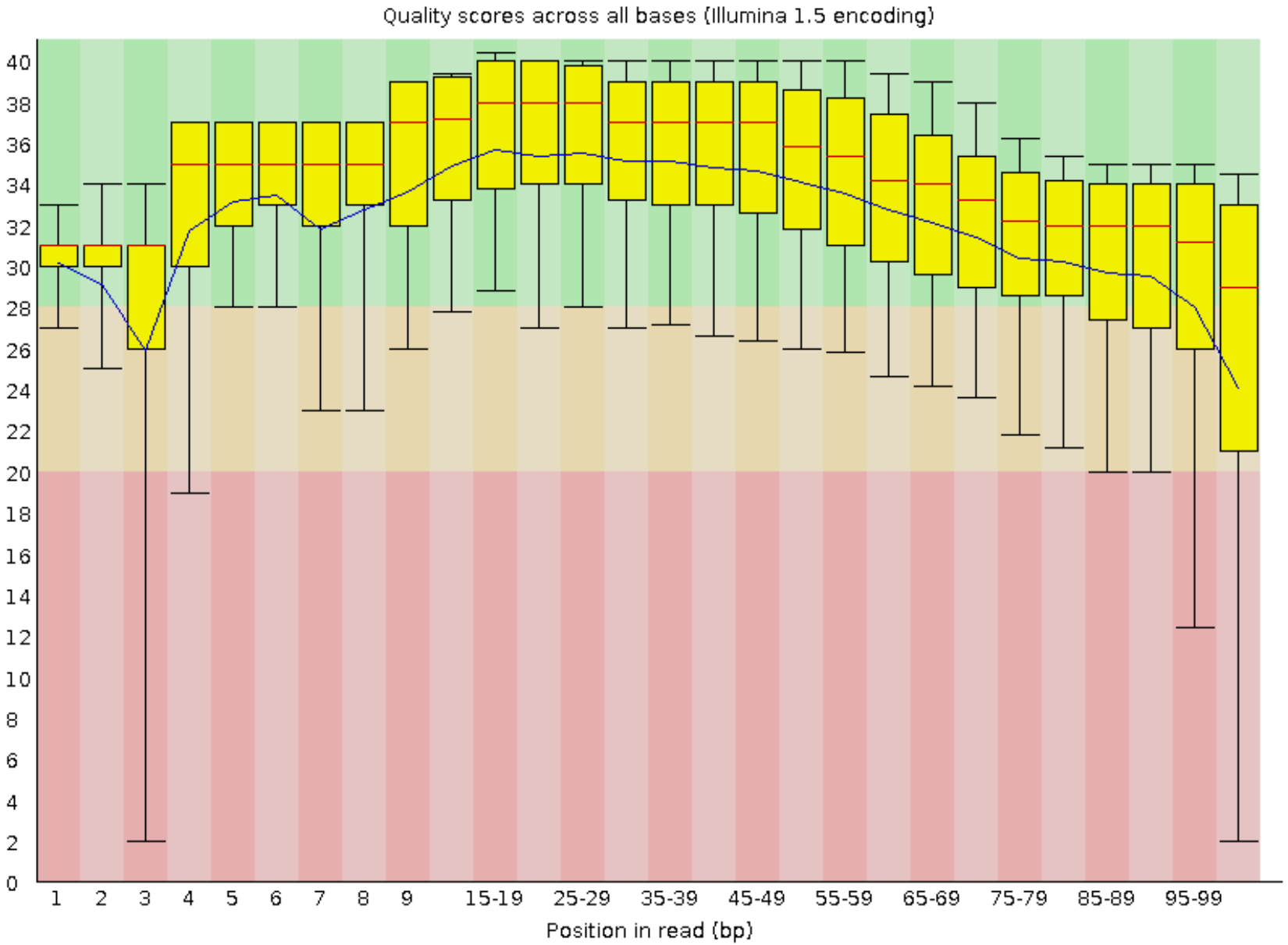

Figure 2.7: FASTQ quality scores of post-filtered Illumina transcriptomic data (generated using FastQC)

The y-axis on the graph shows the quality scores. The higher the score the better the base call. The background of the graph divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). Yellow boxes represent the inter-quartile range from the 25th to 75th percentile, red lines the median value black whiskers the 10th and 90th percentile and the blue line the mean value.

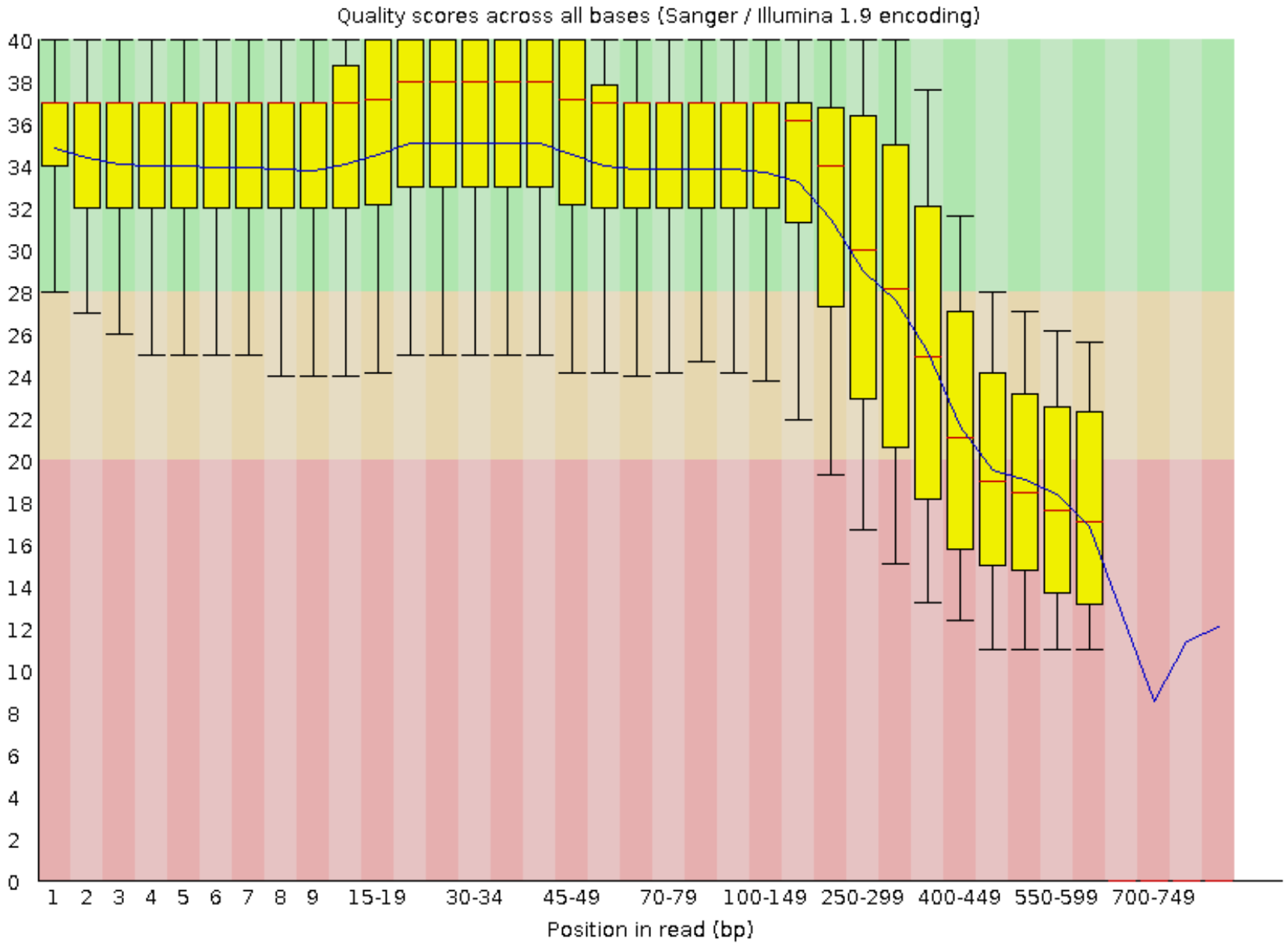

Figure 2.8: FASTQ quality scores of pre-filtered Roche genomic data (generated using FastQC)

The y-axis on the graph shows the quality scores. The higher the score the better the base call. The background of the graph divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). Yellow boxes represent the inter-quartile range from the 25th to 75th percentile, red lines the median value black whiskers the 10th and 90th percentile and the blue line the mean value.

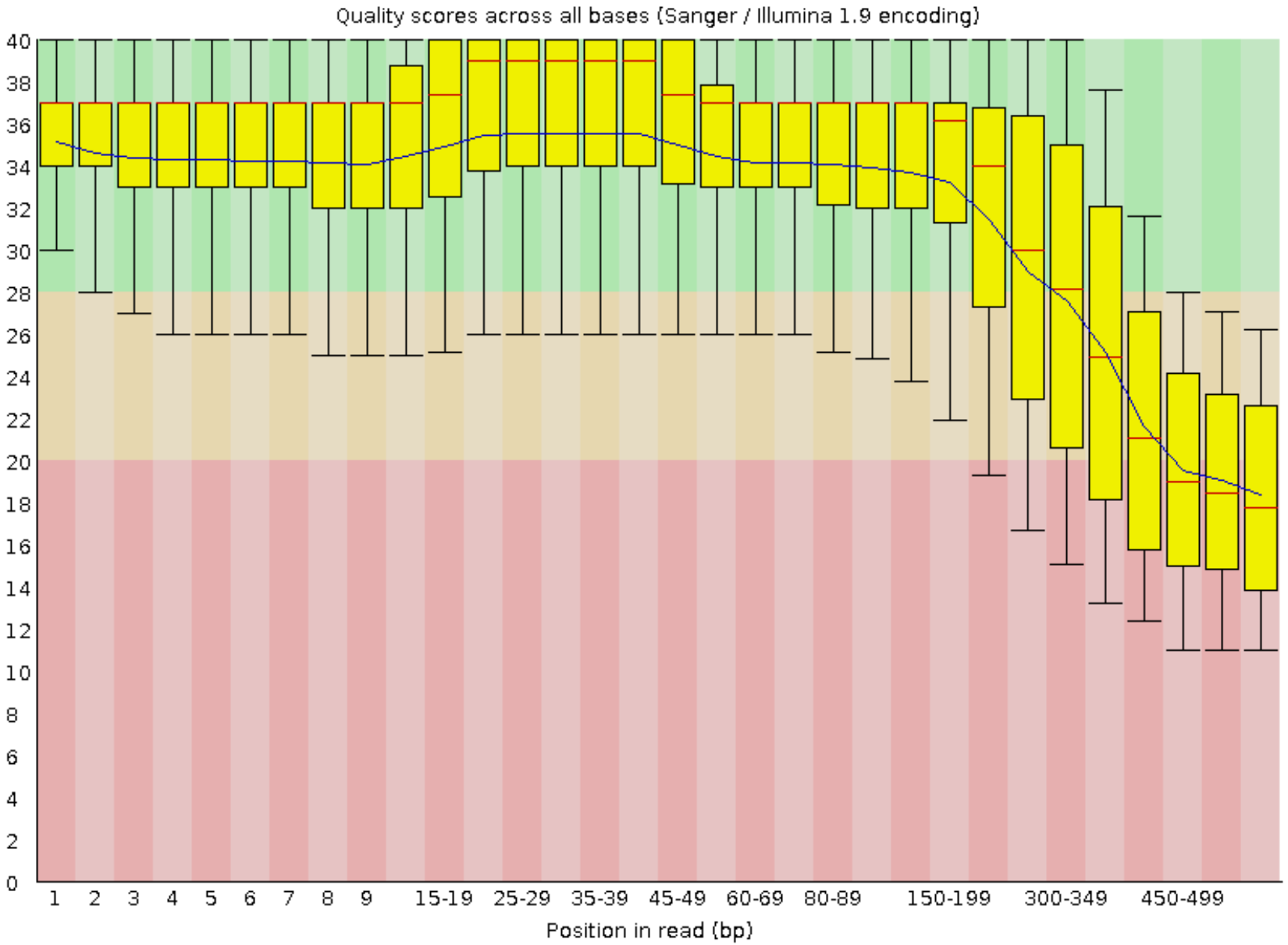

Figure 2.9: FASTQ quality scores of post-filtered Roche genomic data (generated using FastQC)

The y-axis on the graph shows the quality scores. The higher the score the better the base call. The background of the graph divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). Yellow boxes represent the inter-quartile range from the 25th to 75th percentile, red lines the median value black whiskers the 10th and 90th percentile and the blue line the mean value.

Chapter 3

Assembly

3.1 *De novo* assembly using high-throughput sequencing

3.1.1 Genome

NCBI genome statistics as of December 2011 stated that there were currently only 36 complete Eukaryotic genomes in GenBank with many more as drafts or in progress (Table 3.1). Although this data is missing some key complete genomes, the large number of genomes still in draft form indicates the massive amount of time and money that are required to finish a genome, and those in progress are likely to be new HTS projects. These numbers suggest a huge increase in genome projects, due predominantly to the development of HTS removing the bottleneck of sequencing time and cost.

Table 3.1: Genome sequencing project statistics

Organism	Complete	Draft assembly	In Progress	Total
Prokaryotes	1117	966	595	2678
Archaea	100	5	48	153
Bacteria	1017	961	547	2525
Eukaryotes	36	319	294	649
Animals	6	137	106	249
Mammals	3	41	25	69
Birds	-	3	13	16
Fishes	-	16	16	32
Insects	2	38	17	57
Flatworms	-	3	3	6
Roundworms	1	16	11	28
Amphibians	-	1	-	1
Reptiles	-	2	-	2
Other animals	-	20	24	44
Plants	5	33	80	118
Land plants	3	29	73	105
Green Algae	2	4	6	12
Fungi	17	107	59	183
Ascomycetes	13	83	38	134
Basidiomycetes	2	16	11	29
Other fungi	2	8	10	20
Protists	8	39	46	93
Apicomplexans	3	11	16	30
Kinetoplasts	4	3	2	9
Other protists	1	24	28	53
Total	1153	1285	889	3327

Data obtained from http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html (December 2011).

Small labs now have the opportunity to obtain, or even produce for themselves, the volumes of data necessary to sequence entire genomes, a luxury once restricted to only the biggest and wealthiest of institutes. However, the previous methods of assembly and the computer power associated with them can not be scaled to fit this new HTS data, and therefore a new approach to analysing the data is necessary.

This magnitude change in read number brought on by HTS has for the moment ended the reign of the overlap-consensus-layout method. The number of possibilities in an overlap graph with shorter read lengths and high coverage is huge. For example, a million short reads will require a trillion pairwise alignments, and modern machines are producing billions of reads per run.

For the majority of cases, this computational problem is solved with a de Bruijn graph [38], a directed graph structure capable of representing an assembly and crucially capable of scaling up to large assemblies. The first application of this approach to genome assembly was in the EULER assembler in 2001 [130] and has more recently and successfully been applied to short read data in many others such as Velvet [184], ABySS [147], SOAPdenovo [96], ALLPATHS-LG [60] and the commercial CLCBio (http://www.clcbio.com).

The basic process of de Bruijn assembly based on the Velvet version is summarised in Figure 3.1. Here the de Bruijn graph was altered from its original use in 2001 to map k-mers onto nodes instead of arcs and also includes the reverse complementary sequences creating a bi-directed graph. Identical k-mers collapse into the same nodes representing the level of coverage. Many of the assemblers use this level of coverage to guide the assembly process. Nodes with coverage well below that expected can be ignored, and very high coverage nodes which may represent repeats can be approached in a different manner.

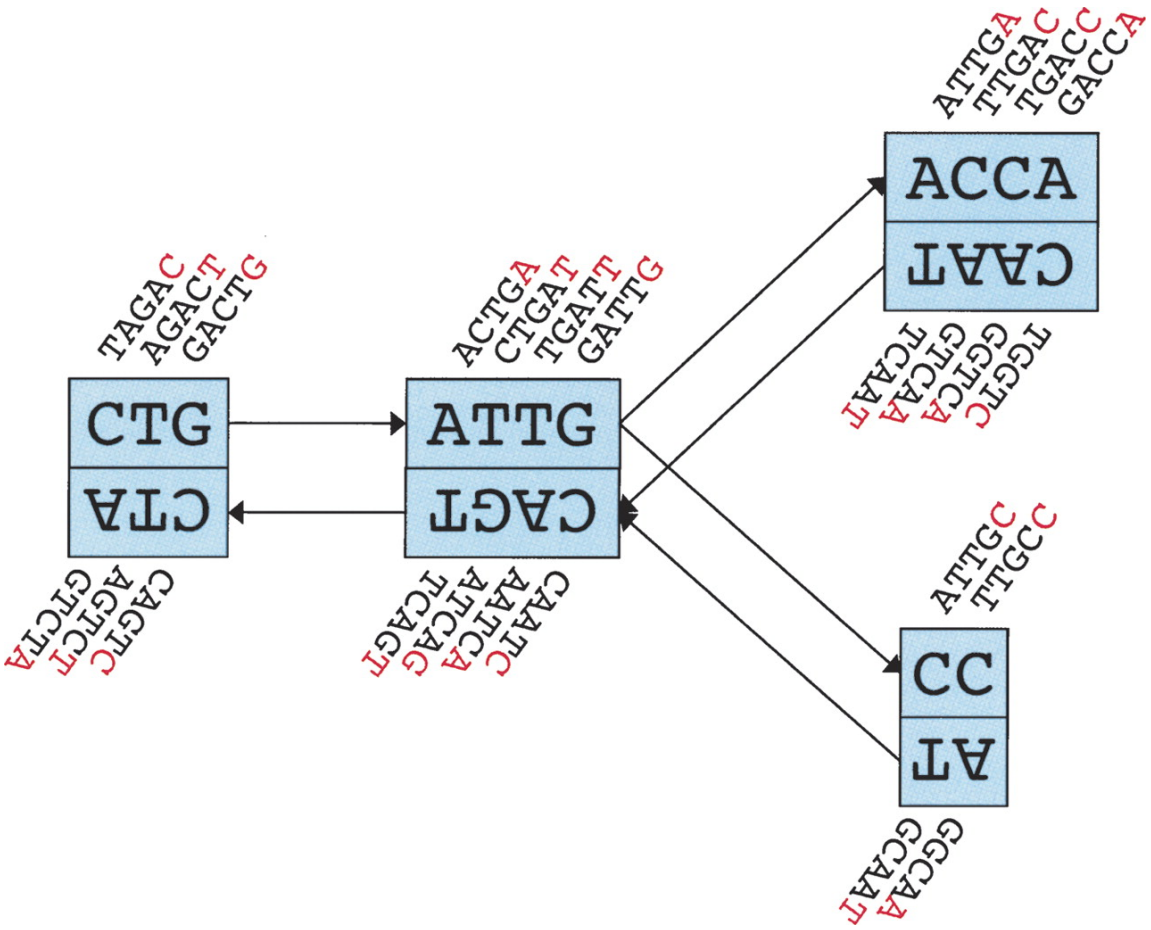

Figure 3.1: Application of a de Bruijn graph in the Velvet assembler (taken from [184])

Each rectangle represents a series of overlapping k-mers (in this case, k=5) with the reverse complement above or below. Each k-mer is represented by its final character. Arcs, indicating presence of linked k-mers in reads, are represented as arrows between the nodes and the last k-mer of an arc's origin overlaps with the first of its destination.

The de Bruijn graph introduces us to the idea of applying k-mers rather than reads to genome assembly. A k-mer in this sense is a string of DNA with length k, which is less than the length of the read. At first it seems counterintuitive to break what is already a very small fragment of the genome into an even smaller piece. However, this approach collapses data into non-redundant units, thus optimising memory requirements and reducing the impact of read errors. Compeau *et al* [29] cover the application of de Bruijn graphs to genome assembly in detail.

Even though a k-mer is relatively short it still contains a high level of uniqueness. Given that there are four possibilities at each base, then 4 to the power of the k-mer length quickly becomes a very big number as k increases. Figure 3.2 shows that even at relatively short k-mer length the percentage of unique k-mers is high, and inversely proportional to the complexity of the genome. Early HTS assembling projects confirmed this, as assembling low complexity genomes, i.e. bacteria, with short read data occurred with much greater success than larger more complex genomes.

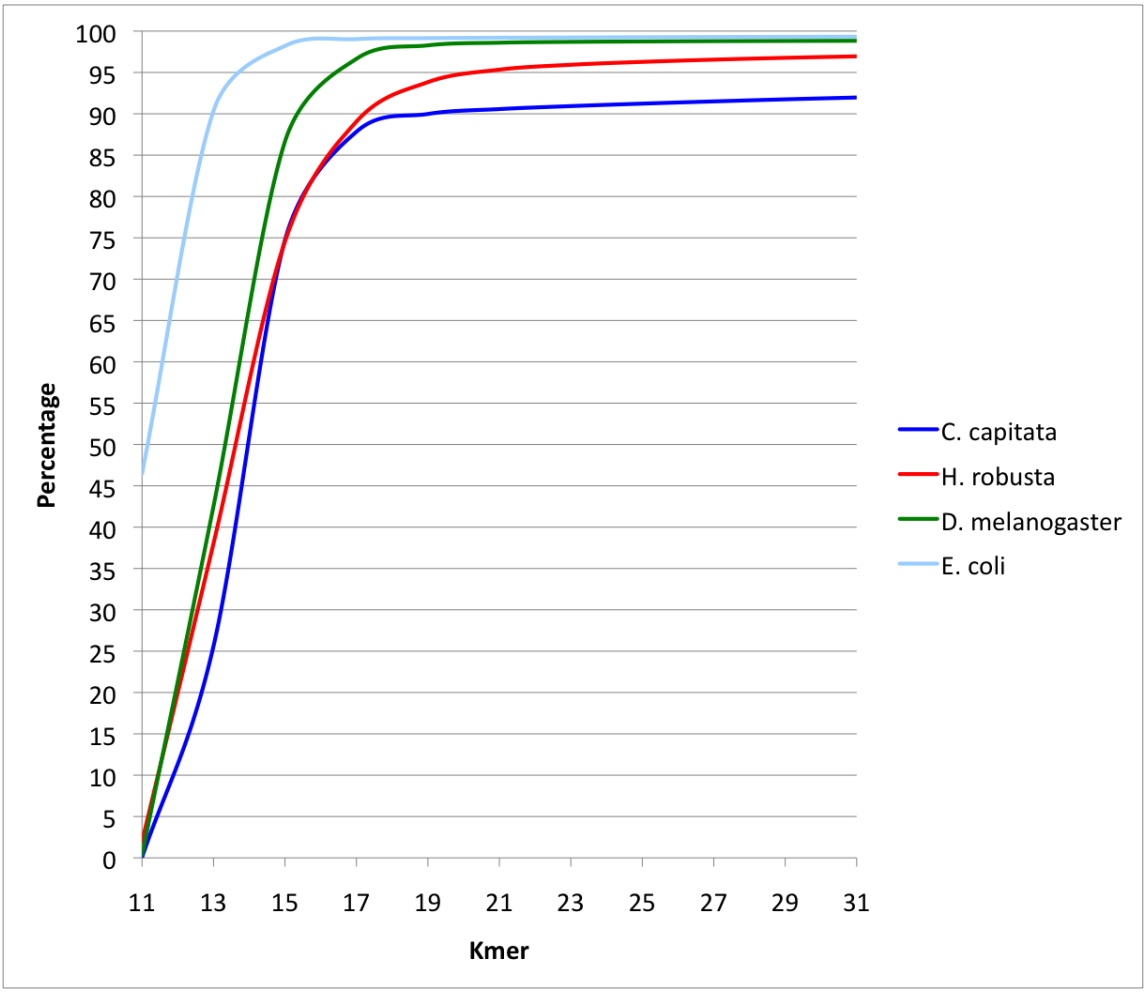

Figure 3.2: Percentage of unique k-mers within a genome

Calculated using Jellyfish [105], using data from http://genome.jgi-psf.org/Capca1/Capca1.home.html (*C. telata*), http://genome.jgi-psf.org/Helro1/Helro1.home.html (*H. robusta*), http://www.fruitfly.org/sequence/release5genomic.shtml (*D. melanogaster*) and http://www.ncbi.nlm.nih.gov.ezproxy.webfeat.lib.ed.ac.uk/nuccore/FN554766 (*E. coli*).

Initially many de Bruijn tools were limited to a k-mer of 31 due to memory limitations on the graph structure. Even though this relatively short k-mer has a high uniqueness, as HTS read lengths and throughput increased so did the usable k-mer length and software was adapted to use this. However at the time of this project the average read length was around 50 bases and optimum k-mer less than 31.

3.1.2 Transcriptome

A transcriptome is the complete set of transcripts that lead to the production of an organism's proteome via translation. Prior to the possibilities of genome sequencing, the preferred method for generating transcription data was to sequence sections of messenger RNA (mRNA) and produce ESTs, a relatively quick and cheap method for identifying coding regions. As of October 1st 2011 there were 70,937,429 ESTs in GenBank from 2,313 species. This massive data set has been generated over the last 25 years and has proved invaluable in the progression of molecular biology, especially when annotating new transcripts, both for training *ab initio* gene finders and annotating by homology the peptide sequences produced by the gene predictions themselves.

HTS brought with it the ability to sequence very large volumes of mRNA simultaneously and at high quality to produce RNA-Seq data. However, unlike ESTs which use Sanger sequencing, the reads generated are relatively short and need sophisticated assembly methods. Initially the methods used to assemble RNA-Seq data were based on the first HTS genome assemblers [115] but it quickly became apparent that new approaches were needed. This was primarily because of the huge variation in coverage per locus that transcriptomic data produces, as genes are transcribed at very different levels. This meant that using an expected coverage to guide the assembly was not going to work.

If a genome is already available then it can be used to guide transcriptome assembly. However, as the expectation for the *L. rubellus* genome was a heavily fragmented assembly and one of the plans for the transcriptome was to help address this, it was decided to assemble the transcriptome *de novo*.

3.2 Review of assemblers

3.2.1 Genome

Between 2005 and 2010 there were 24 *de novo* genome assemblers either written from scratch or modified from old assembly algorithms, and more have been released since [146]. This number reflects both the various types of data that have emerged over the last few years, each with their own error profile and read type, and the pressing need to develop a versatile, reliable and functional assembly tool. A practical comparison of many of them was recently performed by Zhang *et al* [186] and the recent assemblathon [42] compared many of them in a rigorous set of trials, and represents the state of the art in assembly at the present time. There has also been investigation into development of new metrics for comparing the final assemblies as well as comparing new HTS to the older Sanger based assemblies [120].

For this study I will discuss only those methods that were designed for, or were theoretically capable of assembling, large eukaryotic genomes using HTS data and were considered best practice at the time. Table 3.2 lists the genome assemblers that were investigated during this project. The following gives a brief description of each and the outcome of attempted assemblies.

ABySS (Assembly By Short Sequences) [147]

In 2009, ABySS, the first short sequence assembler with built in parallel support, was released making this the only non-commercial assembler at the time capable of large genome *de novo* assembly (without investment in large memory machines). Parallelisation is achieved using a novel distributed de Bruijn graph which spreads the graph over many computer nodes whilst retaining a single structure utilising the message passage interface (MPI). The assembly proceeds in two phases. The initial phase of assembly constructs the distributed graph from overlapping k-mer information. Errors in the graph are removed and then unambiguous nodes are connected. The second phase uses the paired read in-

formation to scaffold the contigs. The first phase is best run on a cluster or in a grid, but as the second phase is fundamentally serial it can be run locally on a desktop machine. This method proved successful due to the distributed graph spreading the huge memory requirements of the de Bruijn graph over many machines.

ALLPATHS

Released in a theoretical form in 2008 [23] and then in a functional form as ALLPATHS-LG in 2010 [60], this is a de Bruijn based assembler which requires multiple insert size mate pair libraries. It was among the best assemblers at the recent assemblathon [42]. However, due to its mate-pair requirement it was not possible to use ALLPATHS-LG for this project.

CLCBio (www.clc.com)

This is the only commercial software capable of assembling large genomes using short reads using 'reasonable' amounts of compute power. The actual methodology remains unknown, although it is de Bruijn based and uses amazingly low amounts of memory and time.

IMAGE (Iterative Mapping and Assembly for Gap Elimination) [170]

Although not an assembler in its own right, this method attempts to close the gaps in a draft assembly by aligning reads to the ends of contigs and running local assemblies of the mapped reads and their pairs. Iterations of the process can then be run to try and improve the number of gaps closed. For this project the use of IMAGE was problematic due to the fragmented nature of the genome and relatively low number of paired reads causing both time and data alignment issues. There is an unreleased parallel version that may solve these issues, but until either this is released or the contiguity of the assembly improves, application of this tool is not possible.

Newbler (www.roche.com)

Roche's commercial Genome Sequencer (GS) De Novo Assembler, also known as Newbler, is a custom built assembler designed to be used with Roche 454 data. It uses the SFF files directly in an overlap-layout-consensus method to create pairwise alignments, resolve branching structures between contigs and then create a consensus sequence for each contig taking into account the quality information. Paired end data can be subsequently used to scaffold the contigs. Like CLCBio this is closed-source software so little is known about its inner workings but as it is designed by Roche for Roche data it was considered the best option available.

Ray [19]

Ray is another assembler that can be run in parallel on a grid and therefore spread memory load. It is designed to be a true hybrid assembler using both the Illumina and Roche type data, and therefore would seem the perfect choice for this project. However, after numerous trials no successful assemblies were generated. No published genomes have been assembled with this software.

SOAPdenovo (Short Oligonucleotide Analysis Package) [96]

Another top scorer in the assemblathon [42], SOAPdenovo has been widely used and proved successful in many genome projects, especially those projects associated with the Beijing Genomics Institute (BGI) such as the panda genome [95]. Again, it is structured around a de Bruijn graph but also incorporates error correction of reads and requires multiple insert size mate-pair data. For this latter reason this assembler could not be used for this project.

Velvet [184]

Velvet is a widely used and excellent assembler, and was the first to really utilise the de Bruijn graph for HTS and produce high quality assemblies. With the ability to include

longer reads as part of the Rock Band [185] scaffolding phase of assembly Velvet would have been an assembler of choice. However, Velvet requires the whole de Bruijn graph to be in memory accessible to a single core and thus, this software would only be appropriate for large genomes if a high RAM machine (greater than 1 TB) was available.

Table 3.2: *de novo* genome assemblers

Assembler	Data	Pairs	Mate-pairs	Structure	Memory
ABySS	Short	Yes	Yes	de Bruijn	Distributed
ALLPATHS-LG	Short	Yes	Yes*	de Bruijn	Single
CLCBio	Short	Yes	Yes	de Bruijn	Single
Image	Short	Yes*	No	Local assembly	Single
Newbler	Long	Yes	No	Overlap-layout-consensus	Single
Ray	Mixed	Yes	No	de Bruijn	Distributed
SOAPdenovo	Mixed	Yes	Yes	de Bruijn	Single
Velvet	Mixed	Yes	Yes	de Bruijn	Single

* data type is essential

Where memory is listed as 'Single' it implies that the algorithm can only use the memory from one machine, whereas 'Distributed means that memory can be spread across multiple machines, e.g. using a compute cluster.

3.2.2 Transcriptome

Trans-ABySS [136]

Based on ABySS, this software pipeline merges ABySS assemblies performed across a range of k-mer values in an attempt to correctly identify transcripts of variable expression.

Oases [143]

Oases is based on procedures similar to Trans-ABySS but using Velvet assemblies. Oases aims to cluster contigs into transcripts and works very quickly with low memory requirements.

Trinity [62]

The only custom built *de novo* transcriptome assembler assessed here, Trinity is a novel method designed for reconstructing a transcriptome from RNA-Seq data. It combines three independent pieces of software. First, **Inchworm** assembles the RNA-Seq data into unique sequences of transcripts. Second, **Chrysalis** clusters the Inchworm contigs into clusters using a de Bruijn graph structure for each cluster. Lastly, **Butterfly** processes each of the individual clusters in parallel and attempts to identify both full-length transcripts and their alternatively spliced isoforms. Even when alterations are made to improve parallelisation of this method, its run time is proportional to the volume of data and can take many days to complete.

3.3 Chosen assembly methods

3.3.1 Genome

After assessing each of the assembly tools in 3.2.1, the only assemblers suitable for assembly of *L. rubellus* with the data and compute power available were CLCBio, ABySS and Newbler. Detailed investigations and trial assemblies led to the final parameter selections

(Table 3.3) for the complete assembly strategy (Figure 3.5). The decision to use multiple assemblers rather than just one was taken with the aim of maximising assembly space whilst using multiple information sources to confirm the reliability of a contig [188]. The final assembly strategy runs in the following way:

Initial assemblies

The filtered set of Roche reads was assembled with Newbler (version 2.3) on a single desktop machine using default parameters. The same data were also combined with the Illumina data and assembled with CLCBio (version 3.1.1), again with default parameters on a single desktop machine. The filtered Illumina data was also assembled independently using ABySS (version 1.2.0) in two stages. The first stage, which forms the de Bruijn graph, was run in parallel on the University of Edinburgh grid (Edinburgh Compute and Data Facility [ECDF] https://www.wiki.ed.ac.uk/display/ecdfwiki/Home) using 300 cores, each with at least 2 GB RAM. The second, which uses the pairing information, was run locally and required less than 10 GB RAM. The parameters used for ABySS were defaults except for a k-mer of 27 and the minimum number of pairs required to consider joining two contigs (n) set at 3.

Merge

The assembled data from ABySS and Newbler were merged using version 2.0.8 of minimus2 (part of the AMOS package http://sourceforge.net/apps/mediawiki/amos/index.php?title=AMOS) using the following parameters OVERLAP=30, CONSERR=0.1, MINID=95 and MAXTRIM=20, where OVERLAP is the minimum required overlap length, CONSERR the maximum consensus error, MINID the minimum overlap percent identity and MAXTRIM the maximum sequence trimming length. This assembly was then merged with the CLCBio contigs, again using minimus2 and the same parameters. This time, however, the singletons identified by minimus2 were discarded under the premise that any ‘real’ data would have been found by both assembly methods.

Redundancy removal

The set of contigs was then collapsed using version 4.0 of CD-HIT [97] set at 99% identity to try and remove any duplicate contigs caused by allelic variation within the worms. After this stage there were still pairs of contigs that shared significant similarity and preliminary annotation identified many instances of identical annotations on these contigs. Therefore an additional reciprocal BLAT [83] analysis was run on all contigs, and those that matched over 90% of their lengths with 90% of their identity were also removed. An unfortunate side effect of this process is the possible removal of distinct sequences that are sufficiently similar to be deemed allelic. In particular, this could result in the removal of 'real' TEs which may only differ by a base or two but are actually the result of a recent transposition event. This loss is not ideal but the removal of allelic content was deemed more important than retaining all copies of TEs.

Filtering

As stated in Section 1.5, the generation of a high quality draft genome requires that efforts have been made to remove contaminating sequences. This was achieved with a custom filtering pipeline. All contigs were searched using BLAST with a bacterial subset of the nt database (`http://www.ncbi.nlm.nih.gov/nuccore`) with a cut off E-value of 1e-5. Positive hits were then BLAST searched with a metazoan subset of the nt database with the same cutoff. The two BLAST results were compared and any contig with a next best hit to a metazoan sequence was retained. The remaining sequences were removed from the assembly.

3.3.2 Transcriptome

Due to time constraints, preliminary studies, correspondence from other lab members and the knowledge that only one of the three softwares above was created specifically for RNA-Seq assembly, the method selected for transcriptome assembly was Trinity. Therefore, using the 48 million FASTX filtered RNA-seq reads the assembly was completed on

a large memory multi-core machine with parameters edge-thr and minimum contig length set at 0.16 and 200 respectively where edge-thr is the Butterfly threshold (controls the complexity of the graphs) and minimum contig length is the minimum length of the final contigs. The assembly produced a very large number of transcripts, probably due to the default settings being very sensitive to read variation, therefore an additional clustering step was performed using UCLUST [44] set at 97% identity (Table 3.3).

Table 3.3: Details of the chosen assembly tools

Stage	Tool	Version	Parameters
Genome			
Initial assemblies	Newbler	2.3	Default
Initial assemblies	CLCBio	3.1.1	k-mer of 25
Initial assemblies	ABySS	1.2.0	k-mer of 27, number of pairs (n) = 3
Merge assemblies	minimus2	2.0.8	OVERLAP=30, CONSERR=0.1, MINID=95 and MAXTRIM=20
Redundancy removal	CD-HIT	4.0	sequence identity threshold (c) = 0.99
Redundancy removal	BLAT	34	Default
Transcriptome			
Assembly	Trinity	2011-08-20	edge-thr = 0.16, minimum length = 200
Redundancy removal	UCLUST	3.0.617	identity (id) = 0.97

3.3.3 Scaffolding the genome

The next logical step in genome assembly is to scaffold the contigs, especially when a genome is in many separate pieces. For HTS the most common solution is the use of mate-pair data but at the time of sequencing this was not an option and hence efforts were made to use the transcriptome. Using a transcriptome to scaffold a genome assembly is not a new idea. Montazavi *et al* [115] used RNA-seq data to scaffold a draft genome of *Caenorhabditis angraria* to great effect, and more recently Riba-Grogunz *et al* [134] published a method for visualising a genome assembly by using the transcriptome to suggest an orientation of the genomic pieces and then viewing this with a Cytoscape [27] plugin. Unfortunately, neither of the these methods were applicable to the *L. rubellus* assemblies probably due to the high fragmentation of the genome (causing issues when mapping either the raw reads in case of Mortazavi method or to assembled transcripts in the Riba-Grogunz method). For these reasons a novel method was created which can be used simply and generally to scaffold an assembly using transcripts.

A custom perl script, SCUBAT.pl (`https://github.com/elswob/SCUBAT`, was written and Figure 3.3 shows an overview of the main stages in scaffolding with SCUBAT. Although this method is prone to creating scaffolds with reduced intron span, it is very useful for extending gene models by joining together gene-linked contigs.

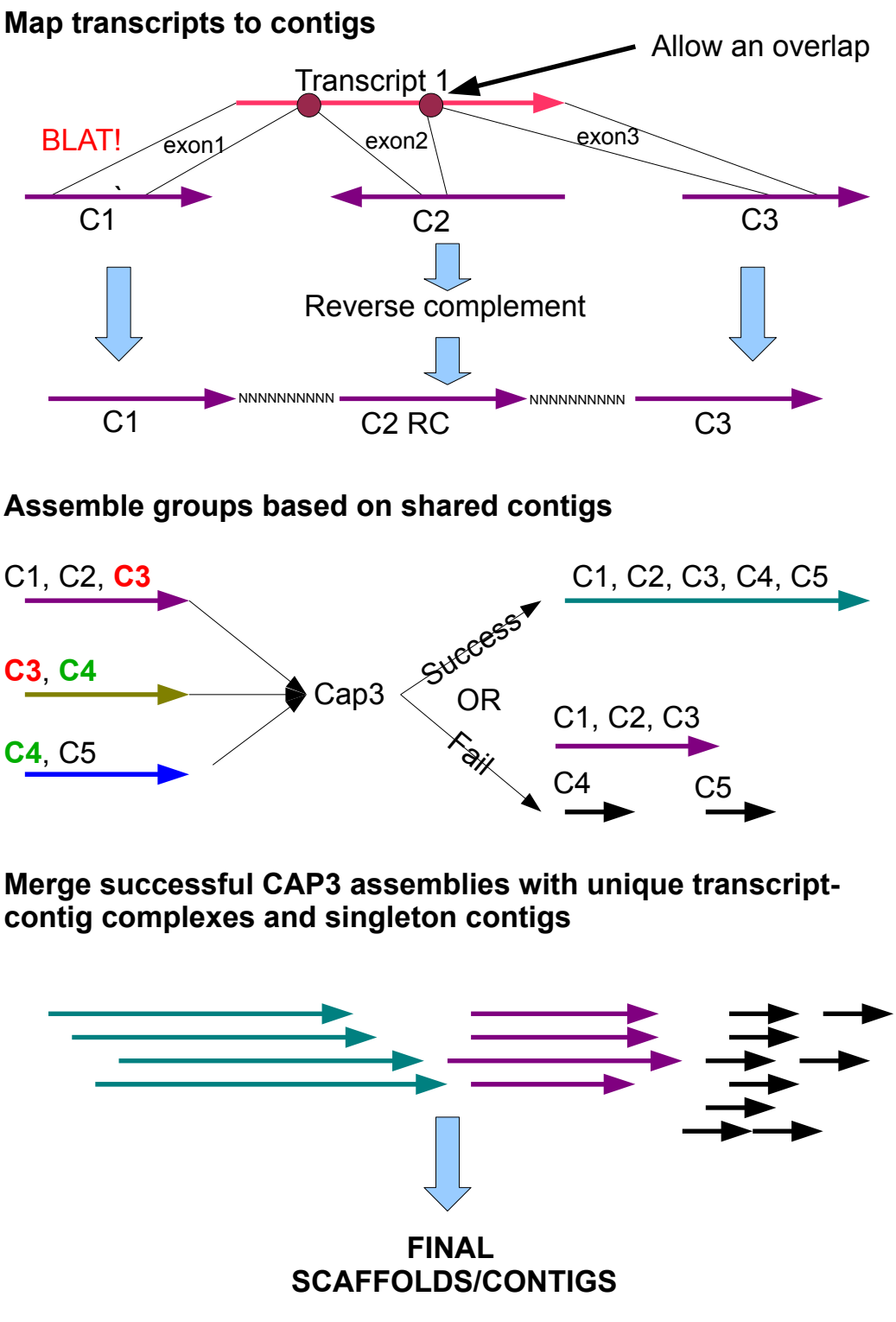

Figure 3.3: Scaffolding contigs using BLAT and transcripts (SCUBAT)

Transcripts are mapped to an assembly using BLAT and fragments joined based on shared hits (exons). Scaffolds are grouped based on shared fragment IDs and assembled using CAP3. Assemblies are assessed for success and a final set of contigs/scaffolds is generated.

Step1: Map the transcripts to the genome assembly

BLAT was used to map Trinity-assembled transcripts and the Lumbribase UniGenes to the genome assembly. This identified consecutive transcript sections (exons) mapping to multiple contigs/scaffolds (Figure 3.4). The mappings are permitted to overlap by a few bases (Table 3.4). In addition, to increase the accuracy of the transcripts used to scaffold, a minimum value for the combined mapping length of the transcript was also incorporated.

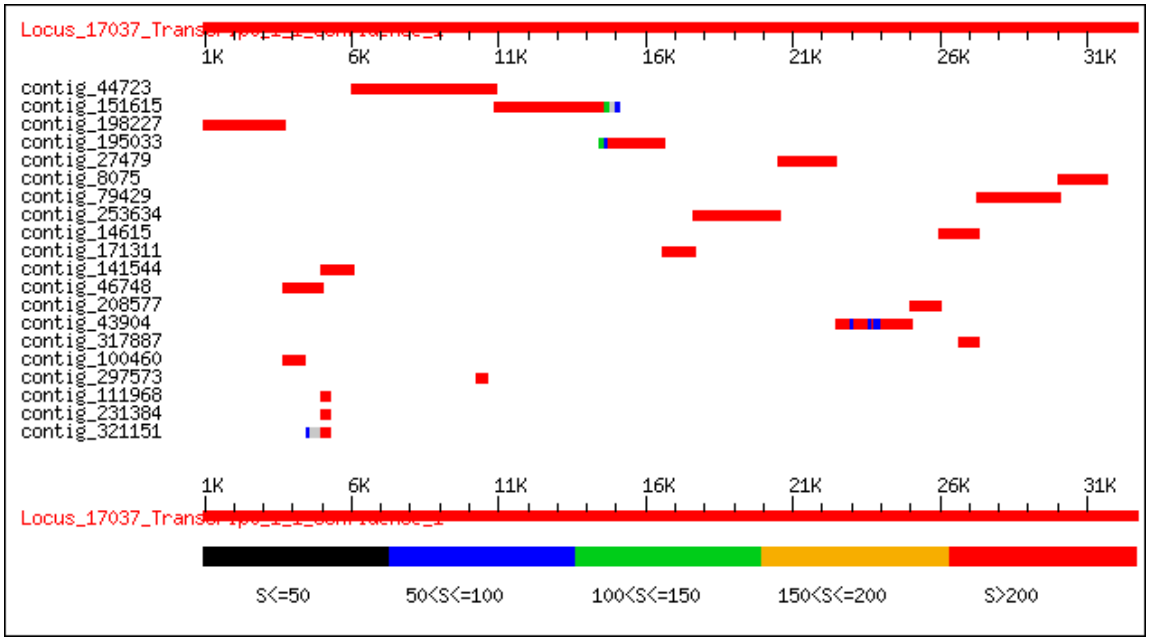

Figure 3.4: Example of transcript BLAST against primary genome assembly

The uppermost red bar is a 31 kilobase transcript, and underneath are the BLAST matches to the exons of that transcript on genomic contigs. This highlights how a single transcript can be used to link multiple contigs.

Table 3.4: BLAT mapping of a 3,393 bp transcript (comp11309_c0_seq1) to the primary assembly.

Contig ID	Length (bp)	Orientation	Transcript start base	Transcript end base	Overlap (bp)
contig_103439	5955	-	19	2110	0
contig_404861	787	-	2106	2325	4
contig_74885	1944	-	2321	2719	4
contig_113361	1402	-	2842	2952	0
contig_97759	1056	-	2950	3188	2
contig_104537	1942	+	3184	3393	4

Step 2: Identify informative split transcripts

All transcripts that mapped to more than one contig were flagged. Each mapping was then analysed to identify those that mapped consecutive non-overlapping sections of the transcript on separate contigs allowing for an overlap buffer zone of 10 bp.

Step 3: Create scaffolds

Each transcript-contig complex was assembled by orientating the contigs based on the BLAT information and adding 10 N's in between the contigs. For example, the contigs in Table 3.3 were scaffolded to form a single contig of over 13,136 bp with all contigs except contig_104537 being reverse complemented.

Step 4: Cluster scaffolds into groups and assemble

The contigs used in each transcript-contig complex were then cross-referenced and any complexes sharing a contig were grouped. Groups were then assembled using CAP3 [72] and default parameters.

Step 5: Filter the assemblies

The assemblies were parsed to identify successful assembly events. Failed assemblies were identified, and the largest complex kept, while the remainder were removed.

Step 6: Create new contig set

A new contig set was generated using the successful CAP3 assemblies, the remaining complexes and any contig not involved.

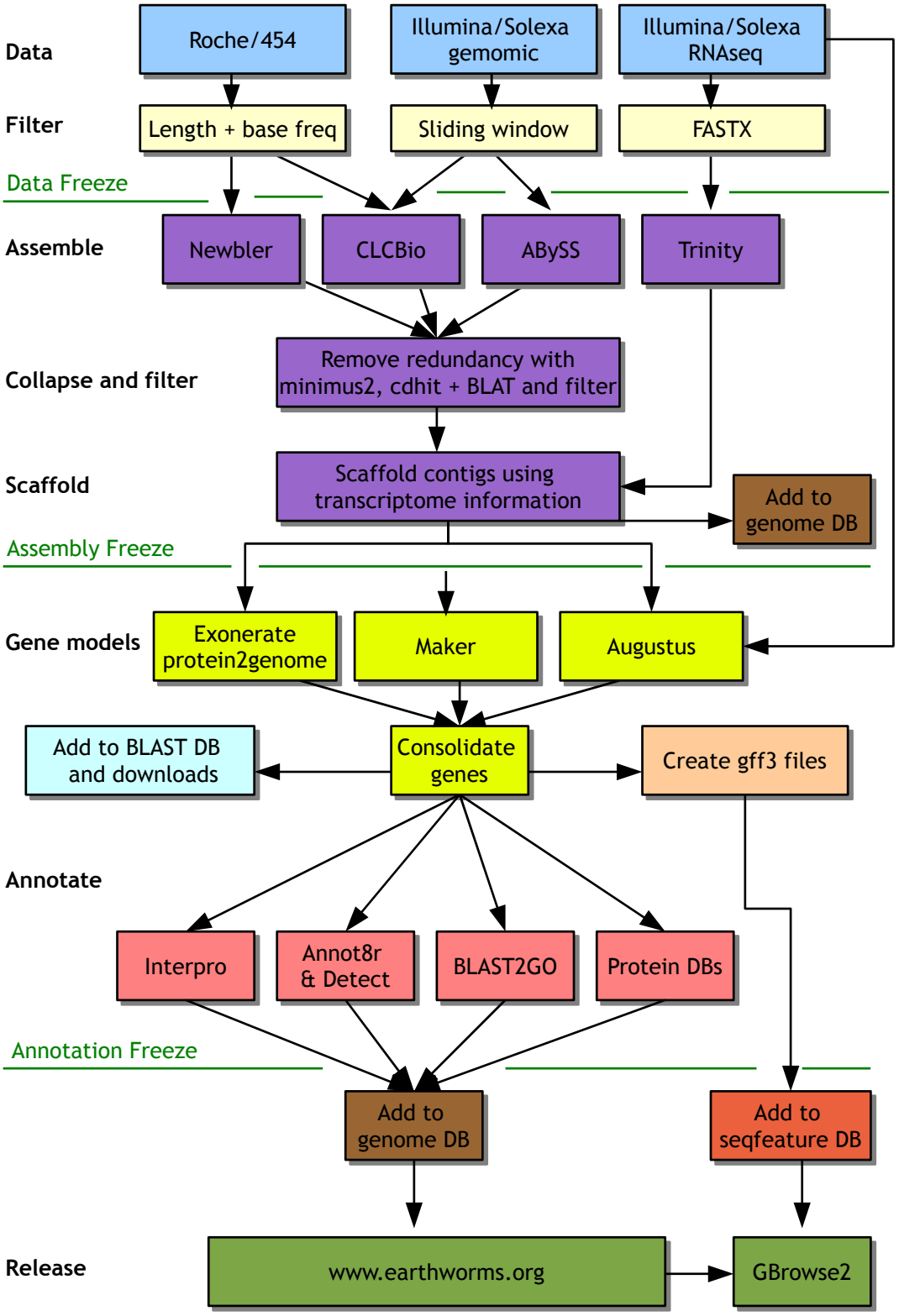

Figure 3.5: Workflow diagram for the assembly and annotation of the *L. rubellus* genome

3.4 Results

3.4.1 Genome

The assembly of the *L. rubellus* B genome was performed following the strategy in Figure 3.5. Combining assemblies increased contig sizes, reduced the overall span and increased the N50 of the assembly (Table 3.5). There was a dramatic reduction in genome size over the three stages, especially when removing putative allelic contigs. This highlights the difficulty of genome assembly even with individuals selected for low heterozygosity. Figure 3.6 presents the three stages of the *L. rubellus* assembly along with the other two annelid genomes in a cumulative assembly graph. This shows that the genome assembly is in many smaller pieces than that of the two JGI genomes as the curves are much more gradual in their inclines, although it is worth noting that the *L. rubellus* genome is substantially larger.

The filtering step identified 68 contigs that were suspected to be bacterial in origin. These had a mean length of 1002 bp and totalled 68 Kb in combined size. This small number suggests that either there was very little bacteria within the sequence data, the merge step removed it, or that the assemblers were very successful in using the low coverage levels of the bacterial data to assemble only *L. rubellus* DNA.

Figure 3.7 shows two dimensional plots comparing GC, coverage and length for the final assembly. The coverage versus length plot reveals a set of scaffolds spreading up the Y axis that are relatively short but have high coverage. These are likely to include repeat regions and transposable elements. The coverage versus GC plot shows most scaffolds centred around 40 % GC as expected, but does reveal a slight grouping at a standard coverage but higher GC content. These could derive from artefacts of short fragments or could be bacterial DNA still present after filtering.

After filtering, the combined read data (Illumina and Roche) equated to an estimated coverage depth of around 38X (Table 2.4). Figure 3.8 shows that the distribution is skewed heavily due to large numbers of high coverage contigs/scaffolds most likely due to repeat

regions within the genome. The median value was closer to 20.

The resulting assembly was a mixture of contigs and scaffolds. The latter were formed during both the initial assembly stages by ABySS and CLCBio and during the transcriptome based scaffolding stage (Table 3.6).

Table 3.5: Genome assembly metrics

Assembly Stage	Number of contigs	Span (Mb)	N50 (bp)*	GC (%)
Individual assemblies				
Newbler	191,000	78	511	40.80
ABySS	868,000	440	572	41.04
CLCBio	1,070,000	590	652	40.36
Initial				
Merged with minimus2	424,000	526	1425	40.45
Intermediate				
Collapsed haploid contigs and filtered	352,000	430	1390	40.51
Final				
Scaffolded with transcripts	315,000	429	1589	40.46

* N50 is a weighted median statistic such that 50% of the assembly is composed of contigs and scaffolds larger than or equal to that value.

Table 3.6: Scaffolds vs Contigs

Type	Number	Span (Mb)	N50 (bases)	GC (%)
Contigs	289,789	334.6	1284	40.75
Scaffolds*	25,343	94.9	6248	39.46

* Any sequence with $\geq$ 10 consecutive N bases

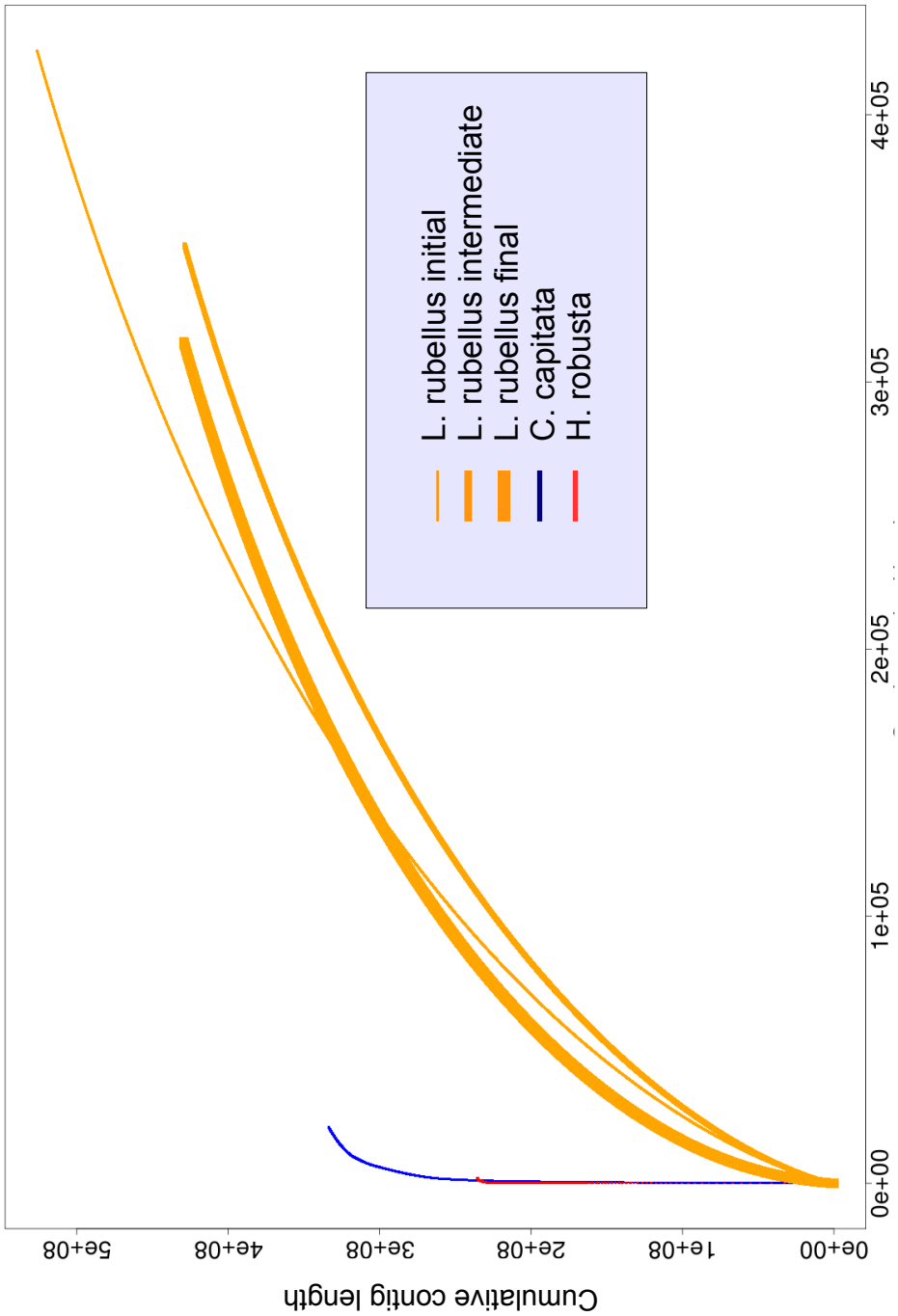

Figure 3.6: Cumulative curves for three annelid genomes

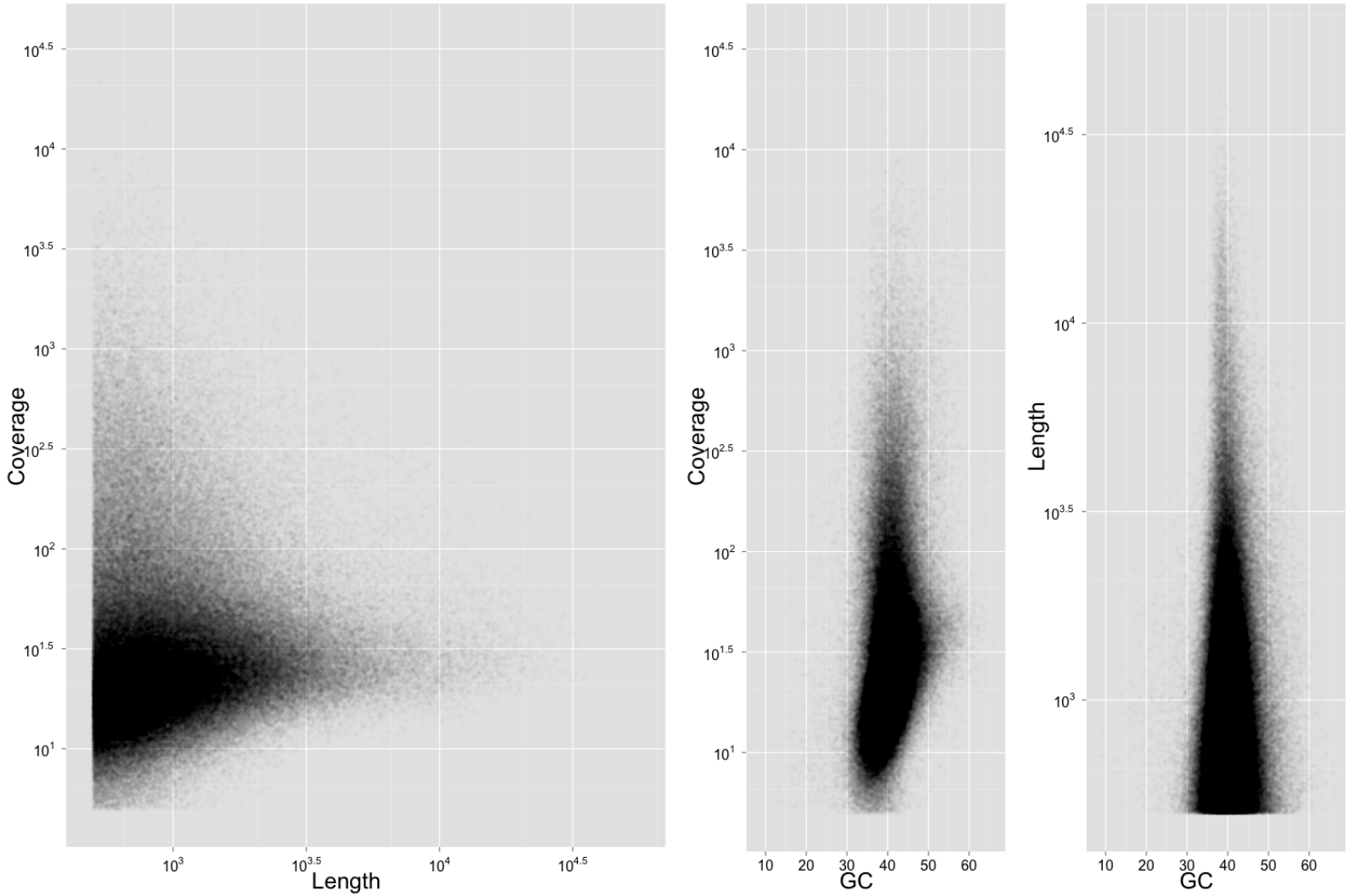

Figure 3.7: GC, coverage and length plots for the final *L. rubellus* genome

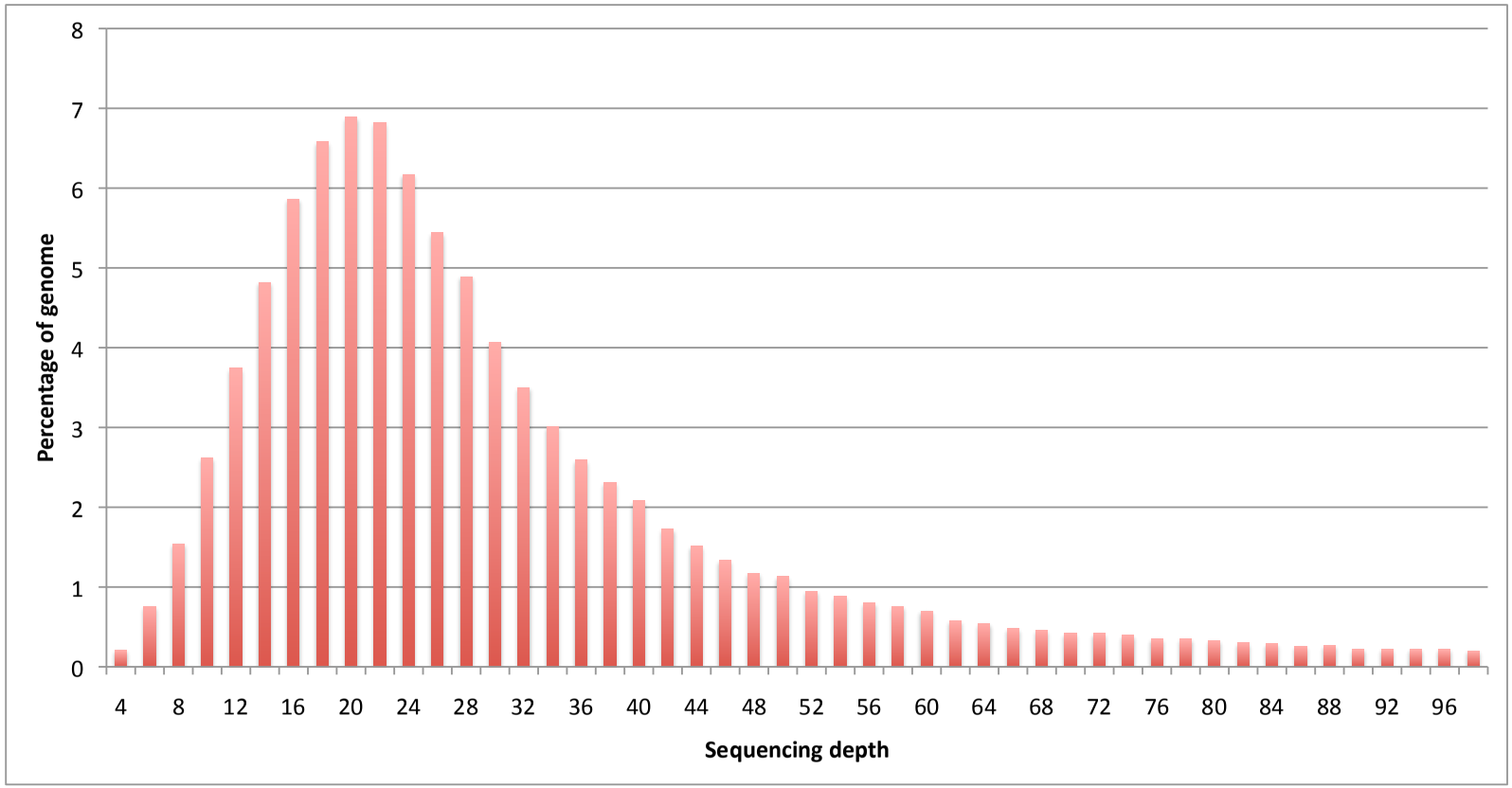

Figure 3.8: Distribution of sequencing depth for the final *L. rubellus* genome assembly.

3.4.2 Transcriptome

Table 3.7 lists the metrics from the transcriptome assemblies. It shows that clustering of the transcripts by UCLUST at 97% removed very little data whilst reducing the number of transcriptome fragments (transfrags) by half. The Lumbribase UniGene coverage of around 80% is expected as the EST data that created them was taken from numerous sources and possibly even numerous species of earthworm due to suspected misidentification, therefore one would not expect them all to be present.

Although designed for assessing the completeness of a genome assembly, the Core Eukaryotic Genes Mapping Approach (CEGMA) [129] can also be used to assess the completeness of a transcriptome as essentially it searches nucleotide sequences for matches to a set of core eukaryotic genes. CEGMA builds a highly reliable set of gene annotations in the absence of experimental data which can be used to both train gene finders and assess the completeness of a genome or transcriptome. This is achieved by searching for the presence of 458 core proteins, in particular a subset of 248 highly conserved core eukaryotic genes (CEGs), in a nucleotide FASTA file, e.g. a genome or transcriptome assembly. In addition to gene models and coding sequence files, a summary output file is given for both complete (70 % of the protein length) and partial (a pre-computed minimum for each gene) models across four protein groups ranging from least conserved (Group 1) to most conserved (Group 4). Metrics include #Prots (the number of CEGs identified), %Completeness (the percent of CEGs identified), #Total (the actual number of CEGs identified), Average (average number of orthologs per CEG, e.g. #Total / #Prots) and %Ortho (percentage of detected CEGs that have more than 1 ortholog).

The high CEGMA values for both complete and partial values vary little across the two transcript sets adding further weight to the quality of the transcriptome. The large number of transcripts is partly due to Trinity creating separate transcript contigs at very low base frequency differences in an aim to capture all low coverage transcripts. Many of the similar alternative transcripts collapsed in the clustering step, but there are still many with over 3% difference. Later mapping steps identify many of these as short contigs

which don't map to the genome and are likely to be incorrect assemblies.

Lastly, there is also the origin of the RNA to consider. The earthworms used originated from Holland and are likely to be a different strain or even species. Even a different strain of *L. rubellus* will show significant variation in the coding regions as demonstrated in the initial AFLP analysis during the selection of the worms (Section 2.1). A best case scenario would place the Dutch worms in clade B, but there could still be up to 2% divergence. A BLASTN of a lumbrokinase gene predicted from the genome (k_17598) against the transcriptome produced a top hit with 94% identity. This would suggest that the worms used for the transcriptome were significantly different to the worm used for genomic sequencing, perhaps accounting for some of the unmapped transcripts. This genetic difference, however, does not reduce the value gained from having the transcriptome for training the gene finders and scaffolding the genome.

Table 3.7: Transcriptome assembly metrics

Method	Number of transcripts	Span (Mb)	N50 (bases)	CEGMA complete (%)	CEGMA partial (%)	UniGene coverage (%) *
Trinity	319241	569	3519	97.98	99.19	80.4
Trinity 97†	163282	120	1257	97.98	98.79	80.0

† Trinity after UCLUST at 97%

* at least 70 % of UniGene present from a BLASTN [5] search

3.4.3 Assembly validation

A measure of the success of an assembly can be obtained from the 'completeness' of its representation of protein coding regions. A measure of completeness can be achieved in two ways:

Transcript-based validation

The Lumbribase UniGenes [126] and the Trinity assembled transcriptome were mapped to the genome using the gapped aligner BLAT [83]. Table 3.8 lists the alignment metrics for the three stages of genome assembly. For both UniGenes and transcriptome there is only a marginal drop in completeness suggesting that the two stages removed very little non-redundant data from the assembly. However, over 10% of the UniGenes and almost 20% of the transcripts were apparently missing from the genome. Perhaps the missing UniGenes are poor quality, comprised of low numbers of ESTs?

Calculations showed the average number of ESTs mapped to the missing UniGenes to be 6.4 compared to an average of 4.0. This number contradicted the hypothesis, but closer inspection revealed many of these to be mitochondrial with many EST members. Mitochondrial DNA is also more divergent than nuclear, therefore, the number of successful mappings would be further confounded by higher sequence divergence across similar strains. Removing sequences with a positive BLAST hit to the mitochondrion removed 171 sequences but decreased the number of ESTs per missing UniGene to 1.5. This low number would suggest that the remaining missing UniGenes are indeed due to poor quality ESTs or other contaminating sequences. It is also worth noting that only 80% of the UniGenes mapped to the transcriptome (3.7) adding further weight to their lack of value.

The missing 20% of the RNA-Seq transcriptome needs explanation. Firstly, these data derived from a different strain of *L. rubellus* (as previously discussed) that may include divergent individuals. Some of the missing 20% could derive from rapidly evolving genes in these very different genomes. The missing 20% were also shorter than the mean length of the whole RNA-Seq transcriptome (344 bp compared to 736 bp) suggesting that many

might in fact be short incorrect sequences generated by Trinity or contaminant DNA fragments.

Table 3.8: Genome completeness metrics

Assembly Stage	CEGMA Complete (%)	CEGMA Partial (%)	CEGs (%)*	UniGenes (%)**	Transcriptome (%)***
Initial	16	29	96.1	89.5	80.8
Intermediate (after collapsing haploid contigs)	16	29	95.6	88.7	80.0
Final (scaffolded with transcripts)	55	67	95.6	88.7	79.8

* calculated based on positive TBLASTN hits to the 458 CEGs used in CEGMA
** calculated based on positive BLAT hits to the 8178 lumbribase EST clusters
*** calculated based on positive BLAT hits to the transcriptome

Single copy core conserved gene-based validation

The CEGMA results (Table 3.8) suggested that the genome from the first two assembly stages was too fragmented. After scaffolding, however, the number of both complete and partial matches to CEGMA proteins rose significantly, confirming that these genes were in fact present but likely fragmented on separate contigs/scaffolds. The proportion of the CEG sequences mapped confirmed this further as the number remained constant. The complete CEGMA report for the scaffolded contigs suggested that the data still contains some redundancy in the form of diploid contigs, as the average number per group was slightly greater than one in all groups (Table 3.9). This is also the case for the two JGI annelid genomes, highlighting the difficulty in producing a genome that represents a haploid set of chromosomes. The CEGMA result for *L. rubellus* was poor as both the other annelids have near complete results for both partial and complete matches (Figure 3.9). There is, however, a dip in Group 2 for the two JGI annelids, which is unusual as the groups of core genes should become more readily observed from left to right.

Table 3.9: CEGMA completeness reports for *L. rubellus* and the two other annelid genomes

Species	Group	#Prots	%Completeness	#Total	Average	%Ortho
L. rubellus	Complete	137	55.24	149	1.09	8.76
	Group 1	29	43.94	30	1.03	3.45
	Group 2	32	57.14	35	1.09	9.38
	Group 3	37	60.66	41	1.11	10.81
	Group 4	39	60.00	43	1.10	10.26
	Partial	167	67.34	189	1.13	10.78
	Group 1	37	56.06	38	1.03	2.70
	Group 2	38	67.86	42	1.11	10.53
	Group 3	46	75.41	59	1.28	19.57
	Group 4	46	70.77	50	1.09	8.70
C. telata	Complete	234	94.35	263	1.12	10.68
	Group 1	62	93.94	72	1.16	14.52
	Group 2	51	91.07	58	1.14	11.76
	Group 3	59	96.72	63	1.07	5.08
	Group 4	62	95.38	70	1.13	11.29
	Partial	239	96.37	286	1.20	17.15
	Group 1	64	96.97	78	1.22	17.19
	Group 2	52	92.86	61	1.17	15.38
	Group 3	60	98.36	70	1.17	15.00
	Group 4	63	96.92	77	1.22	20.63
H. robusta	Complete	229	92.34	256	1.12	9.61
	Group 1	60	90.91	64	1.07	6.67
	Group 2	48	85.71	52	1.08	6.25
	Group 3	59	96.72	65	1.10	10.17
	Group 4	62	95.38	75	1.21	14.52
	Partial	231	93.15	259	1.12	9.52
	Group 1	61	92.42	65	1.07	6.56
	Group 2	48	85.71	52	1.08	6.25
	Group 3	59	96.72	65	1.10	10.17
	Group 4	63	96.92	77	1.22	14.29

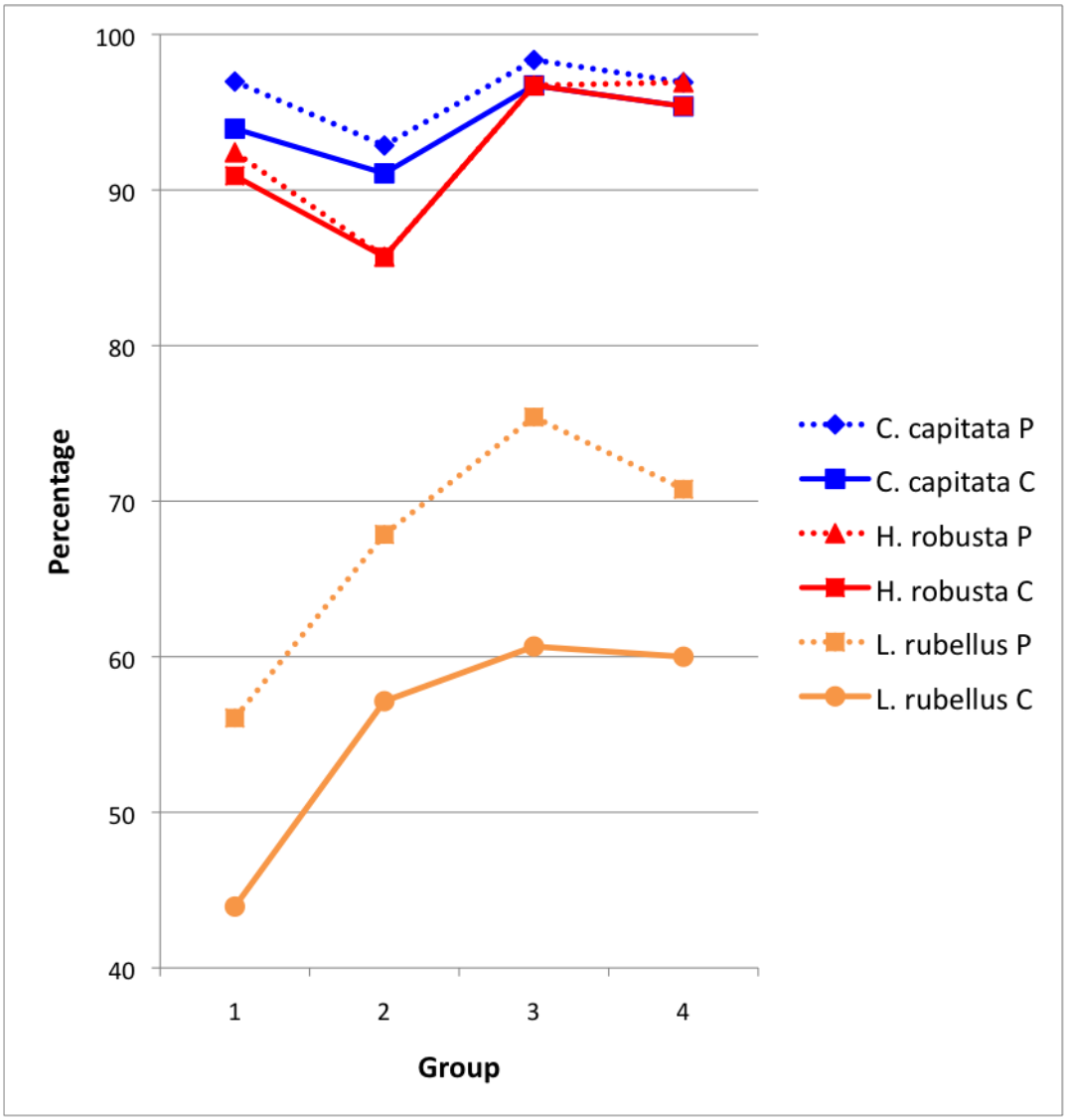

Figure 3.9: CEGMA summary results for *L. rubellus* and the two other annelid genomes

P and C represent the partial and complete results. Group is a CEGMA quantifier whereby the core genes are placed in groups which become more conserved as the group number increases.

3.5 Conclusions and discussion

The conclusion of these assembly, quality control and validation checks was that the genome is indeed a high quality draft [24]. The span of contigs and scaffolds was roughly correct (estimated 420 Mb actual 429 Mb), given that repeat regions may have collapsed and some amount of allelic variation may still exist. The protein-coding sequence completeness was good but not outstanding. The statistics are comparable to that of the panda genome [95] before it was scaffolded with mate-pairs. The initial panda assembly was generated using short read data with an average read length of 52 bases at 39X coverage and formed contigs with an N50 length of 1.5 kb, compared to this project (47 bases, 55X and 1.59 Kb respectively). Post scaffolding, the N50 of the panda genome rose to 1.3 Mb, a huge improvement and a positive example of how an assembly can be improved with additional data.

3.5.1 Assembly improvements

Aside from the improvements in read length and quality that have continued since the original sequencing data was generated in 2008, there are other measures that could be taken to improve the assembly.

Pre-assembly screening

Although all mitochondrial reads were removed from the assembly, this approach can be taken a step further by using GC content and coverage in general. Preliminary assemblies can be assessed to highlight contigs that are not from the organism of interest based on GC content and expected coverage. These contigs can then be used to screen the sequence data prior to in-depth assemblies, ensuring that the assembly proceeds with a cleaner set of reads. Although very little bacterial data was detected in the filtering step it would still help to remove this data in the preliminary assemblies.

Read correction

Some assemblers such as SOAPdenovo [96] already implement a read correction phase, prior to assembly, whereby all reads are subjected to some form of error correction. Other software exists for performing this step prior to assembly [81][180][142] and report marked improvements in assembly after this step [179]. This would reduce the variability in the data which may reduce memory requirements sufficiently to make more assembly tools available.

Additional read distance information

Possibly the most important of the improvements would be the addition of information that links reads across large distances such as large insert mate-pairs which would improve an assembly by bridging the gaps in an assembly caused by repeat regions. Recently the Pacific Biosciences single molecule sequencing technology (www.pacificbiosciences.com) has developed a 'strobing' method, in which the sequencing reaction is turned on and off at known intervals over a large length of DNA producing many linked reads. This approach would provide invaluable information for assembly yet it remains to be seen if it proves to be successful as the error rate is currently very high at around 15%. Alternatively, another next generation sequencing technology, Oxford Nanopore, promises much longer reads (http://www.nanoporetech.com/). If this happens and read sizes increase sufficiently, then the older overlap-layout-consensus assembly approach may return and utilise new assemblers such as the String Graph Assembler from Simpson and Durbin [146].

Chapter 4

Annotation

4.1 Introduction

Genome annotation is the process of identifying regions of interest and attempting to assign them a name or function. Regions of interest include protein-coding genes, RNA genes and repetitive elements. Subsequent to the definition of protein-coding genes, these can be additionally annotated with functional information such as domain content, similarity to other proteins, enzyme classification, etc. There are two computational approaches which are not mutually exclusive: (a) identify from scratch (*ab initio*) and (b) evidence-based. The former uses complex algorithms often incorporating external information via the training of search algorithms, while the latter relies on evidence from both previous annotations and biological information such as mRNA. As more genomes are sequenced and more annotations become available this evidence based approach becomes more powerful as the chances of significant homology increases. Ideally all annotations would be verified manually. However, this is impossible for all but the smallest of genomes due to human resource limitations.

4.1.1 Repetitive Elements

Identifying the repetitive sequence content of a genome is often the first step in annotation. Once identified, these regions can be masked, preventing subsequent annotation searches from analysing these areas. A repetitive element (RE) is by definition any element that is present more than once (see Section 1.6.3). Again, finding them can be achieved with or without evidence. Without evidence involves identifying REs within the genome by reciprocal sequence analysis. However, this can lead to the identification of false positives, e.g. common genes/domains. Evidence based methods use databases of known REs, which derive predominantly from model organisms. Perhaps because of the ambiguity of identifying REs there is a plethora of software for the task [91] and each genome project appears to use a slightly different combination of methods.

Almost all methods use data from the main database for REs, RepBase Update [79]. This is also used by the most popular repeat finding software, RepeatMasker [152]. However, unless working on an organism which already has a well documented set of REs, then the best approach may be to combine both *ab initio* and evidence based methods. Currently, there are two approaches to this.

Creating a custom library for RepeatMasker

RepeatMasker uses a modified version of the RepBase Update library. The standard format is a modified FASTA format including a unique name for the sequence followed by a standardised ID, with the sequence on a new line, for example:

```
>unique_name #LINE/RTE-BovB
ACGTACGTACGTACGTACGTACGT
```

A custom database in this format can be generated and appended to the RepBase Update library. Novel REs can be identified in many ways, for example using RECON [11], RepeatScout [132], Censor [80] and PILER [45]. Annotating these to a specific TE group can be done manually using BLAST or with specific software such as TEclass [2]. To

avoid over annotating sequence as being repetitive, and thus masked, a check of the custom library against a protein database or gene set is advisable.

All in one Pipelines

Software tools have been written that automate the process of generating a new repeat library and then using this to annotate the repetitive content of a genome. The most popular of these is RepeatModeler [151] which first runs RECON, RepeatScout and Tandem Repeats Finder [16], filters and classifies the results, then runs RepeatMasker.

REPET [54] is another complete package which aims to identify and classify repeats within a genome. The first stage (TEdenovo) runs a reciprocal genome comparison using BLASTER, it then clusters common elements using RECON, GROUPER and PILER building a multiple alignment for each element. A consensus sequence is then derived from each cluster, which is then classified according to known RE features. Finally redundancy within the repeat set is removed. The second stage (TEannot) uses this repeat library or any other to mine the genome for all copies of the elements.

4.1.2 Non-coding RNA

A non-coding RNA (ncRNA) is a functional RNA molecule that is not translated into a protein. Examples include transfer RNA (tRNA), ribosomal RNA (rRNA), small nuclear ribonucleic acid (snRNA) and microRNA (miRNA).

4.1.3 Protein-coding gene prediction

Often the major interest in a genome is the encoded proteome. Yet, despite many years of research, revealing this is still a major task, and probably the next bottleneck once the current issues surrounding genome assembly have been resolved. For this reason, there is an abundance of gene finding algorithms, 18 of which are discussed in detail by Picardi *et al* [131]. Many of these were investigated, but only those methods that were used in this project are discussed here. These were chosen based on their previous application and

functionality. As with finding repeats, there are two standard techniques, with and without direct evidence.

ab initio gene finders

Ab initio gene finders provide a quick and often accurate set of gene predictions. The standard format of a gene is shown in Figure 4.1. Conserved features form the basis for *ab initio* gene finding: a signal to start, splice sites to signal exon/intron boundaries and a signal to stop. UTRs are much harder to predict as they show high variation in both length and sequence content, but can be inferred from training.

However, these markers alone are insufficient to identify genes efficiently or accurately, as codon bias and splice signals vary widely across the animal kingdom. In addition to donor 5' and acceptor 3' splice sites, introns also contain a branch site toward the 3' end. However, these three signals have evolved dramatically [144] and do not always conform to the GT-branch-AG rule [21]. To combat this, extra evidence is needed, often obtained by training gene finders with genes or transcripts, ideally from the same species. The data used for training can be obtained from the output of a naive application of a gene predictor or from a secondary gene prediction method such as CEGMA, which creates a set of gene predictions based on core eukaryotic genes.

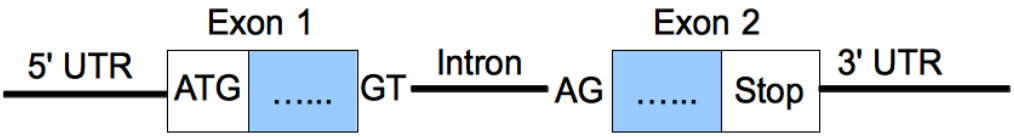

Figure 4.1: Standard gene format

UTR is untranslated region, stop is any of the three stop codons, TAG, TAA or TGA and GT and AG are the standard splice sites which occur at the start and end of introns, the number of which varies widely from species to species and gene to gene. The dots in the blue boxes represent exon sequence.

SNAP (Semi-HMM-based Nucleic Acid Parser) [87]

SNAP has been successfully used on many genomes [93][41][22][127]. Essentially it is a hidden Markov Model (HMM) based predictor which benefits from extensive and flexible training, making it excellent at finding genes in novel genomes. Training occurs via a simple set of steps which use a starting set of gene predictions (available from CEGMA analysis) to generate a HMM. A HMM in this sense is a model which attempts to identify patterns of bases which match a predefined set of states. SNAP can be trained in a boot-strapping fashion by running once, training itself on the output, running again and so on.

AUGUSTUS [157]

AUGUSTUS, like SNAP, uses HMMs and is designed to identify protein-coding genes in eukaryotes, but unlike SNAP it can also predict 5'UTR and 3'UTR regions. It too has been used in many large genome projects [135] [58] [121] and benefits from extensive training to generate either a hints file or a species specific HMM.

Evidence based gene finders

The most widely implemented evidence based finder of protein-coding regions is BLAST [5]. Although not designed as a gene finder, it is the first step in many evidence based gene annotation pipelines, using one of the many protein databases such as the curated databases at UniProt [9]. These same databases can be used in a slightly more involved way to try and identify complete gene structures in new genomes. Exonerate [150] uses an exhaustive search method to align sequences as opposed to the quicker heuristic alternative used by BLAST. This makes it slower but more accurate at identifying a complete alignment. Part of the exonerate package is protein2genome, which, as the name suggests, can be used to align protein sequences to genomic data whilst abiding by the general rules of gene structure, i,e, start, stop and splice sites. If, however, the genome to be annotated is large, and the protein to be aligned is also large, the computational requirements of Exonerate/protein2genome are immense. To minimise this, regions of interest can be pre-

defined by preforming a BLASTX search of the genome using the protein set, identifying the regions that generate an alignment and then using those regions (with some extensions either side of the BLAST results) for the protein2genome alignment.

This method of creating gene predictions, based solely on protein similarity, is not without its risks and had been deprecated as an option from the annotation pipeline MAKER2 [70] when annotating eukaryotes until very recently (version 2.22). However, in the case of *L. rubellus*, perhaps due to the fragmented genome causing issues with the traditional methods, preliminary annotations identified numerous occasions whereby 'real' genes were being identified by protein2genome and missed by MAKER2 (Section 4.2.3).

Combined pipelines for gene finding

Large scale, integrated annotation pipelines for eukaryote genomes have been devised [32] and often research groups create their own based on in-house best practice and algorithms/models designed for the species they work on. The GMOD (Generic Model Organism Database) project (`http://gmod.org`) aims to produce a suite of tools to enable the generation of a database and genome browser for any organism. Part of this suite is the MAKER2 package [70], a genome annotation pipeline that is highly customisable and easy to configure. It works by combining *ab initio* methods with evidence from other sources such as ESTs, transcripts and protein alignments. Figure 4.2 shows an overview of the pipeline. The first stage is the annotation and masking of repeats. The second stage gathers the first round of information from *ab initio* algorithms such as SNAP [87]. Alignments are then generated against three databases: transcripts from the same or a very closely related species (using BLASTN), alternative transcripts form more distantly related species (using TBLASTX) and a set of proteins (using BLASTX). This alignment information is then refined using exonerate est2genome and protein2genome, and then all alignment information is fed back to the *ab initio* predictors as new parameter estimations. This set of gene predictions along with all the other information is then used to produce a final set of gene predictions, in both GFF3 format and nucleotide and peptide sequences.

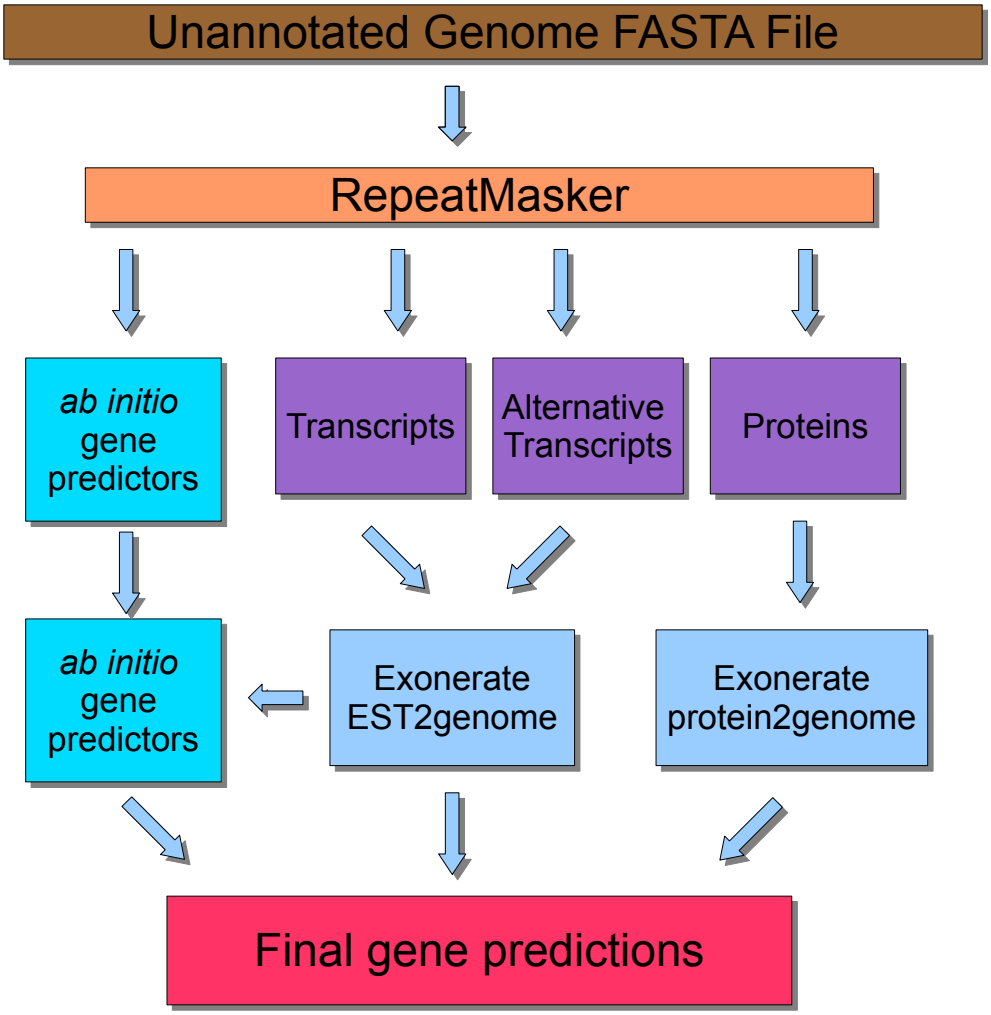

Figure 4.2: MAKER2 pipeline

A FASTA file is masked for repeats using RepeatMasker. *Ab initio* predictors are trained with the alignment information of up to three sources: transcripts from the same or a very closely related species (using BLASTN), alternative transcripts form more distantly related species (using TBLASTX) and a set of proteins (using BLASTX). Exonerate polishes the predictions and a final set of gene predictions is produced.

A GFF3 (General Feature Format v3) file is a standard text file with nine tab delimited columns (Table 4.1) and was developed as part of the "Sequence Ontology Project (SO) [48], replacing the GFF2 file format developed at the Sanger institute (www.sanger.ac.uk). An example of a *L. rubellus* GFF3 file can be seen in Figure 4.3. GFF3 is a standard format and widely used across the genomic communities (although confusing variations exist, e.g. SNAP and EVidenceModeler [67] have their own variants). The main advantage of GFF3 over its predecessors is the mechanism for representing hierarchical information groups. Most commonly, this is applied to a gene object, with a three tier system. The gene attribute is the parent to mRNA which is then parent to exon and CDS.

Table 4.1: GFF3 format

Column	Name	Description
1	seqid	The ID of the sequence object, in this case the contig or scaffold ID
2	source	A free text qualifier used to distinguish the source of the feature. Figure 4.3 shows how this is used to identify the different gene predictors (AUGUSTUS, MAKER, protein2genome), and the separate stages of MAKER2 (BLASTN, TBLASTX, BLASTX, est2genome, protein2genome, RepeatMasker and SNAP)
3	type	This is a fixed named that must adhere to a term from the Sequence Ontology database [48], e.g. gene, mRNA, exon, CDS, match, match_part
4	start	The start base of the feature
5	end	The end base of the feature
6	score	A floating point number for the score of the feature
7	strand	The strand of the feature relative to the sequence object, '+' for positive and '-' for minus strand
8	phase	Only for CDS features, the phase indicates where the feature begins with reference to the reading frame, "0", "1" or "2" bases
9	attributes	Accepts multiple tag=value pairs separate by semicolons used to provide information for the feature, e.g. ID, Name, Parent, Target

```
##gff-version 3
scaffold_m4078  .       contig  1       33158   .       .       .       ID=scaffold_m4078;Name=scaffold_m4078;
scaffold_m4078  p2g     gene    946     32158   .       +       .       ID=p_21141;Name=p_21141:Q9ULT8;
scaffold_m4078  p2g     mRNA    946     32158   .       +       .       ID=p_21141-RF;Name=p_21141:Q9ULT8;Parent=p_21141;
scaffold_m4078  p2g     exon    946     1076    .       +       .       ID=p_21141:946-1076;Name=p_21141:Q9ULT8;Parent=p_21141-RF;
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
scaffold_m4078  AUGUSTUS        gene    776     8690    0.02    +       .       ID=g63
scaffold_m4078  AUGUSTUS        mRNA    776     8690    0.02    +       .       ID=g63.t1;Parent=g63
scaffold_m4078  AUGUSTUS        exon    776     1076    .       +       .       Parent=g63.t1
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
scaffold_m4078  maker   gene    553     27797   .       +       .       ID=maker-scaffold_m4078-snap-gene-0.2;Name=maker-scaffold_m4078-snap-gene-0.2;
scaffold_m4078  maker   mRNA    553     27797   .       +       .       ID=maker-scaffold_m4078-snap-gene-0.2-mRNA-1;Parent=maker-scaffold_m4078-
scaffold_m4078  maker   exon    553     1076    42.247  +       .       ID=maker-scaffold_m4078-snap-gene-0.2-mRNA-1:exon:500;Parent=maker-scaf
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
scaffold_m4078  repeatmasker    match   4911    4950    206     +       .       ID=scaffold_m4078:hit:1172;Name=species:%28TTAGGG%29n%20genu
scaffold_m4078  repeatmasker    match_part      4911    4950    206     +       .       ID=scaffold_m4078:hsp:3603;Parent=scaffold_m4078:hit:1172;Nam
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
scaffold_m4078  blastn  expressed_sequence_match        32971   33066   44      +       .       ID=scaffold_m4078:hit:1182;Name=comp286_c2_seq16_len
scaffold_m4078  blastn  match_part      32971   33066   44      +       .       ID=scaffold_m4078:hsp:3613;Parent=scaffold_m4078:hit:1182;Name=comp286_
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
scaffold_m4078  tblastx translated_nucleotide_match     30636   32095   107     +       .       ID=scaffold_m4078:hit:1198;Name=gi|161117513|gb|EY48458
scaffold_m4078  tblastx match_part      30636   30704   107     +       .       ID=scaffold_m4078:hsp:3669;Parent=scaffold_m4078:hit:1198;Name=gi|1611175
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
scaffold_m4078  blastx  protein_match    946     32158   219     +       .       ID=scaffold_m4078:hit:1199;Name=Q9ULT8;
scaffold_m4078  blastx  match_part      946     1074    219     +       .       ID=scaffold_m4078:hsp:3674;Parent=scaffold_m4078:hit:1199;Name
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
scaffold_m4078  est2genome      expressed_sequence_match        32939   33113   476     -       .       ID=scaffold_m4078:hit:1200;Name=comp286_c2_seq
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
scaffold_m4078  protein2genome  protein_match   5160    13246   2035    +       .       ID=scaffold_m4078:hit:1217;Name=Q9ULT8;
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
scaffold_m4078  snap_masked     match   26378   32161   147.483 +       .       ID=scaffold_m4078:hit:1218;Name=snap_masked-scaffold_m4078-
scaffold_m4078  snap_masked     match_part      26378   26389   4.369   +       .       ID=scaffold_m4078:hsp:3759;Parent=scaffold_m4078:hit:1218;
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
##FASTA
>scaffold_m4078
TATACACACACAGACACAGATAGACAGATGTATAGGCAGACAATAAGAAGAAGACGAAGC
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
```

Figure 4.3: An example section of a GFF3 file

Dashed lines represent where the data has been cut for ease of viewing

4.2 Chosen annotation methods

4.2.1 Repeat finding method adopted

After trying many of the different software available REPET was chosen for two reasons. Firstly, of the two pipeline options this was the most successful in running, and secondly, it is also developed to run on Sun Grid Engine (SGE: http://www.oracle.com/us/sun/index.htm), meaning it could be parallelised and run on the ECDF grid. This latter point is always an advantage when working on large data as large amounts of RAM (over 32 GB) and large numbers of processors on a single desktop machine are still expensive. Unfortunately, the REPET pipeline failed during the TEannot stage, due to the huge numbers of tasks that were required causing memory issues on the ECDF grid. However, the library that was created was transformed into a RepeatMasker library as above. Many of the sequences within the REPET library were unclassified, so were compared to RepBase Update version 20110419, and significant hits were used to assign sequences in the library to RepBase RE classes. Redundancies between this set and the RepBase Update library were removed and the two were combined and run against the *L. rubellus* genome with RepeatMasker.

4.2.2 ncRNA finding method adopted

RNA families were identified by comparison to version 10.1 of the RFam database [64][65][55][36] using the script rfam_scan.pl (ftp://ftp.sanger.ac.uk/pub/databases/Rfam/tools/rfam_scan.pl).

4.2.3 Gene finding method adopted

For each CEG identified in a genome, CEGMA produces a gene model and GFF file. This GFF file can be used as the training set and can be converted into the two files required for the snap parameter estimation steps using cegma2zff (part of the MAKER2 package) to produce a snap specific ZFF file. Following a simple five step procedure, the ZFF files

are used to create the HMM.

Initial predictions using MAKER2 gave reasonable numbers of genes. However, overlap with exonerate and AUGUSTUS predictions was not encouraging (see Section 4.2.4) suggesting either that AUGUSTUS and/or exonerate predictions were incorrect, or were correct and should be included in the gene set. Until version 2.22 released in January 2012, MAKER2 had excluded using the protein2genome output directly for gene prediction in eukaryotes. The fact that this has been included in the latest version adds weight to its inclusion in this project. For this reason, the three methods were run independently and the output from AUGUSTUS and protein2genome were parsed and added to the MAKER2 produced GFF3 files to generate one GFF3 for each contig/scaffold, e.g. figure 4.3.

4.2.4 Functional annotation method adopted

The peptide sequences produced from each of the three gene prediction sets were annotated with functional information. Annot8r [140], BLAST2GO [31][30], InterProScan [133] and DETECT [75] were used to assign Gene Ontology (GO), Kyoto Encyclopaedia of Genes and Genomes object identifiers (KEGG object IDs), Enzyme Commission IDs (EC) and domain annotations. Alignment to Lumbribase UniGenes, Annelid ESTs, *H. robusta* proteins, *C. telata* proteins and the SwissProt and NCBI nr protein databases were also performed using BLAST. As many of the predictions were short, all gene predictions with a coding sequence less than 50 amino acids were removed.

4.2.5 Manual annotation

Despite major improvements in genome annotation, the best method is still to annotate manually. Computational methods are excellent and constantly improving but nothing beats an expert in a particular field manually editing a set of genes. With this in mind, an annotation workshop was held in July 2009 where the project collaborators came together to discuss best practices and begin to manually annotate their gene families/areas of interest. GBrowse2 is excellent for visualising a genome, but has no option to edit

the annotations on a particular scaffold. The two genome editing softwares considered suitable for this task were Artemis [137] and Apollo [92], both of which have been used by many genome projects and can work directly on GFF3 files. As Apollo is part of the GMOD project, the continuity between the GFF3 file and the software meant that it was more logical to work with Apollo (Figure 4.4).

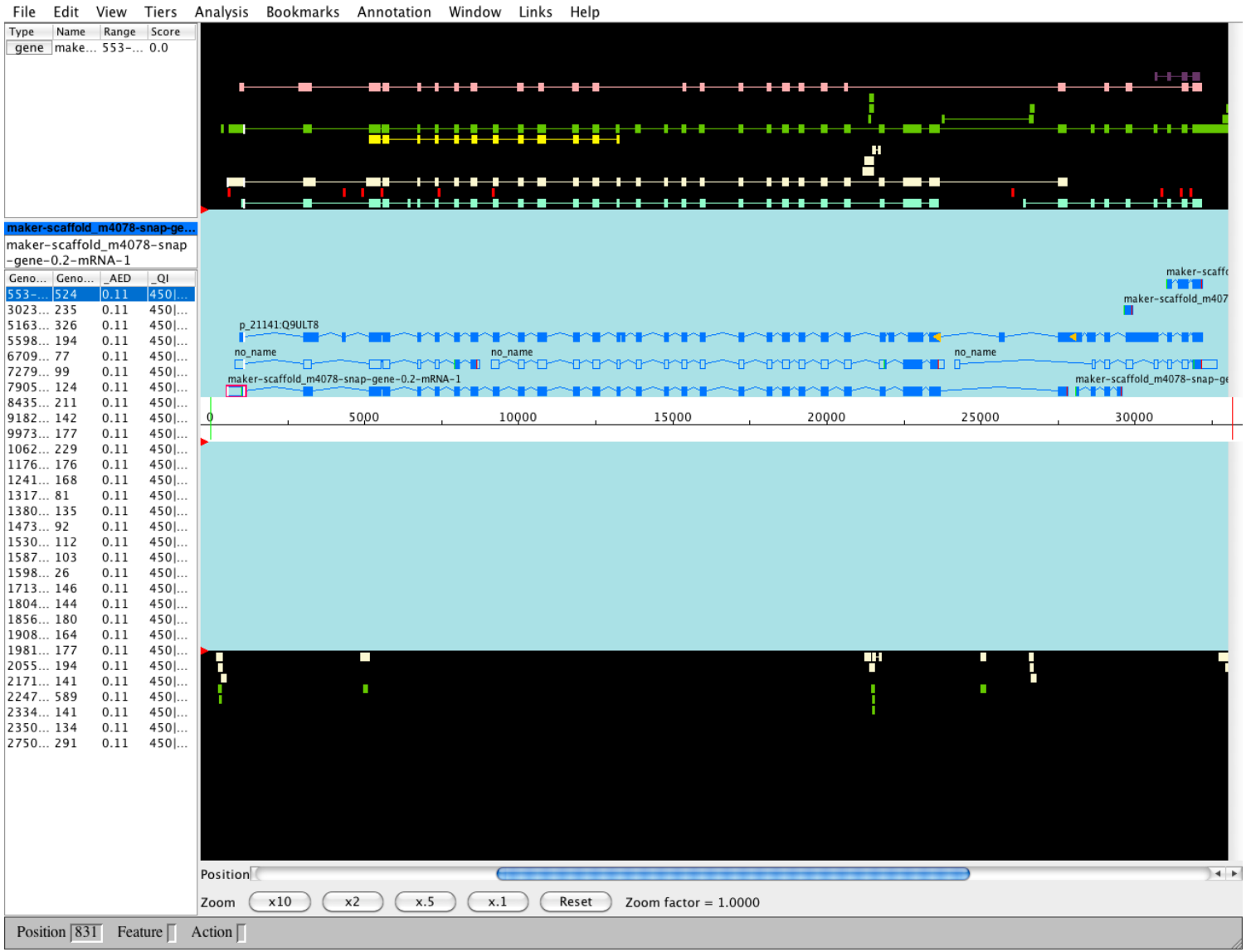

Figure 4.4: Screenshot of scaffold_m4078 loaded in Apollo

Custom '.tiers' and '.style' files match up to the source attributes from the three gene prediction sources. Features in the black area above the central contig/scaffold section are those features on the positive strand and those in the lower black section are on the minus strand. The features in the central section are gene features. These can also be created from the features in the black sections by a simple drag and drop process where they can be refined using the exon editor. Right click on a gene object opens the 'Annotation info editor' window which is used to add extra annotation information when manually annotating.

4.2.6 Databases, wiki and website

The GMOD project provides a second database schema called Chado, [118] an extensive set of tables containing ontology information that is designed to be interoperable with Chado-compliant databases. This schema can also hold data from GFF3 files and can be linked to a GBrowse2 system, albeit with much slower response times in web browsing. To avoid this issue, a Chado database can be converted to a GBrowse2-friendly SeqFeature database via regular data dumps. Chado is thus used as a master database and SeqFeature as a slave. The other attraction of this schema is its ability to link directly to Apollo on a remote database, Annotation can theoretically be performed anywhere and fed directly back to the database. The problem with this method is the lack of versioning, as any changes made are permanent, although a solution appears to be in development (http://gmod.org/wiki/WebApollo). There is also a steep learning curve in understanding the complex interactions within the Chado database.

A solution to this problem was developed in the lead up to the annotation workshop. The GFF3 file for each contig/scaffold was made available for download on the contig_details.php page for that particular contig. This could then be uploaded into a local instance of Apollo and edited. The annotation information was then added in the 'Annotation info editor' using the following agreed formats (Table 4.2).

Table 4.2: Apollo manual annotation notes

Apollo annotation info editor field	Contents
ID value/DB Name	Use orthologues from *C. elegans* Wormbase (www.wormbase.org), *D. melanogaster* FlyBase (www.flybase.org) or *M. musculus* MGI (www.informatics.jax.org/).
Comments and Properties	#NAME your name #BLAST blast similarity info underpinning identification: blast database, top hit ID, evalue and score #GO GO terms if relevant and additional #EC EC terms if relevant and additional #KEGG KEGG terms if relevant and additional #DOMAIN Pfam or other domain presence (give domain ID from InterPro, and coordinates of domain match #FILE the name of a file of data relevant to this annotation (e.g. an alignment file, or a tree file, or an informative BLAST report) #TEXT other comments that do not fit in to the above #PUB PMID:number

The custom information format added to each manual annotation within the Apollo annotation info editor. Notes are appended to the 'attributes' field of the GFF3 file.

Each new model created a new feature in the GFF3 file with a source attribute of '.' and the information added in the 'Annotation info editor' was added to the final 'attributes' column as text. This new GFF3 file was then uploaded to the database using an upload section, which used code within the PHP page gff_uploader.php to first remove the previous data from the SeqFeature database before adding the new GFF file. This was necessary as the SeqFeature database has no primary key for contig/scaffold ID therefore duplicate data can exist. The gff_uploader.php script also launches an rdiff-backup command (http://www.nongnu.org/rdiff-backup/index.html) which created a versioned backup of the old data whilst adding the new annotations. This process allowed instantaneous updates of the SeqFeature database, such that any changes were immediately visible in GBrowse2 rather than having to wait for the next data dump. This was a useful feature which reduces the chances of two people simultaneously editing the same gene. It also allowed older versions to be accessed by the administrator in case of an incorrect annotation being added to the database.

Two structured query language (SQL) databases were built to hold the annotation data: a custom built PostgreSQL (Figure 4.5) and a MySQL SeqFeature database (part of the BioPerl package [156]. The PostgreSQL database contained data describing the assembly fragments, the annotations and information from external databases for InterProScan, Enzyme Commission and KEGG pathway information. A brief description of each table is given in Appendix A.

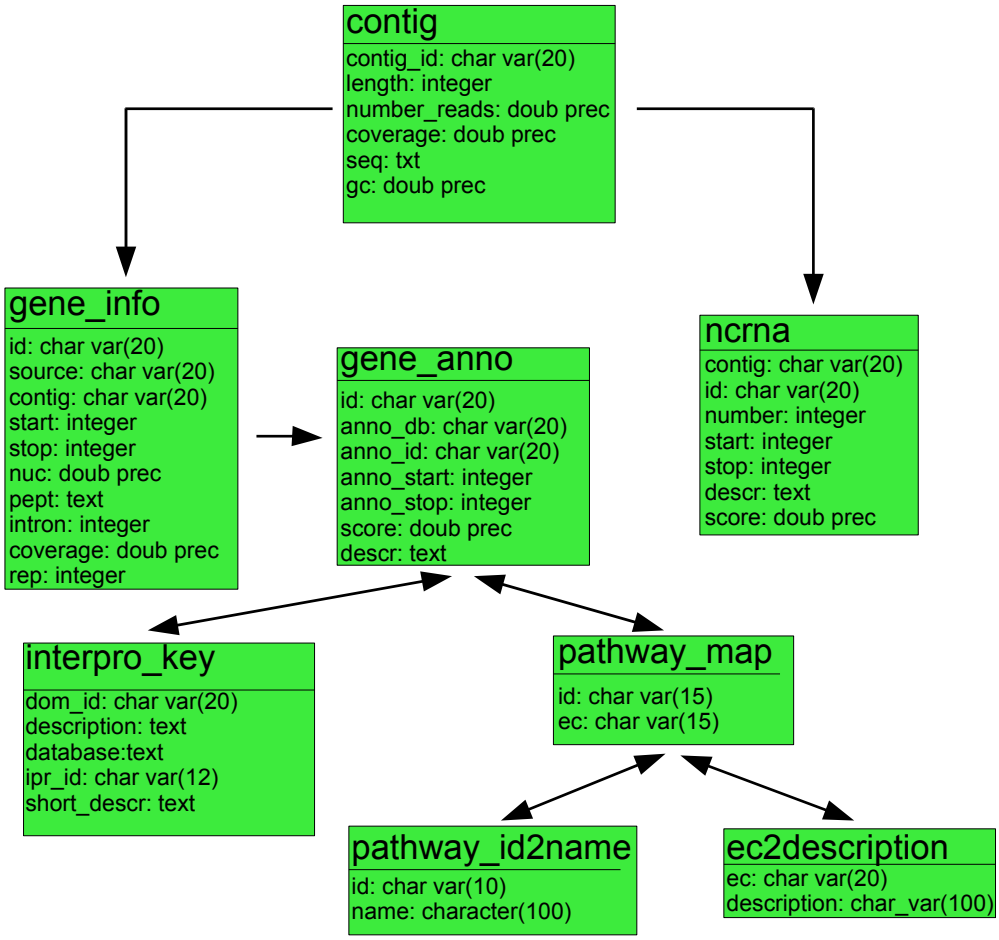

Figure 4.5: PostgreSQL genome database

The PostgreSQL database was linked to a custom built PHP:Hypertext Preprocessor (PHP) and Common Gateway Interference (CGI) front-end (Figure 4.11) through which collaborators and interested parties could browse the genome and its annotations as they were developed. A community portal for the *L. rubellus* genome project was constructed (see www.earthworms.org; Figure 4.12). The site is controlled with a username/-password login (which will be removed after publication) and a priority system whereby only certain aspects of the site are open to collaborators.

The genome sequence and annotations were exposed for browsing using the MySQL database and GBrowse2 [159]. This database stored the data from the GFF3 files in ten tables that link to the genome browser. GBrowse2 is highly customisable with a large community of users and active development.

4.3 Results

The selected annotation process and how it fits in to the overall assembly/annotation pipeline is shown in Figure 3.5. Due to the fragmented nature of the genome, even after scaffolding with the transcriptome, it was assumed that some genes would be split over multiple fragments making the process of gene finding more complex.

4.3.1 Repeat finding

Figure 4.6 shows the summary output from RepeatMasker. Even though the percentage of bases masked (26.43) was a reasonable value, the majority of the repeats were unclassified. This suggests that either the repeats are real and novel to *L. rubellus* or that they are false positives. The latter is plausible due to the nature of the genome assembly method. The de Bruijn graph structure used in the assembly may have generated nodes for REs that without significant variation will have collapsed similar regions in to the same node. This effect was anticipated from the beginning of the project and, as such, it was always assumed that the final set of REs would be underrepresented.

Preliminary investigations with MAKER2 were attempted using the custom repeat libraries. However, these produced more fragmented gene models than using the default 'Annelida' subset of the RepBase Update library.

```
file name: contigs.fa
sequences:        315201
total length:  429562379 bp  (428259863 bp excl N/X-runs)
GC level:         40.63 %
bases masked:  113513641 bp ( 26.43 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
SINEs:             1512       185620 bp    0.04 %
      ALUs            0            0 bp    0.00 %
      MIRs            0            0 bp    0.00 %

LINEs:            86066     10530405 bp    2.45 %
      LINE1           8         4336 bp    0.00 %
      LINE2       18556      2727093 bp    0.63 %
      L3/CR1       2032       349632 bp    0.08 %

LTR elements:     16576      1865288 bp    0.43 %
      ERVL          183        16643 bp    0.00 %
      ERVL-MaLRs      0            0 bp    0.00 %
      ERV_classI      0            0 bp    0.00 %
      ERV_classII   525        57709 bp    0.01 %

DNA elements:     45166      5193264 bp    1.21 %
     hAT-Charlie    687        65055 bp    0.02 %
     TcMar-Tigger     0            0 bp    0.00 %

Unclassified:    714624     83127448 bp   19.35 %

Total interspersed repeats:100902025 bp   23.49 %


Small RNA:           48        10275 bp    0.00 %

Satellites:        4106       519221 bp    0.12 %
Simple repeats:  181892     10888901 bp    2.53 %
Low complexity:   76657      4525913 bp    1.05 %
==================================================

* most repeats fragmented by insertions or deletions
  have been counted as one element


The query species was assumed to be homo
RepeatMasker version open-3.3.0 , default mode

run with blastp version 2.0MP-WashU [04-May-2006] [linux24-x64-I32LPF64 2006-05-11T11:17:04]
The query was compared to classified sequences in "L_rub_custom_repeat.lib"
RepBase Update 20110419, RM database version 20110419
```

Figure 4.6: RepeatMasker output

4.3.2 ncRNA finding

Noncoding RNAs identified in *L. rubellus* are shown in Tables 4.3 and 4.4. Of immediate interest were the 51 matches to the Clostridiales-1 family, an RNA structure of unknown function normally present in bacteria from the order Clostridiales, species of which have been documented in the guts of animals. This suggests either that there was still a large amount of bacterial sequence in the dataset, or, in fact, that the earthworm has acquired this RNA sequence. One of the highest scoring occurrences of this RNA was on scaffold_m520, which has RNA-seq transcripts spanning its entire length and is predicted to encode *L. rubellus* Antimicrobial peptide lumbricin-1 (O96447) [25]. This suggested that the scaffold was indeed from *L. rubellus* but also contained an RNA structure more commonly associated with bacteria, or was a false positive identification. Only 6 of the 51 structures were on a contig/scaffold that contained a gene prediction making it hard to be certain as to their origin. For this reason, to fully investigate the presence of this ncRNA within the *L. rubellus* genome will require experimental analysis or a more contiguous assembly as only then can the origin of a fragment be certain.

Table 4.3: RFam search results part 1

ID	Family	Description	Number
RF00001	5S_rRNA	5S ribosomal RNA	9
RF00002	5_8S_rRNA	5.8S ribosomal RNA	5
RF00003	U1	U1 spliceosomal RNA	8
RF00004	U2	U2 spliceosomal RNA	11
RF00005	tRNA	tRNA	388
RF00007	U12	U12 minor spliceosomal RNA	1
RF00009	RNaseP_nuc	Nuclear RNase P	1
RF00012	U3	Small nucleolar RNA U3	2
RF00015	U4	U4 spliceosomal RNA	1
RF00017	SRP_euk_arch	Eukaryotic type signal recognition particle RNA	5
RF00020	U5	U5 spliceosomal RNA	3
RF00026	U6	U6 spliceosomal RNA	3
RF00027	let-7	let-7 microRNA precursor	8
RF00029	Intron_gpII	Group II catalytic intron	4
RF00032	Histone3	Histone 3' UTR stem-loop	9
RF00045	SNORA73	Small nucleolar RNA SNORA73 family	3
RF00053	mir-7	mir-7 microRNA precursor	3
RF00069	SNORD24	Small nucleolar RNA SNORD24	1
RF00074	mir-29	mir-29 microRNA precursor	2
RF00087	SNORD26	Small nucleolar RNA SNORD26	1
RF00089	SNORD31	Small nucleolar RNA SNORD31	3
RF00091	SNORA62	Small nucleolar RNA SNORA62/SNORA6 family	1
RF00103	mir-1	mir-1 microRNA precursor family	2
RF00188	SNORD103	Small nucleolar RNA SNORD103/SNORD85	1
RF00237	mir-9	mir-9/mir-79 microRNA precursor family	4
RF00239	mir-124	mir-124 microRNA precursor family	5
RF00241	mir-8	mir-8/mir-141/mir-200 microRNA precursor family	4
RF00251	mir-219	mir-219 microRNA precursor family	2
RF00261	IRES_L-myc	L-myc internal ribosome entry site (IRES)	2
RF00270	SNORD61	Small nucleolar RNA SNORD61	1
RF00306	snoZ178	Small nucleolar RNA Z178	12
RF00321	snoZ247	Small nucleolar RNA Z247	1

Table 4.4: RFam search results part 2

ID	Family	Description	Number
RF00323	snoR79	Small nucleolar RNA R79	1
RF00440	SNORD37	Small nucleolar RNA SNORD37	1
RF00446	mir-133	mir-133 microRNA precursor family	5
RF00485	K_chan_RES	Potassium channel RNA editing signal	38
RF00548	U11	U11 spliceosomal RNA	1
RF00563	SNORA53	Small nucleolar RNA SNORA53	1
RF00619	U6atac	U6atac minor spliceosomal RNA	1
RF00657	mir-184	microRNA mir-184	4
RF00667	mir-33	microRNA mir-33	1
RF00672	mir-190	microRNA mir-190	1
RF00694	mir-137	microRNA mir-137	12
RF00696	mir-203	microRNA mir-203	3
RF00700	mir-375	microRNA mir-375	1
RF00708	mir-450	microRNA mir-450	4
RF00832	mir-71	microRNA mir-71	5
RF00834	mir-268	microRNA mir-268	2
RF00885	MIR821	microRNA MIR821	2
RF00907	mir-941	microRNA mir-941	1
RF00920	MIR444	microRNA MIR444	41
RF00929	mir-574	microRNA mir-574	3
RF01005	MIR530	microRNA MIR530	2
RF01056	Mg_sensor	Magnesium Sensor	1
RF01059	mir-598	microRNA mir-598	446
RF01063	mir-324	microRNA mir-324	2
RF01170	snoU61	Small nucleolar RNA U61	3
RF01226	snoZ5	Small nucleolar RNA Z5	2
RF01289	snoR17	Small nucleolar RNA snoR17	1
RF01699	Clostridiales-1	Clostridiales-1 RNA	51
RF01848	ACEA_U3	ACEA small nucleolar RNA U3	2
RF01852	tRNA-Sec	Selenocysteine transfer RNA	1
RF01853	mtDNA_ssA	Mitochondrial DNA control region secondary structure A	1
RF01960	SSU_rRNA_eukarya	Eukaryotic small subunit ribosomal RNA	2

4.3.3 Gene finding

MAKER

For *ab initio* gene predictors only SNAP was used as it proved impossible to include the RNA-seq hints file generated for AUGUSTUS. Instead of a boot-strapping approach, the output of CEGMA was used as the training data for SNAP.

To reduce the computational load at the alignment stage, where possible the databases were customised to only include the relevant sequences. For example, of the 132,559 annelid transcripts (NCBI annelid ESTs and *H. robusta* and *C. telata* transcripts), only 19,052 mapped to the genome (identified by BLAST) therefore only these were provided. For the SwissProt [10] data (September 2011) used for the protein alignments, the data set was reduced from 531,473 to 16,910 sequences, and for the closely related transcript data 103,911 of the 163,282 *L. rubellus* Trinity transcripts that mapped to the genome were provided.

Additional gene finding

For AUGUSTUS both a hints file and custom model were generated, the former using PASA (Program for Automated Sequential Assignment) [178], ESTs and the augtrain.pl pipeline, the latter using the RNA-seq reads mapped to the genome using BLAT. AUGUSTUS was then run independently on an unmasked version of the genome on a single desktop machine.

Exonerate was run using the same reduced SwissProt database as above. After BLAST alignment to the genome these sequences mapped to 44,678 contigs. Protein2genome was run using these two subsets of data with the required length of protein mapping set at 50%.

4.3.4 Functional annotation

While MAKER2 has a sophisticated and thorough approach to gene finding, comparing its output to that of AUGUSTUS and protein2genome revealed many missing, credible

gene models (Table 4.5). The difference in annotation frequencies across the three prediction methods was explored further (Figure 4.7). There was a significant amount of non-overlapping annotation between the different gene finders gene sets. The GO, EC, KEGG and InterProScan annotations all shared a similar distribution, but the transcript mapping is significantly different showing a marked bias towards the MAKER2 predictions. This probably reflects the emphasis MAKER2 placed on the transcript evidence compared to the other methods. For these reasons a method to combine the separate gene predictions was required which removed the redundancy between the separate prediction techniques. Softwares designed for this purpose were investigated, notably Evidence Modeler [67] and Glean [49]. Evidence Modeler uses a strange GFF alignment file style and a cryptic weights file for prioritising evidence, and even with much parameter tweaking gave disappointing results splitting believable gene models. Glean uses GFF2 files, and required far too much work to downgrade the GFF3 files from the three gene prediction methods, and so was not assessed.

Table 4.5: Gene annotation metrics

	Source of gene model					
	AUGUSTUS	protein2genome	MAKER2	Manual	Total	Filtered
# of models	17158	16295	17965	47	51465	44648
GO	**37.01**	**87.13**	**61.80**	**63.83**	**61.56**	**58.87**
annotation	[2450]	[5692]	[4423]	[50]	[6286]	[6022]
KEGG	**11.09**	**35.52**	**25.47**	**27.66**	**23.86**	**22.87**
annotation	[727]	[1421]	[1565]	[9]	[1931]	[1909]
EC	**16.42**	**42.01**	**37.96**	**27.66**	**32.05**	**36.95**
annotation	[695]	[999]	[1138]	[13]	[1286]	[1286]
IPR	**39.43**	**81.81**	**72.00**	**70.21**	**64.25**	**61.97**
annotation	[2010]	[3473]	[3969]	[42]	[4647]	[4596]
ESTs	**24.19**	**28.28**	**34.40**	**21.28**	**29.04**	**27.97**
mapped	[2240]	[1733]	[2871]	[29]	[3569]	[3538]
Transcripts	**91.74**	**99.30**	**99.10**	**91.49**	**96.70**	**96.32**
mapped	[40930]	[25290]	[49562]	[717]	[61205]	[60989]
Any annotation	**96.3**	**99.9**	**99.4**	**93.6**	**98.5**	**98.3**

Numbers in bold are the percentage of models with that annotation type. Numbers in square brackets are the number of unique annotations for the prediction group.

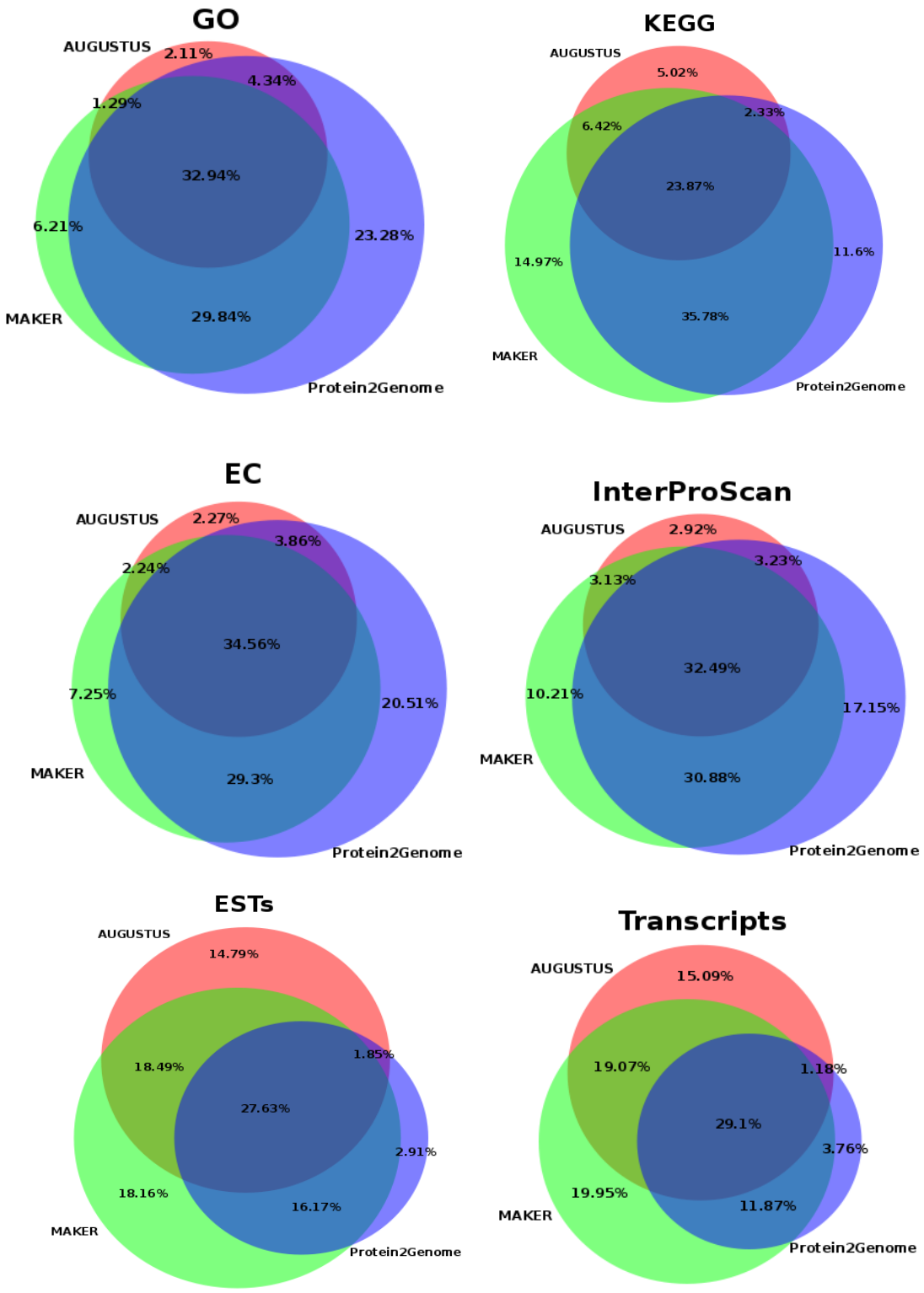

Figure 4.7: Gene annotation comparisons

Figures show area-proportional annotation associations for 6 annotation methods across the three gene prediction methods. Figures created using BioVenn [74].

Therefore, a novel approach to combine the separate gene predictions was created using a custom perl script (Appendix B). Each gene prediction was sorted by assembly fragment ID and then by start base, creating groups of distinct, overlapping gene predictions. If the group only contained one prediction, it was added to the final set. If a group contained multiple predictions, then a hierarchical selection procedure was implemented. All MAKER2 genes were selected (as this was the most 'trusted' method). If a prediction started before and overlapped with the MAKER2 prediction it was selected also, and if there were only protein2genome and/or AUGUSTUS predictions then AUGUSTUS was selected. This resulted in a filtered set of 44,648 gene models. The drop in gene numbers between the total and filtered gene sets (Table 4.5) reflected the hierarchical structure of the annotations.

The final number of genes is significantly higher than other similar genomes. This elevated number was expected due to the fragmented genome, resulting in genes split over contigs (Figure 4.9). There were many more short genes predicted for *L. rubellus* compared to the other two annelids. Protein2genome was the major source of short gene models (Figure 4.8) which is to be expected as it simply uses peptide alignment to identify potential exons.

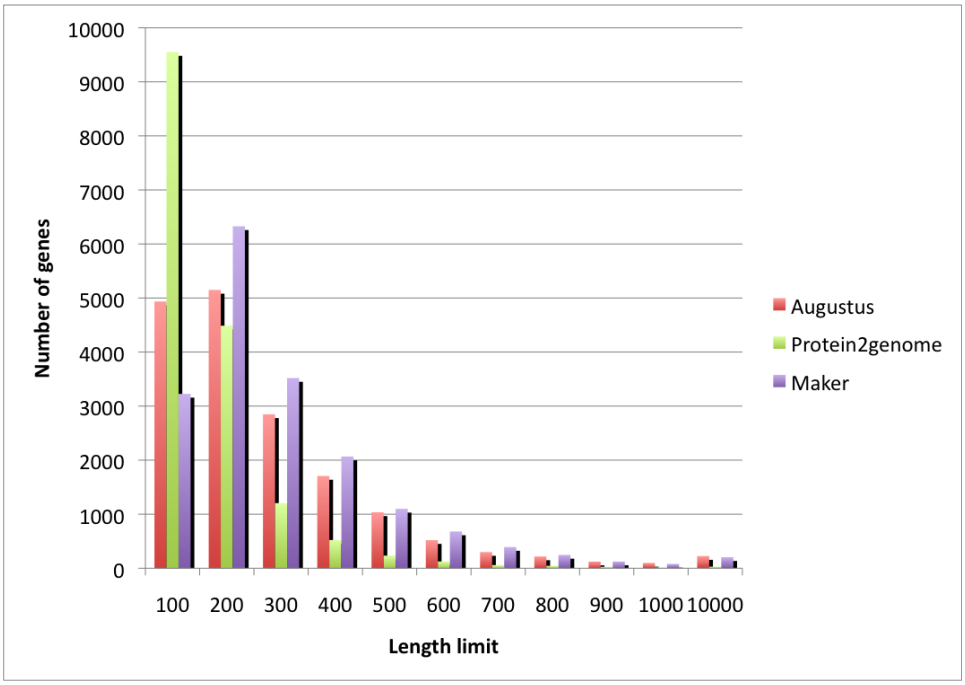

Figure 4.8: *L. rubellus* gene prediction lengths ($\geq$ 50 aa) compared between gene finding programmes

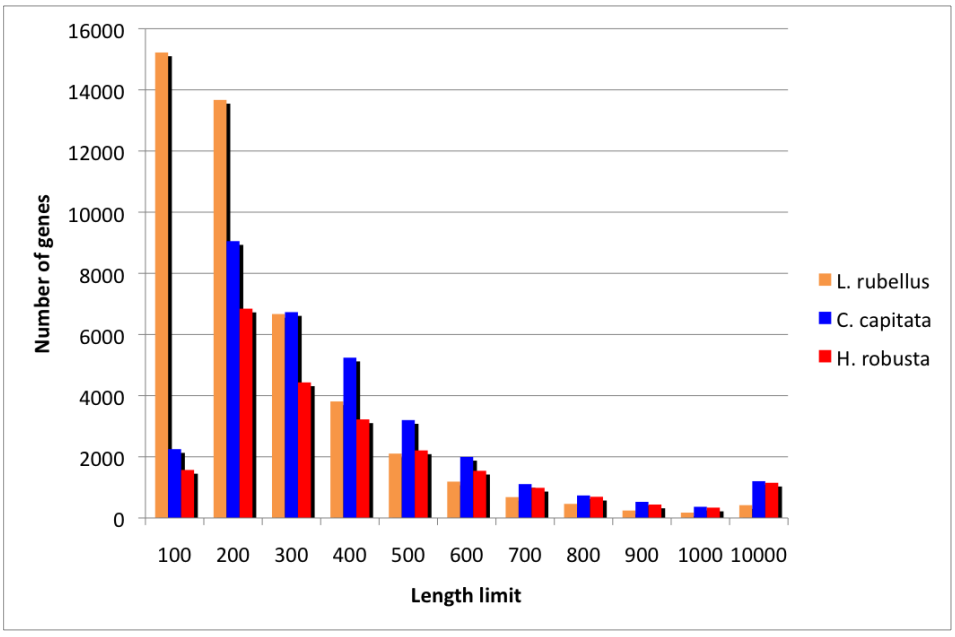

Figure 4.9: Annelid gene length comparisons ($\geq$ 50 aa)

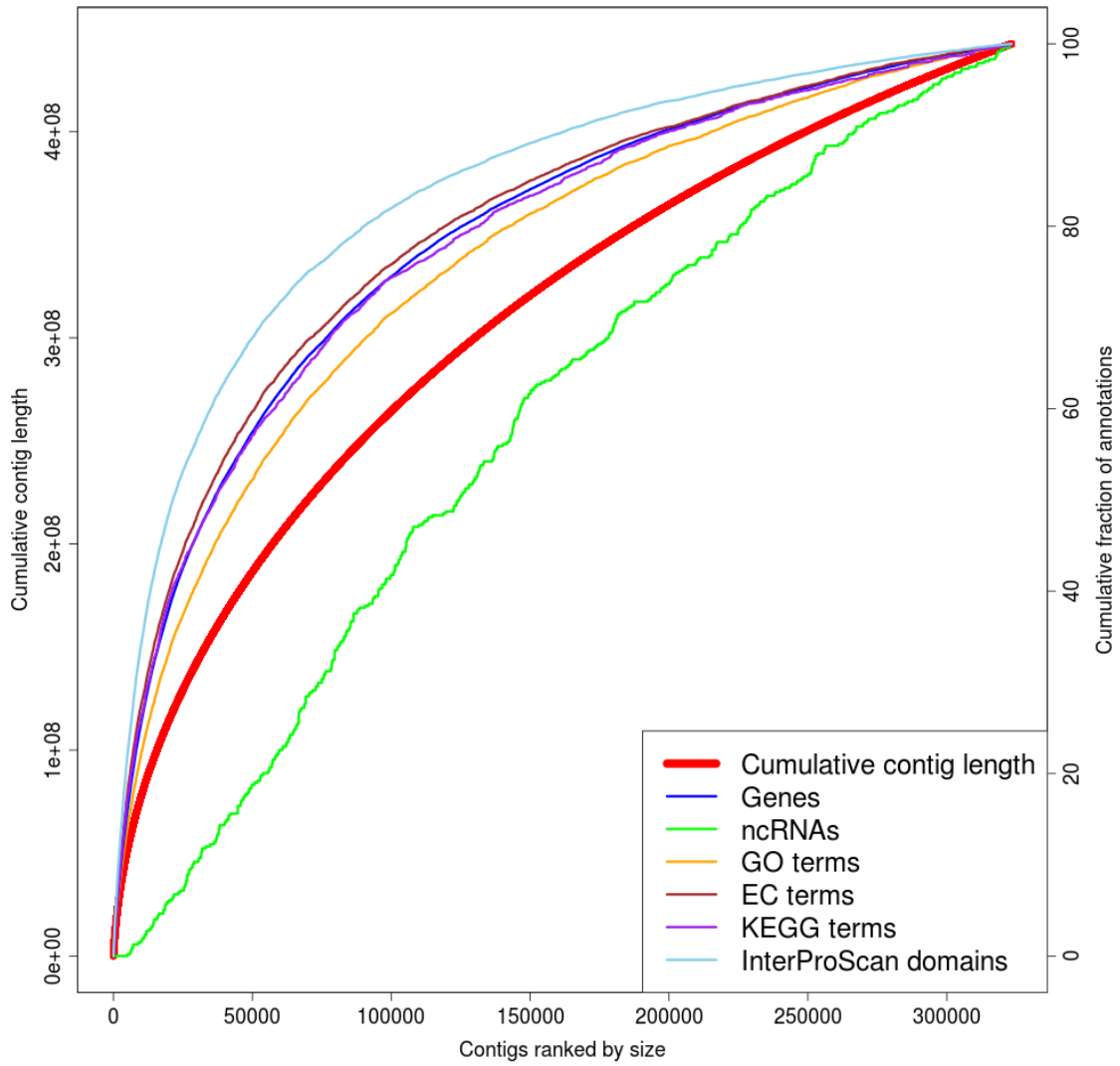

Figure 4.10: Comparison of cumulative assembly length and fraction of annotations

Contigs are ranked by size from largest to smallest. The cumulative contig length is calculated at each contig point which in turn is used to calculate the fraction of annotations.

The final assembly and annotations are summarised in Figure 4.10. A similar pattern was found for both the assembly and annotation data; the majority of the data was in the first longest 50% of contigs. At this point 75% (332 Mb) of the genome was present with the majority of the annotations (86% genes, 64% ncRNAs, 83% GO, 87% EC, 85% KEGG and 90% InterProScan domains). The major trend demonstrated by the annotation lines was to be above the contig line, meaning annotations are more likely to be found in longer contigs. In fact, the 100,000 shortest contigs contain around 12% of the sequence data and only 2-6% of the protein coding annotations. The only annotation line under the contigs was the ncRNAs which was almost linear. This suggests that these are sparser and as likely to be found in short as in long contigs, perhaps due to their size and repetitive nature.

4.3.5 Database, website and wiki

A community web portal for the *L. rubellus* genome project was constructed (www.earthworms.org). There are currently over 60 registered members, many of whom actively use the site to research a wide range of biological topics. The site has a simple construction (Figure 4.11) but provides intuitive and detailed search facilities (Figure 4.12), and includes a GBrowse2 genome browser (Figure 4.13). This resource was invaluable during the project as visualising the data as it emerged enabled key decisions to be made, such as identifying overlap issues between the three gene predictors, tweaking parameters in each of the gene predictors, and guiding manual annotations.

The two databases and associated front ends provide a combined setup which allows a complex set of search queries using the PostgreSQL DB and an in depth browsing facility using the SeqFeature MySQL DB which is not designed for any more than basic text searching. Designing a new database for the genome data was therefore beneficial as it allowed expansion of searching methods and data customisation.

An additional component of the project's web presence is a community Wiki, hosted at the University of Edinburgh (https://www.wiki.ed.ac.uk/display/LR/

The+Lumbricus+rubellus+genome+Wiki). This contains additional, private data for collaborators and served as a hub for reporting the progress of the genome project.

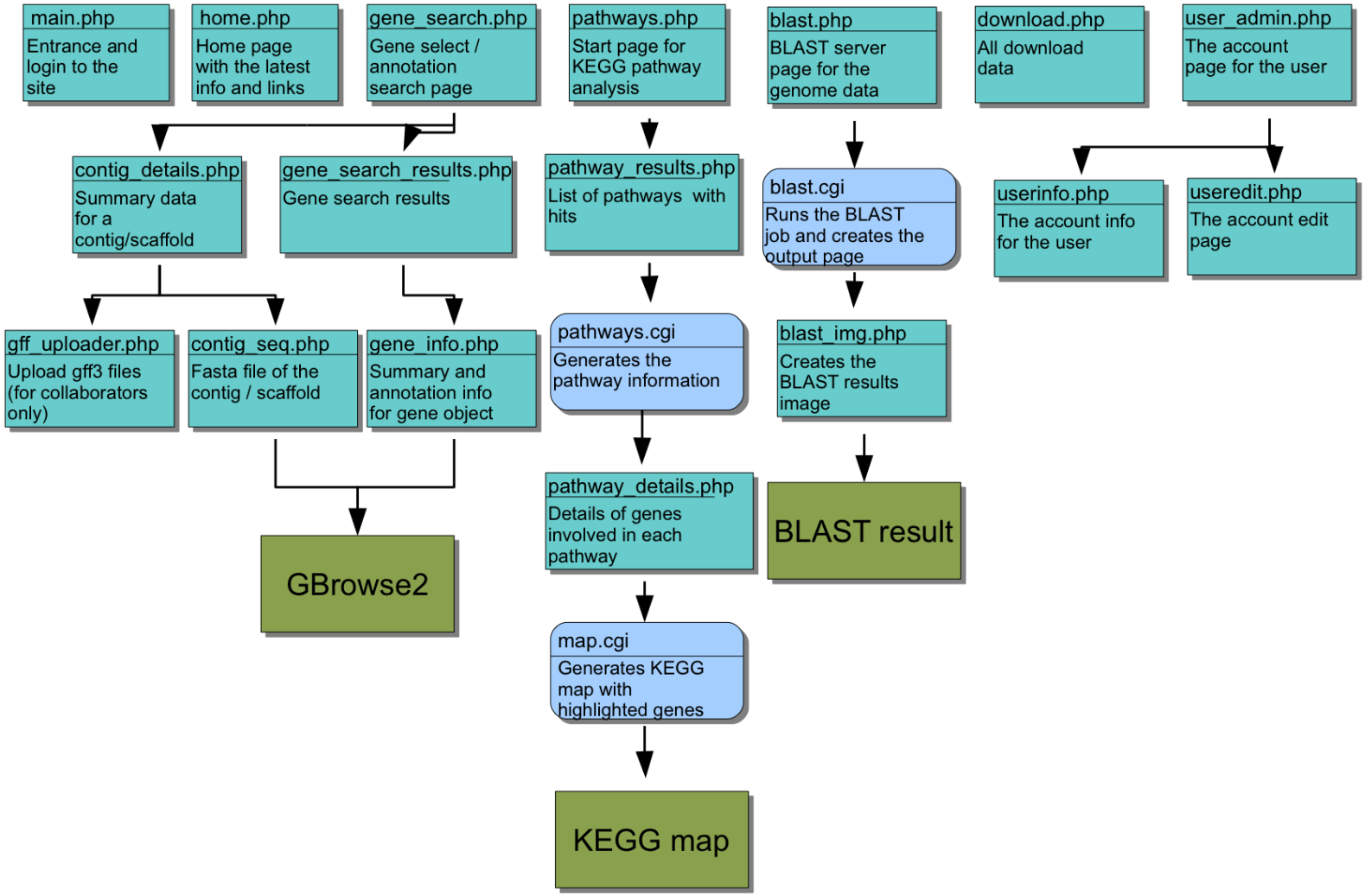

Figure 4.11: *L. rubellus* genome project web site structure

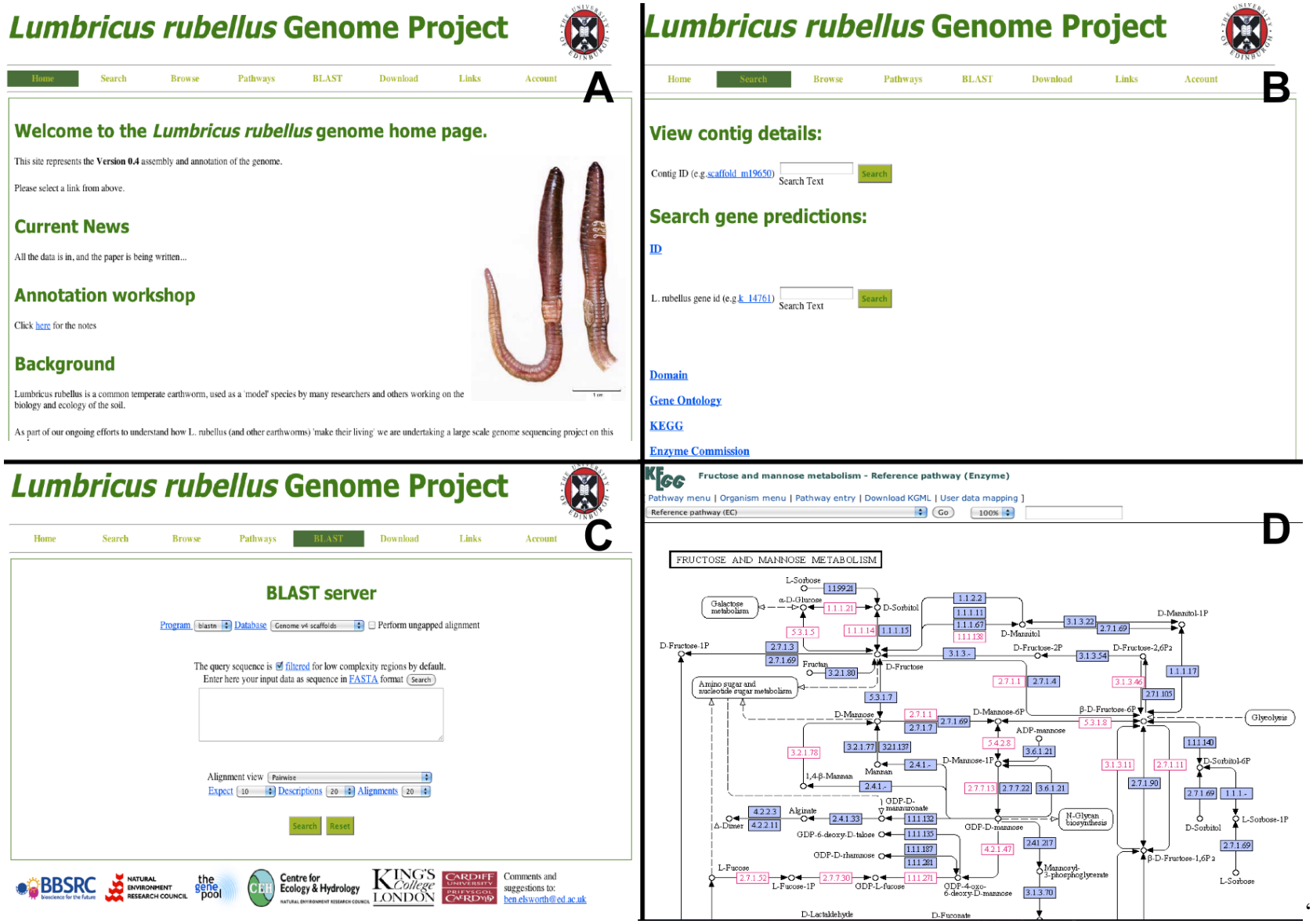

Figure 4.12: Screenshots from www.earthworms.org

Panel 'A' displays the Home page, 'B' the Search portal, 'C' the BLAST server for searching the contigs/scaffolds and annotations, and 'D' is an example of the end results of the Pathways section with the *L. rubellus* genes mapped onto a KEGG map (highlighted in red).

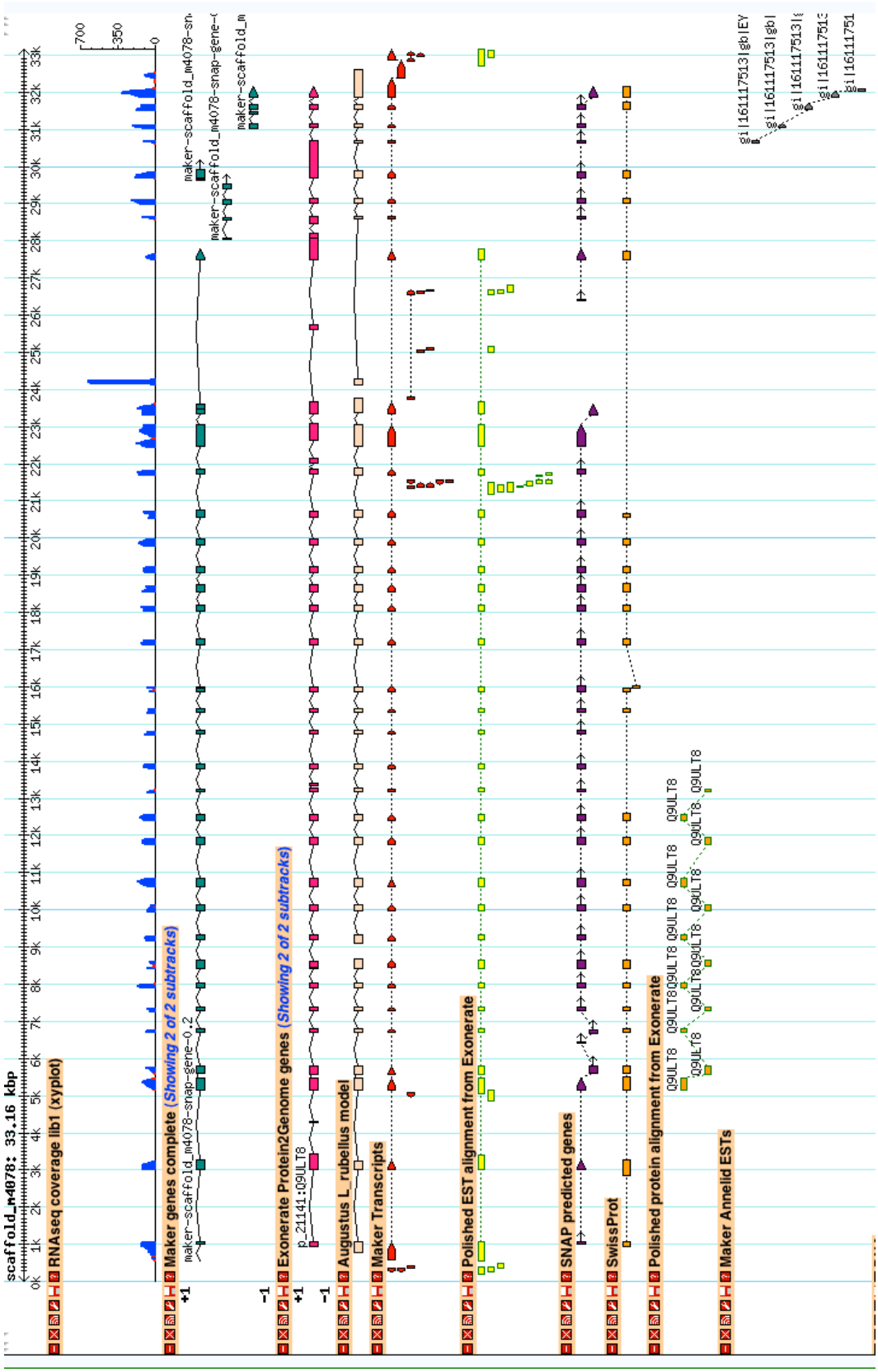

Figure 4.13: GBrowse2 screenshot of scaffold_m4078

4.4 Discussion

4.4.1 Genome annotation, assessment and validation

Section 3.4.1 discussed the completeness of a genome. This idea can be extended to the gene set and can also be used to compare the two, as one would hope the annotated genes contain the same protein-coding evidence as was found in the genome itself. Table 4.6 lists the completeness metrics for the scaffolded genome and the filtered and non-filtered gene sets.

There was a very noticeable drop in the number of mapped EST-derived transcripts between the genome and the predicted gene model sequences. The issue of mitochondrial reads was discussed in section 3.4.1. Examination of the number of ESTs per UniGene for the "missing" EST UniGenes (after removal of mitochondrial UniGenes) showed they had low coverage (1.8 ESTs per UniGene compared to an average of 4). There is also the likelihood that many of the UniGenes may derive from UTRs, and as these were missing from many of the gene predictions, they will not be mapped. Only 576 of the 4560 'missing' EST UniGenes had a positive match to either UniProt or other annelids ESTs suggesting that many of the missing UniGenes were error prone, contaminants, or highly diverged.

Of the 102,077 RNA-seq derived transcripts that did not map to the predicted genes, 96,089 (92%) had no BLAST matches to NCBI nr protein database (E-value cutoff 1e-5). Again, the transcripts that did not map to the final gene set were shorter than the average for the whole transcriptome (484 bp compared to 736 bp). The non-mapping transcripts may be UTR fragments, contaminants, or noncoding RNA transcripts missed in the gene predictions.

Improvement of the genome annotation could be achieved in many ways. First, a more accurate repeat library would help minimise false positive annotations. Second, better evidence would improve the training of gene finders and checking completeness, e.g. a more refined and complete transcript set. Third, a more contiguous genome would vastly improve the quality of the gene predictions as identifying genes would be less error

prone if they were complete and on single scaffolds. Lastly, a more advanced method to merge gene predictions would help produce a high quality set of predictions as combining predictions from multiple sources can be a powerful way of identifying the 'true' set of annotations.

Table 4.6: Genome vs gene completeness

Data set	CEGs (%)*	ESTs (%)**	Transcripts (%)**
Scaffolded genome	95.6	88.7	79.8
All gene models	93.4	43.9	37.5
Filtered gene models	92.4	43.5	37.4

* calculated based on positive BLAST hits to the 458 CEGs used in CEGMA, BLASTX for the genome and BLASTP for the peptide sequences of the gene models.

** calculated based on positive BLAT hits for the genome and BLASTX for the peptide sequences of the gene models.

4.4.2 Comparing annotations

The multiple annotation methods were combined to create one set of annotations for both the Gene Ontology (GO) terms and Enzyme Commission (EC) numbers. It was informative to compare the sources of these annotations.

Gene Ontology annotation comparisons

Three separate GO term predictors were used: annot8r, BLAST2GO and InterProScan (the latter also being used within BLAST2GO). Annot8r uses a subset of UniProt which has been annotated with GO terms. Positive BLAST matches to these that exceed a chosen E-value suggest that the query sequence can be annotated with the GO term. BLAST2GO works in a similar way in that an initial BLAST is performed and the results are used to map GO terms based on a look-up table matching UniProt to GO terms. It is slightly more involved as it includes a step that adds weight to the more 'reliable' annotations based on their origin, e.g. lower weight is given to electronic annotation. BLAST2GO also has the additional benefit of being able to incorporate InterProScan data, and adds this information during the GO annotation step. To compare the three, an area-proportional Venn diagram for the three separate sources of data was created (Figure 4.14). The overlap between the three methods was less than optimal as one predictor (BLAST2GO) contained more than half of the predictions. The variation is likely to be due primarily to the search space used by the three methods and the look-up tables they use to annotate positive mappings. The two GO terms uniquely identified by InterProScan and annot8r were based on quite short gene predictions of 75 and 111 amino acids, compared to a mean of 206 amino acids for the complete set of predictions. This may explain their absence from BLAST2GO predictions. To ensure comparable and largely complete metrics, all analyses in Chapter 5 used only BLAST2GO GO annotations.

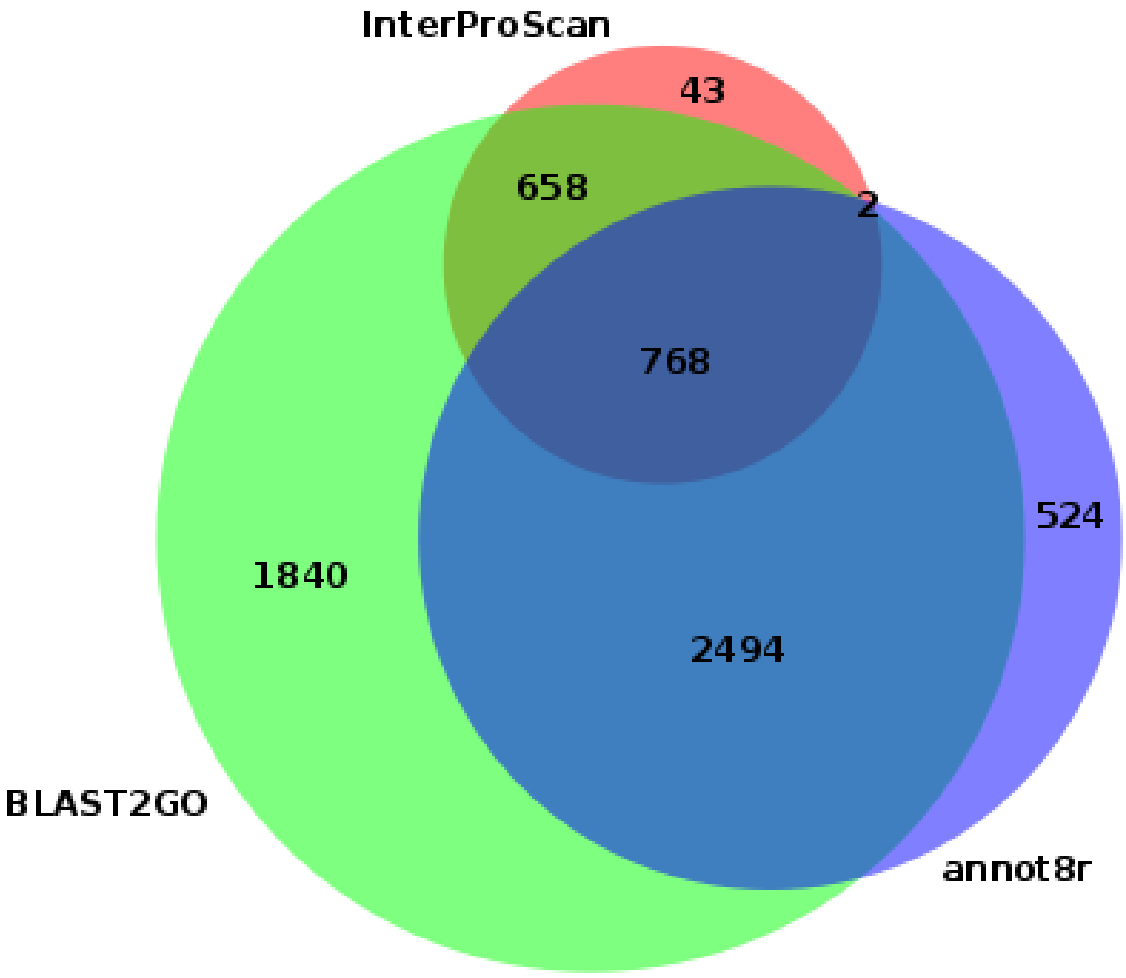

Figure 4.14: Comparison of GO annotations

Area proportional Venn diagram showing the numbers of unique GO terms predicted by BLAST2GO, InterProScan and annot8r and how the predictions overlap between the three methods.

Enzyme annotation comparisons

DETECT (Density Estimation Tool for Enzyme ClassificaTion) is a method for enzyme prediction that accounts for sequence diversity across enzyme families. Again, its method of assigning predictions to genes is similar to, but a little more in-depth, than annot8r. In DETECT, a subset of SwissProt proteins is generated based on those with known EC annotations, the proteins are then globally aligned to generate sequence profiles for each EC number, and for those alignments with 30 members or more a probability profile is generated. Profiles were available for 585 EC categories. Sequences of interest are then compared to the SwissProt subset and EC annotations are assigned whilst taking into account the probability profile. Figure 4.15 displays the overlap between DETECT EC terms, annot8r EC terms and BLAST2GO predictions mapped to EC terms using a GO to EC conversion from the Gene Ontology Consortium [8]. As expected DETECT predicted far fewer EC annotations (as only 585 categories were included) and the vast majority were predicted from annot8r as this programme has no such restriction. Both DETECT and BLAST2GO did however predict EC annotations not found by annot8r, but these had an average E-value of 0.045 (compared to a mean E-value for annot8r of 1.9e-7) implying they were weak predictions. For subsequent analysis annot8r EC predictions were used.

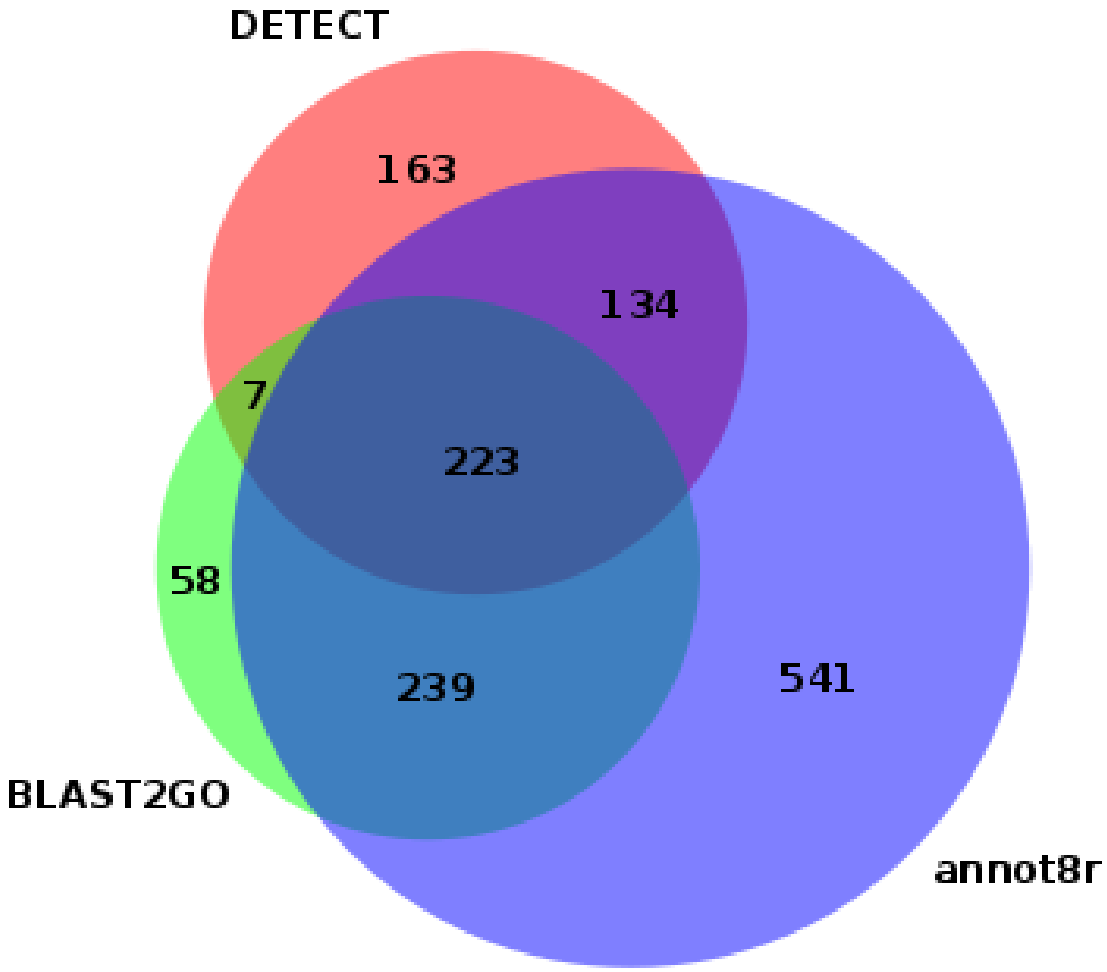

Figure 4.15: Comparison of EC annotations

Area proportional Venn diagram showing the numbers of unique EC terms predicted by DETECT, annot8r and InterProScan and how the predictions overlap between the three methods.

4.5 Annotation summary

The genome of *L. rubellus* has been used to generate the first set of repetitive elements, ncRNAs and protein-coding gene predictions. Although some of the protein-coding gene predictions are likely to be split over multiple contigs, it is believed that the vast majority of coding regions have been identified. This has been confirmed by the completeness metrics, as although CEGMA scores were low, over 90% of CEGs were identified in the gene models. Improvements to the predictions would be achieved by greater contiguity in the genome and a more refined set of transcripts.

Chapter 5

Investigations

The fourth data freeze of both the *L. rubellus* genome and its annotation was made on the 29th September 2011. This release will also be the basis of the main *L. rubellus* genome paper in preparation.

A project of this scale offers a wealth of opportunities for investigation of new and exciting biology, ranging from transposon evolution, gene loss and gain, ncRNA families, novel domains, and so on. However, only a small subset of these aspects could be investigated. In addition, as part of the genome paper, my collaborators are looking at several aspects including the immune response, Hox genes, metal transporters and drug response (as part of an overall aim to discuss the toxicological aspects of the earthworm). This final chapter covers my explorations into the biology of *L. rubellus* made possible by the genome and its annotation.

Although the set of filtered gene models has been proven to contain a comparable level of data (Table 4.6), for the following investigations all the gene models were used except in the comparative genomics section (Section 5.1).

These investigations used the postgreSQL and SeqFeature databases as well as directed BLAST searches, alignment, mapping to biochemical pathways and phylogenetic analyses as appropriate.

5.1 Comparative genomics

Comparing one genome to others can help illuminate interesting biology. If the species being compared are closely related, then a more intricate examination of their similarities and differences can be performed, highlighting more of the subtle nuances within. The aim of this investigation is to identify which elements define a species, in this case, what makes *L. rubellus*, or earthworms in general, unique? As mentioned previously in Chapter 1 there are two other unpublished annelid genomes, a leech, *H. robusta*, and a polychaete, *C. telata*. While these other annelids may not be close relatives to the earthworm, they are still useful and informative comparators.

This investigation utilised two methods for comparing the predicted gene sets from the three annelid genomes. Firstly, the proteins from the three species were clustered into putative gene families and the families of interest were annotated where possible. Secondly, differences in annotation were derived based purely on the individual annotations for the peptide sequences. The combination of these two methods ensured that the chances of a false positive annotation was minimised in the first method, and the effects of erroneous sequence predictions were minimised in the second.

5.1.1 Sequence comparisons

To perform comparative analyses, the predicted genes from *L. rubellus* were clustered into putative gene families based on sequence similarity along with the gene sets from *H. robusta* and *C. telata*. This approach reduced the effect of fragmented contigs and genes, as gene fragments should have clustered along with the more complete genes of the other species.

There are a number of tools designed for clustering, including TribeMCL [51], OrthoMCL [94], InParanoid [125]) and OrthoInspector [98]. OrthoMCL was used as it was well represented in the recent literature [158][163], and had the native ability to compare more than two protein sets (unlike InParanoid). OrthoMCL clustering proceeds in five phases. First, the protein sets are filtered. Here the default parameters of minimum

length 10 and maximum percent stop codons of 20% were used. Second, an all against all BLAST comparison of the protein sequences with a default E-value cutoff of 1e-5 is performed. Third, the BLAST results are parsed and a percent match length for each pairwise hit is computed. Fourth, all potential pairwise alignments are identified as inparalogs, orthologues or co-orthologues. Inparalogs are pairs of proteins from one species that have mutual hits better or equal to all hits to other species, while orthologues are pairs of proteins across two species that have hits as good or better than between proteins from the same species and co-orthologues are pairs of proteins across two species that are connected through orthology and inparalogy. The weights of the alignments are normalised based on the number of inparalogue pairs in each data set. The final phase is an implementation of the MCL clustering algorithm [51] filtered on the BLAST match matrix. The only additional parameter given to the MCL algorithm is the inflation value, which is suggested to be set at 1.5 as this "appears to balance sensitivity and selectivity: exhibiting consistency close to the maximum observed value, while excluding a minimum number of sequences" [94]. The inflation value affects the tightness of the clustering, and decreasing the value decreases the tightness of the clustering and leads to more clusters than at higher values.

TribeMCL was applied to the output of OrthoMCL using gene sets from *L. rubellus*, *H. robusta* and *C. telata* (Figure 5.1). A majority of clusters contained representatives from all 3 species at all inflation values. The *L. rubellus*-only clusters were more numerous than those containing only *H. robusta* or *C. telata*.

Figure 5.1 illustrates the effect the inflation value had on clustering. An interesting observation was the relative stability of the *L. rubellus* clustering compared to the other two annelids, which showed an increase in single species cluster abundance as the inflation value was increased. This suggests that, as the tightness of the clustering decreases, it was the other annelid sequences which were leaving the three annelid clusters, especially sequences from *C. telata*.

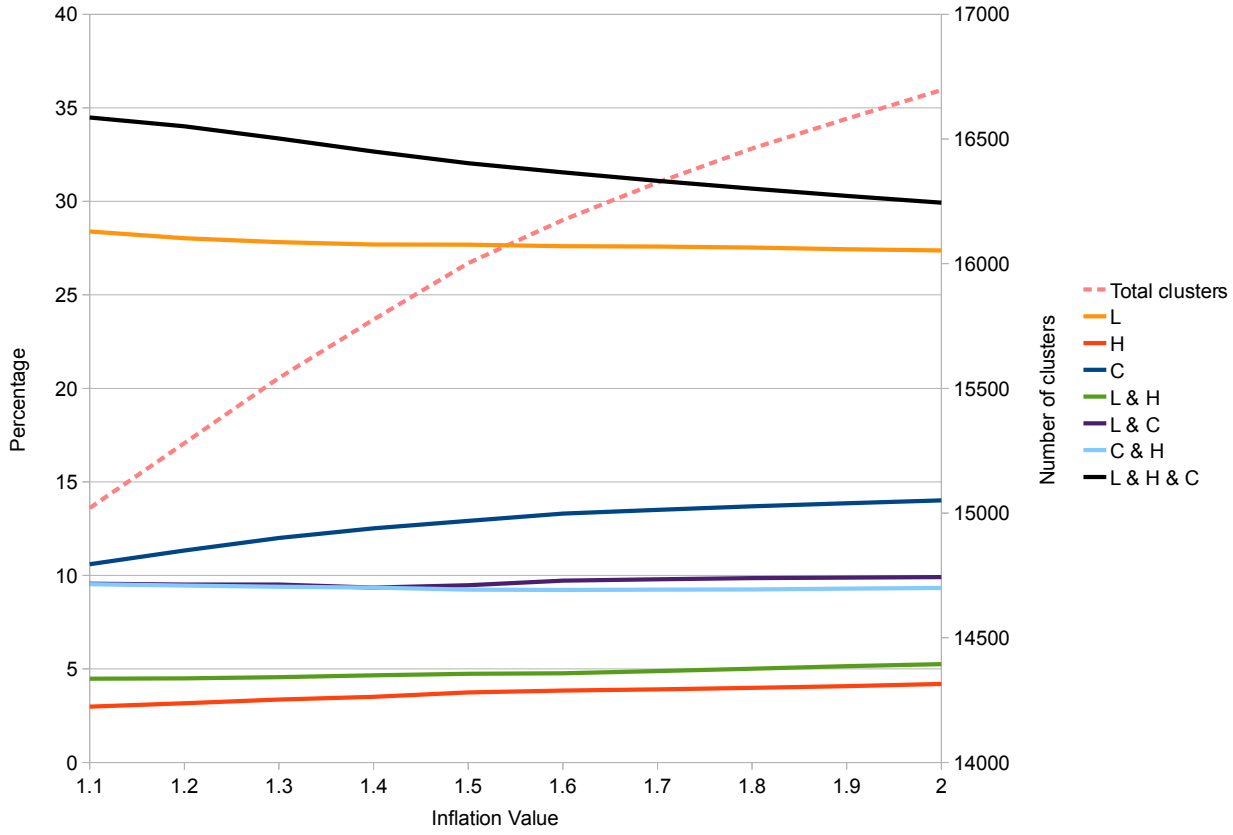

Figure 5.1: Plot of inflation value and percent of clusters per species group

L = *L. rubellus*, C = *C. telata* and H = *H. robusta*. Where an '&' is present, this represents a cluster with sequences from multiple species.

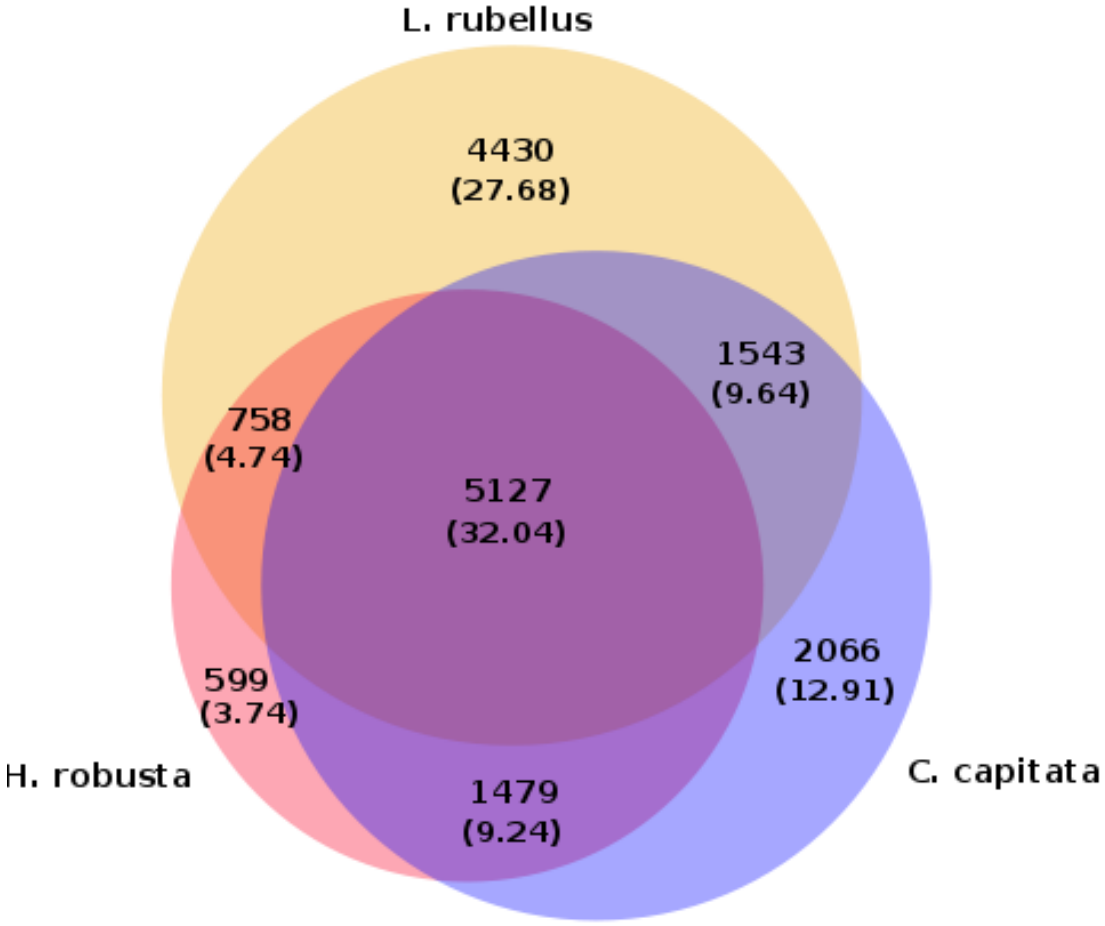

Figure 5.2: Area proportional venn diagram of OrthoMCL clusters

Counts and percentages (in brackets) of species composition for the 16,002 OrthoMCL clusters at inflation value 1.5. For example, 5,127 (32.04%) of clusters contained proteins originating from all three species and 4,430 (27.68%) of clusters contained proteins only from *L. rubellus*.

Table 5.1: Proportion of genes per OrthoMCL cluster at inflation value 1.5

	# genes	singletons	single	double	triple
L. rubellus	44648	**48.3** [1]	**25.6** [2.6]	**8.6** [1.7]	**17.5** [1.5]
H. robusta	23432	**40.1** [1]	**14.2** [5.5]	**15.4** [1.6]	**30.3** [1.4]
C. telata	32415	**27.2** [1]	**31.7** [5.0]	**17.9** [1.9]	**23.2** [1.5]

Singletons refers to genes not placed in any cluster, single is a single species cluster, double is a cluster containing genes from two species, and triple is a cluster containing genes from all three species. Numbers in bold are the percentage of genes within that group type, numbers in square brackets are the mean number of genes per cluster.

Figure 5.2 shows the frequency statistics for the 16,002 clusters from the orthoMCL output at an inflation value of 1.5. This analysis suggests a closer relationship between *L. rubellus* and *C. telata* as genes from these two species were present together in a higher percentage of clusters than either was with *H. robusta*, contrary to the expected phylogenetic relationships (as discussed in section 1.2). The large proportion of *L. rubellus* specific gene families may have arisen through redundancy that produced false positive paralogues, and is discussed below.

L. rubellus orthoMCL clusters

Figure 5.2 shows that a large number of the clusters are *L. rubellus* specific. However, Table 5.1 shows that within this set of clusters there are actually a lower percentage of sequences than in the *C. telata* only clusters. This is due to the lower number of sequences per cluster at 2.6 and 5 respectively. This implies that many of these are indeed not paralogous clusters but redundant, overlapping gene predictions caused by the merging of the three gene prediction methods. Indeed of the 4,430 *L. rubellus* specific clusters 3,491 have just two members and a further 511 have 3. Many of these are also from the same genome fragment: of those with two members 1695 (48.5%), for three members 15 (3%) and one 4 member cluster, suggesting there are indeed redundant gene predictions especially within the 2 member clusters. Removing these would reduce the *L. rubellus* specific clusters from 4,430 to 2,719, a number more in keeping with that of *C. telata* (Figure 5.2). The ten most populous *L. rubellus*-only clusters are shown in Table 5.2.

Table 5.2: Top 10 *L. rubellus* specific OrthoMCL clusters

Abundance rank	# genes	Mean length	Annotations	Descriptions
1	101	178	n/a	n/a
2	59	234	EC:1.14.14.1 [57] GO:0004497 [35] IPR001128 [34] Q4V8D1 [21]	unspecific monooxygenase monooxygenase activity Cytochrome P450 Cytochrome P450 2U1
3	40	102	n/a	n/a
4	37	112	n/a	n/a
5	34	97	n/a	n/a
6	30	126	IPR002350 [25]	Proteinase inhibitor I1, Kazal
7	29	103	EC:1.1.1.1 [28] GO:0005488 [27] IPR016040 [25]	alcohol dehydrogenase binding NAD(P)-binding domain
8	28	266	n/a	n/a
9	26	76	n/a	n/a
10	25	290	GO:0055114 [10] IPR003819 [9] Q9LIG0 [25]	oxidation-reduction process Taurine catabolism dioxygenase TauD/TfdA Clavaminate synthase-like protein

Numbers in square brackets are the number of elements with that annotation.

The cluster with the most members had relatively little annotation. Its members have have no BLAST hits to SwissProt or nr, and infrequent domain annotations. Ten constituent genes were annotated with a a Cadherin domain (IPR002126) and 4 with an Innexin domain (IPR000990). Three genes were annotated with GO:0007156, homophilic cell adhesion. These annotations suggest that this gene family is associated with cell-cell interactions, adhesions and the formation of ion channels. There were many matches to the Lumbribase UniGenes suggesting that this is indeed a common gene family within *L. rubellus* but its true role remains uncertain.

The second most populous cluster encodes a family of *L. rubellus*-specific cytochrome p450 (CYP) enzymes. CYP enzymes catalyse the oxidation of organic substances such as lipids and xenobiotic toxic chemicals such as drugs. These genes are of particular interest given the use of *L. rubellus* in ecotoxicological screening and research.

Many of the peptide sequences for each species did not cluster and remain as singletons (Table 5.1). These are either erroneous gene predictions, novel genes for that species or contaminants that have escaped the screening process. The mean length of *L. rubellus* singletons was 162 amino acids compared to non-singletons at 255 amino acids, suggesting that many were the result of incomplete or incorrect gene predictions. Different gene prediction tools yielded different proportions of singletons (Table 5.3). The majority of the AUGUSTUS and protein2genome sequence predictions were not clustered with any other sequences. Either these tools were more adept at finding novel gene objects, or they have created fragments of genes (exons) that are too short to cluster. Indeed, most of these predictions had supporting evidence from the transcriptome (Table 4.5) suggesting they are real genes, just too small to be clustered. The MAKER2 predictions showed a lower proportion of predictions failing to cluster, adding further weight to its credibility as a gene prediction method. This increased quality of prediction is the result of the MAKER2 pipeline, which masks repeats, combines both *ab initio* and alignment evidence and screens gene models, which neither exonerate protein2genome or AUGUSTUS do.

Table 5.3: Number of *L. rubellus* genes derived from each prediction method classified as orthoMCL singletons at inflation value 1.5

Cluster type	MAKER2	Manual	Exonerate protein2genome	Augustus
Singleton	7147	18	6521	7908
Non-singleton	10818	29	4385	7822
% singleton	40	38	60	50

Assuming that the shortest singletons were the least reliable, analysing the gene set in descending order of length may help maximise the likelihood of observing 'real' *L. rubellus* unique genes. The longest singleton was the gene object k_10917, a 2,909 amino acid protein predicted by MAKER2, which spanned the first 20 kb of a 50 kb scaffold and contained 24 introns. It had *ab initio* evidence from SNAP, transcript alignment evidence with high RNA-seq peak data and many SwissProt alignments. The main annotations for the gene object suggested it was a collagen-like protein, however the top BLAST matches from both SwissProt and nr predictions were both bacterial, albeit with low scores. Three of the next four longest singleton predictions were un-annotated while one was annotated as containing a C2H2-type zinc finger domain.

While singleton sequences are by definition dissimilar, the common annotations across the entire set were summarised to explore the main types of proteins represented (Table 5.4). It was striking that there was little evidence from the SwissProt and nr databases. Zinc finger-related annotations were common across the *L. rubellus* singletons, perhaps indicating that some may derive from retrotransposons.

There were many seven transmembrane-spanning segment annotations in the singletons. 7TM/GPCR proteins are commonly involved in chemosensation and olfaction, and are frequently observed to have undergone taxon specific amplifications in sequenced genomes. For example, in *C. elegans* there are over 1,200 7TM/GPCR-type genes, 6% of the total gene count. The abundance of 7TM/GPCR genes thus hints at the richness of the sensory capacity of *L. rubellus*.

L. rubellus also appeared to have a complex and unique kinome, as there were several hundred kinase-like proteins and nearly two hundred phoshpotases in the singleton set.

Table 5.4: Top annotations across several modalities for the the 21,594 *L. rubellus* singletons not clustered by OrthoMCL

Annotation	Rank	ID	Number	Description
InterProScan	1	IPR007087	830	Zinc finger, C2H2
	2	IPR000276	495	GPCR, rhodopsin-like, 7TM
	3	IPR011009	489	Protein kinase-like domain
Enzyme	1	2.7.11.1	604	non-specific serine/threonine protein kinase
	2	2.1.1.43	420	Histone-lysine N-methyltransferase
	3	3.1.3.48	171	Protein-tyrosine-phosphatase
KEGG	1	K10408	143	dynein heavy chain, axonemal
	2	K01769	61	guanylate cyclase
	3	K01539	44	sodium/potassium-transporting ATPase subunit alpha
SwissProt	1	Q9VZW5	33	FMRFamide receptor [*D. melanogaster*]
	2	P35500	31	Sodium channel protein para [*D. melanogaster*]
	3	Q6ZMW2	25	Zinc finger protein 782 [*H. sapiens*]
nr	1	328705603	21	zinc finger protein 62 homolog [*A. pisum*]
	2	256087133	17	transient receptor potential channel [*S. mansoni*]
	3	325296793	15	hyperpolarizaion-activated cyclic nucleotide-gated cation channel [*A. californica*]
GO	1	GO:0005515	1405	protein binding
	2	GO:0005524	1100	ATP binding
	3	GO:0008270	896	zinc ion binding

5.1.2 Annotation comparisons

The proteomes of the three annelids were compared through the spectrum of functional annotations. All three proteomes were annotated using the same tools, annot8r for EC, BLAST2GO for GO terms and InterProScan for domain identification. The use of three different methods for assigning function to the annotation reduced the risk of systematic bias.

While many softwares exist for identifying enriched terms within a gene annotation set, e.g. BiNGO [103], GOrilla [43] and GoMiner [183], these tend to be based on the use of model organisms with well defined gene lists and annotations. For non-model organisms the annotations may exist but standardised tables linking them to official gene names do not. Therefore, for this project, a novel method was developed that can use any set of annotations for three data sets to produce easily interpretable enrichment plots.

One frequently used comparative genomic approach is to order a set of annotations by frequency and then compare these frequencies across species (Figure 5.3). This method is useful for the identification of significant differences in highly ranked annotations.

InterPro ID	Description	Rank		
IPR011009	Protein kinase-like domain	1	4	2
IPR015880	Zinc finger, C2H2-like	2	6	6
IPR013087	Zinc finger C2H2-type/integrase DNA-binding domain	3	39	38
IPR017452	GPCR, rhodopsin-like superfamily	4	1	9
IPR000276	GPCR, rhodopsin-like, 7TM	5	2	11
IPR002290	Serine/threonine- / dual-specificity protein kinase, catalytic domain	6	9	3
IPR015943	WD40/YVTN repeat-like-containing domain	7	16	10
IPR020683	Ankyrin repeat-containing domain	8	3	17
IPR013783	Immunoglobulin-like fold	9	31	12
IPR002110	Ankyrin repeat	10	5	24
IPR020635	Tyrosine-protein kinase, catalytic domain	11	13	4
IPR005821	Ion transport domain	12	126	44
IPR009057	Homeodomain-like	13	33	5
IPR001757	ATPase, P-type, K/Mg/Cd/Cu/Zn/Na/Ca/Na/H-transporter	14	365	124
IPR016040	NAD(P)-binding domain	15	12	35
IPR001680	WD40 repeat	16	21	15
IPR001245	Serine-threonine/tyrosine-protein kinase catalytic domain	17	50	29
IPR018249	EF-HAND 2	18	32	26
IPR011992	EF-hand-like domain	19	30	22
IPR006201	Neurotransmitter-gated ion-channel	20	43	65

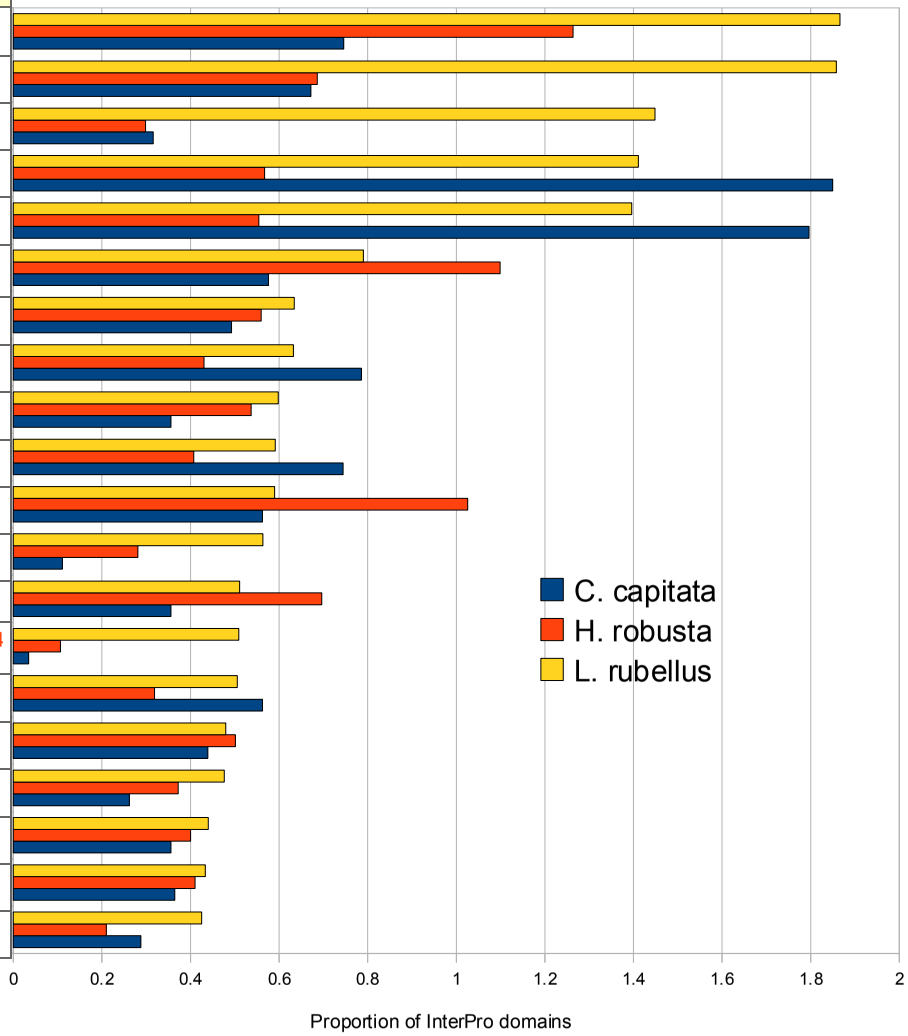

Figure 5.3: Top 20 *L. rubellus* InterPro domains compared to two other annelids

Domains are ranked by proportion of annotations. Only parent InterPro IDs are displayed to remove redundancy based on the parent-child relationships (ftp://ftp.ebi.ac.uk/pub/databases/interpro/ParentChildTreeFile.txt).

The limitation of such an approach lies in the restriction to a top ranking set of annotations. Therefore, it was decided to analyse all the annotations simultaneously using ternary plots. To generate the plots, for any annotation, a list of IDs was generated that included the numbers of genes with each annotation, e.g. a redundant list of GO terms, EC numbers, InterProScan domains, etc. The three lists were compared and for each ID the number present per list was counted and then normalised by dividing it by the total number of IDs for that list and multiplying by 100 to generate a percentage. This resulted in a four column data set (Table 5.5). The data file was loaded into R and plotted using the 'ternaryplot' function (part of the VCD package http://cran.r-project.org/web/packages/vcd/index.html). Each ternary plot was configured with the setting 'prop_size=TRUE' which plotted each point at a size proportional to the combined sum of all three entries.

Table 5.5: Ternary plot data format example

Data ID	*C. telata*	*H. robusta*	*L. rubellus*
1.1.1.1	3.53	1.08	2.72
1.1.1.10	5.16	1.69	1.05
1.1.1.100	7.29	2.63	3.63
1.1.1.101	3.03	0.94	0.91
1.1.1.102	4.65	1.42	1.189

Numbers represent the percentage of annotations from each of the annelid data sets for each data ID.

The resulting ternary plot has two useful features. The size of the points represents the relative abundance of each annotation term. The position in the plot indicates the relative contributions from the three species. For example, a large point near the *L. rubellus* vertex indicates an annotation that is abundant with a majority contribution from *L. rubellus*, suggesting a gene expansion in *L. rubellus* compared to the others.

The top ten predictions based on proportion in *L. rubellus* are labelled on the plots and also presented in more detail in tables. The top sets were selected by ranking predictions by cumulative prediction percentage across all three species, and selecting those that were present in at least one other species and had at least 50% of the combined contribution originating from *L. rubellus* .

Enzymes (EC annotations)

The annot8r enzyme predictions (EC) for the three species showed that the vast majority of enzymes had roughly equal representation across the three annelids (Figure 5.4). Some EC classifications were however, overrepresented in *L. rubellus* (Table 5.6). Many of the enzymes in the *L. rubellus* top ten were involved in ion transport, particularly of metals. There was further evidence for the expansion of the cytochrome P450 enzymes (e.g. unspecific monooxygenase) as also identified in *L. rubellus* unique clusters. This expansion may reflect the earthworm's lifestyle, in particular the various types of xenobiotics it comes into contact with compared to the other two annelids, especially when originating from such a unique habitat as an abandoned lead mine. *L. rubellus* had a relative overabundance of proteins annotated as Xenobiotic-transporting ATPase, also suggesting a greater focus on dealing with environmental toxins in the earthworm. The abundance of polypeptide N- acetylgalactosaminyltransferase in *L. rubellus* may relate to mucus production in this species.

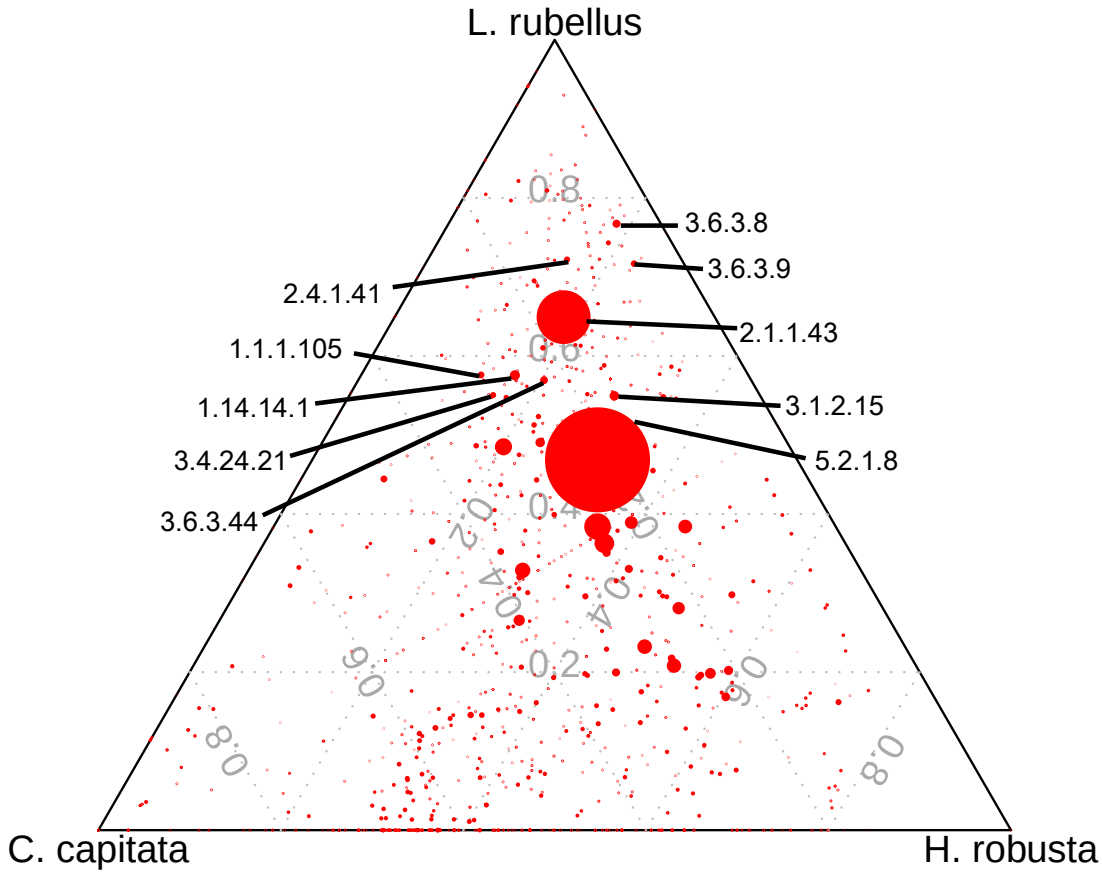

Figure 5.4: Ternary plot for enzyme prediction

The top ten predictions based on proportion in *L. rubellus* are labelled.

Table 5.6: The ten most significant *L. rubellus* enzyme predictions in the three-annelid comparison

EC number	Common name	*C. telata*		*H. robusta*		*L. rubellus*		Proportion
		%	#	%	#	%	#	in *L. rubellus*
3.6.3.8	Ca^{2+}-transporting ATPase	0.06	18	0.24	38	1.00	153	76.74
2.4.1.41	Polypeptide N-acetylgalactosaminyltransferase	0.11	31	0.13	21	0.63	96	72.20
3.6.3.9	Na^{+}/K^{+}-exchanging ATPase	0.05	15	0.22	35	0.69	106	71.68
2.1.1.43	Histone-lysine N-methyltransferase	1.95	552	2.18	344	7.64	1165	64.93
1.1.1.105	All-trans-retinol dehydrogenase (NAD^{+})	0.30	84	0.13	21	0.58	89	57.61
1.14.14.1	Unspecific monooxygenase	0.45	128	0.30	47	1.02	155	57.56
3.6.3.44	Xenobiotic-transporting ATPase	0.28	80	0.25	40	0.71	108	56.94
3.4.24.21	Metalloendopeptidases	0.25	70	0.13	21	0.47	71	55.06
3.1.2.15	Ubiquitin thiolesterase	0.26	74	0.47	75	0.90	137	54.96
5.2.1.8	Peptidylprolyl isomerase	0.23	65	0.43	68	0.66	101	50.09

Gene ontology annotations

Analyses of the BLAST2GO gene ontology predictions (Figure 5.5 and Table 5.7) again revealed a general similarity between the species. However, there were outliers overrepresented in *L. rubellus* annotations. These were mainly focused around ATPase activity, ion transport and microtubule based function, with the largest relative expansion being dynein complex (GO:0030286) with 204 annotations compared to 7 and 6 for *C. telata* and *H. robusta* respectively, which is again associated with microtubule motor activity.

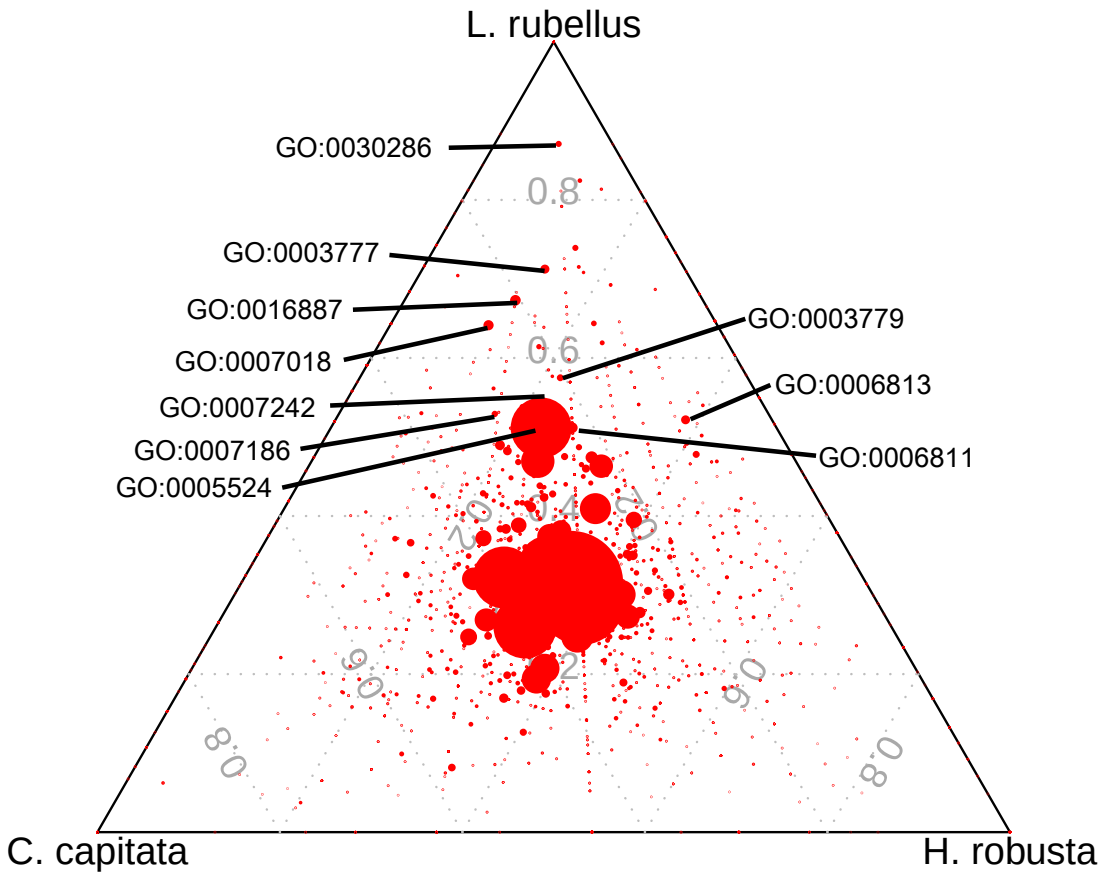

Figure 5.5: Ternary plot for gene ontology predictions

The top ten predictions based on proportion in *L. rubellus* are labelled.

Table 5.7: The ten most significant *L. rubellus* comparative gene ontology term predictions in the three-annelid comparison

GO term	Common name	*C. telata*		*H. robusta*		*L. rubellus*		Proportion
		%	#	%	#	%	#	in *L. rubellus*
GO:0030286	Dynein complex	0.02	7	0.02	6	0.31	204	87.09
GO:0003777	Microtubule motor activity	0.10	27	0.08	19	0.45	296	71.25
GO:0016887	ATPase activity	0.16	40	0.09	20	0.52	345	67.31
GO:0007018	Microtubule-based movement	0.19	47	0.08	17	0.47	312	64.15
GO:0003779	Actin binding	0.09	22	0.09	19	0.25	163	57.50
GO:0007242	Intracellular signal transduction	0.13	32	0.11	24	0.28	183	54.10
GO:0007186	G-protein coupled receptor signalling pathway	0.11	28	0.06	14	0.20	132	52.92
GO:0006813	Potassium ion transport	0.06	16	0.24	51	0.33	214	52.18
GO:0006811	Ion transport	0.17	44	0.20	44	0.40	260	51.18
GO:0005524	ATP binding	1.43	350	1.28	266	2.83	1864	51.17

InterProScan Domains

Perhaps the most telling of the proteome comparisons can be performed with domain annotations. Analyses of the InterProScan domain annotations across the three annelid species (Figure 5.6 and Table 5.8) highlighted the ATPase, P-type, K/Mg/Cd/Cu/Zn/Na/Ca/Na/H-transporter domain (IPR001757). This domain had a very high *L. rubellus* proportion at 78.31% and a large number of predictions (310 compared to 20 and 42 for *C. telata* and *H. robusta* respectively). This domain is representative of P-type ATPases, two of which were identified in the earlier enzyme analysis, Ca^{2+}-transporting ATPase (EC:3.6.3.8) and Na^{+} /K^{+}-exchanging ATPase (EC:3.6.3.9).

Other annotation highlights include large numbers of Zinc finger annotations (IPR013087, IPR015880 and IPR007087) two potassium channel related domains (IPR003091 and IPR003131) and another domain associated with microtubule motor activity (IPR001752).

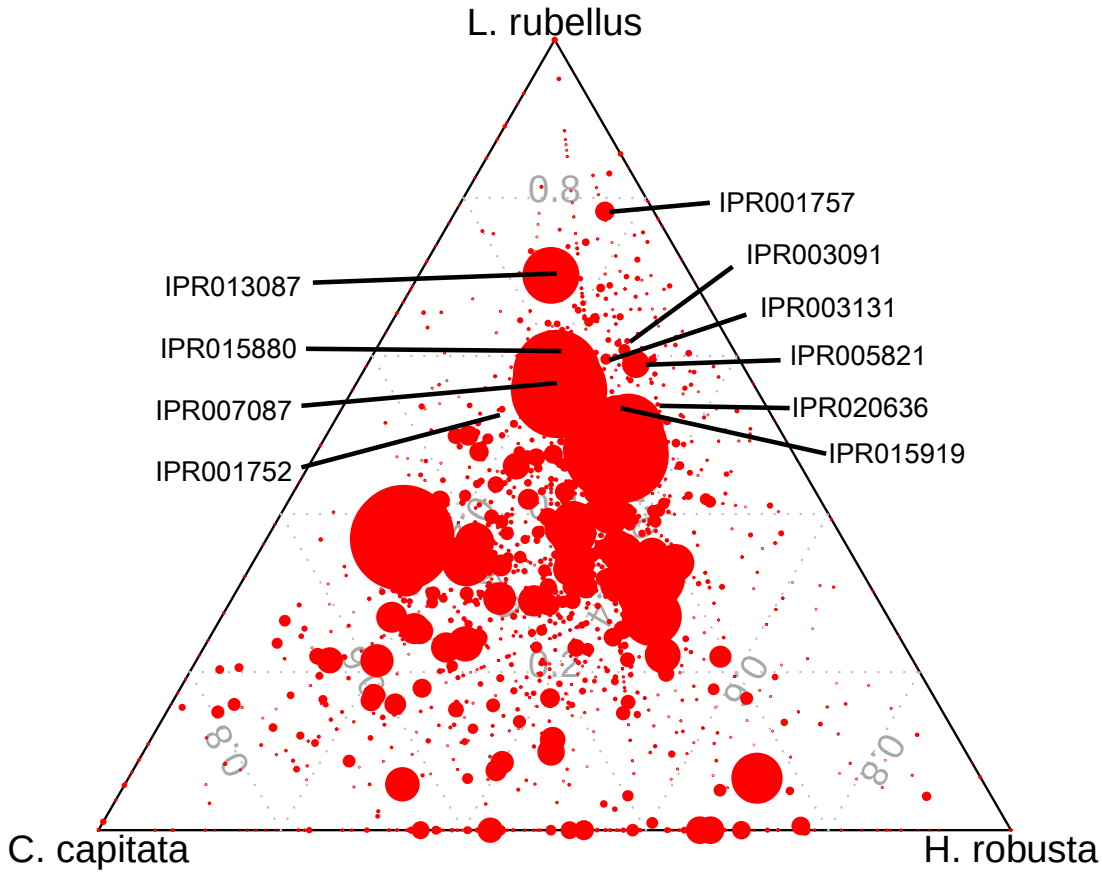

Figure 5.6: Ternary plot for InterProScan domains

The top ten predictions based on proportion in *L. rubellus* are labelled.

Table 5.8: The ten most significant *L. rubellus* comparative InterProScan domain predictions in the three-annelid comparison

Domain ID	Common name	*C. telata*		*H. robusta*		*L. rubellus*		Proportion
		%	#	%	#	%	#	in *L. rubellus*
IPR001757	ATPase, P-type, K/Mg/Cd/Cu/Zn/Na/Ca/Na/H-transporter	0.03	20	0.11	42	0.51	310	78.31
IPR013087	Zinc finger, C2H2-type/integrase, DNA-binding	0.32	182	0.30	118	1.45	882	70.21
IPR003091	Voltage-dependent potassium channel	0.04	25	0.10	39	0.22	134	60.76
IPR003131	Potassium channel, voltage dependent, Kv, tetramerisation	0.05	28	0.09	34	0.20	121	59.61
IPR005821	Ion transport	0.11	64	0.28	111	0.56	343	58.96
IPR015880	Zinc finger, C2H2-like	0.67	387	0.69	271	1.86	1131	57.77
IPR007087	Zinc finger, C2H2	0.77	443	0.80	317	1.97	1202	55.68
IPR020636	Calcium/calmodulin-dependent/calcium-dependent protein kinase	0.08	46	0.20	80	0.32	194	53.02
IPR001752	Kinesin, motor domain	0.08	48	0.07	29	0.17	105	52.39
IPR015919	Cadherin-like	0.13	73	0.23	92	0.36	220	50.12

Annotation comparisons summary

The ternary plot analyses identified many functional areas of the proteome that were significantly expanded in *L. rubellus* compared to the other two annelids. The most common annotations highlighted in all three analyses were metal ion ATPase related domains and their enzymatic functions, key components of metal homeostasis. For both the enzyme and domain data the most significant was Calcium ATPase, but this term was the 11th most significant for the GO data. There were 195 gene objects in the *L. rubellus* gene prediction set with Ca^{2+}-transporting ATPase (EC:3.6.3.8) annotations, 266 with calcium ion transport (GO:0006816) annotations, and 310 with ATPase, P-type, K/Mg/Cd/Cu/Zn/Na/Ca/Na/H-transporter (IPR001757) annotations. However, a search for the ATPase, P-type, calcium-transporting domain (IPR005782) only returned two gene objects.

Morgan and Morgan investigated the interactions between calcium and lead metabolism in *L. rubellus* [112][113] and recent work by Andre *et al* [6] identified multiple variants of the sarcoplasmic/endoplasmic retriculum calcium ATPase (SERCA) in *L. rubellus*. They propose that "Pb-adaptation traits may be inextricably linked to regulators of Ca physiology" [6]. Of the two classes of variants detected by Andre *et al*, 15 genes corresponding to variant 1 (GI:260181324) and 20 genes corresponding to variant 2 (GI:260181326) were identified in the gene predictions. Considering that the worms sequenced to generate this genome originated from the same lead mine as those studied by Morgan and Morgan and Andre *et al*, this result was somewhat expected. However, *L. rubellus* had many more variants of this ATPase than the two other annelids. This feature could be common to all *L. rubellus*, or could be an adaptive response to lead unique to the earthworms at this heavily contaminated site.

The analysis of the InterPro domains identified an expansions of Zinc finger domains in *L. rubellus*. Lineage specific expansion of C2H2-like Zinc finger domains is well known, for example in plants [50] and mammals [165], and a proposed function for the non-conserved unique domains is transcriptional regulation. Expansion of Zinc finger genes has even been linked to speciation "via modulation of expression of genes influencing re-

productive fitness or behavior" [88]. It is hypothesised that even single amino acid change within seemingly repetitive Zinc finger regions can lead to differential regulation of genes involved in reproduction. This expansion of domains may therefore be linked to the proposed cryptic species complex from the site in Wales from which the worm originated [6]. These speciation events are common in taxa found to be thriving in specialised environments, therefore if the next earthworm to be sequenced is from a less extreme landscape a detailed analysis of the causes underlying speciation can be performed and any links to Zinc finger domains can be identified.

In addition to the proposed Calcium ATPase function, the high levels of potassium related transport proteins across all three ternary plots may be related to earthworm muscle fibres as suggested by Volkov *et al* [172] where Na^+-K^+-ATPase is identified as an important electrogenic factor in earthworm longitudinal muscle fibres.

Annotations unique to *L. rubellus*

Annotations that were only present in *L. rubellus* compared to the other two annelids suggest an evolutionary gain for the earthworm or a complete loss for the other two species. To avoid false positive annotations, analysis was focused on those *L. rubellus* specific annotations with the highest frequency (Table 5.9).

L. rubellus appeared to have 94 'Antifreeze protein, type I' domains (IPR000104), while *H. robusta* and *C. telata* had none. However, all of the gene objects with this domain annotation were predicted using the AUGUSTUS gene predictor, and IPR000104 is a very short low-complexity domain. It is thus likely that this set of genes was the result of over prediction by AUGUSTUS, the probability of which is increased by a complete lack of RNA-Seq evidence on any of the gene predictions.

In the EC annotations, *L. rubellus* had two unique glycoside hydrolase enzymes. The first, mannan endo-1,4-β-mannosidase (EC:3.2.1.78), also known as Endo-β-1,4-mannanase, is a polysaccharide-degrading enzyme associated with the degradation of plant cell walls. The enzyme was predicted by both annot8r and DETECT. Thorough inspection revealed no homolog in the other two annelids, suggesting that this annotation was indeed unique

to *L. rubellus*. A study of cellulolyitc systems in animals [174] identified only three occurrences of endo-β-1,4-mannanase, all in molluscs (the blue mussel *Mytilus edulis*, sea snail *Haliotis discus* and freshwater snail *Biomphalaria glabrata*). The enzyme is also present in the gastropod *Haliotis discus hannai* [124] and the sea hare *Aplysia kurodai* [182]. However, like all cellulases, it is still much more commonly associated with bacteria. The *L. rubellus* gene prediction (k_10972) was apparently missing the first exon. This exon had been predicted by SNAP but not used by MAKER2 and it was therefore added manually using Apollo to create a full length sequence.

The second glycoside hydrolase enzyme unique to *L. rubellus* was Glucan endo-1,6-beta-glucosidase. Closer inspection of the contig containing this domain (contig_162141) reveals a short contig (1.4 Kb) a high GC content (56.35%) and a very close similarity to bacteria. Therefore, it can be assumed that this is a bacterial sequence that has slipped through the filtering process.

A conserved cellulolytic glycosyl hydrolase was previously identified in the Lumbribase UniGenes [37] but this gene is split over a number of contigs and is therefore not represented in the predicted genes set. Since then a protein sequence for the earthworm *Eisenia andrei* was submitted to GenBank in 2007 for the homolog of Endo-β-1,4-glucanase (`http://www.ncbi.nlm.nih.gov/protein/ACE75511.1`). Therefore it is indeed possible that the mannanase enzyme is present and unique to *L. rubellus* with respect to the other annelids.

Studies into the microbiome of the earthworm [13] have identified many varied glycosyl hydrolases. Therefore, to assess the likely origin of the *L. rubellus* homolog similar protein sequences were downloaded from GenBank, aligned using CLUSTAL [167] (see Appendix C), and the alignment subjected to phylogenetic analysis in MrBayes [73] using the following parameter set: include 110-705; lset nst=6 rates=invgamma; mcmc ngen=1000000 samplefreq=1000; sump burnin=250; sumt burnin=250; (Figure 5.7).

Table 5.9: Most common unique *L. rubellus* annotations

Class	Annotation ID	Name	Number
InterPro domain	IPR000104	Antifreeze protein, type I	94
	IPR020703	Tyrosine-protein kinase, non-receptor Src64B	7
	IPR009311	Interferon-induced 6-16	7
	IPR004342	EXS, C-terminal	7
	IPR015577	Interferon-induced Mx protein	5
	IPR001004	Adrenergic receptor, alpha 1C subtype	5
EC number	1.7.1.6	Azobenzene reductase	1
	3.1.3.27	Phosphatidylglycerophosphatase	1
	3.2.1.75	Glucan endo-1,6-beta-glucosidase	1
	3.2.1.78	Mannan endo-1,4-β-mannosidase	1
KEGG pathway	K05680	ATP-binding cassette, subfamily G (WHITE), member 4	12
	K04138	adrenergic receptor alpha-2A	12
	K04422	mitogen-activated protein kinase kinase kinase 13 [EC:2.7.11.25]	9
	K04635	Guanine nucleotide binding protein (G protein), alpha 11	7
	K02522	inositol 1,4,5-triphosphate receptor, invertebrate	7
GO annotation	GO:0010045	Response to nickel cation	9
	GO:0008061	Chitin binding	9
	GO:0004691	cAMP-dependent protein kinase activity	9
	GO:0046958	nonassociative learning	8
	GO:0042800	histone methyltransferase activity (H3-K4 specific)	8
	GO:0016742	hydroxymethyl-, formyl- and related transferase activity	8
	GO:0016309	1-phosphatidylinositol-5-phosphate 4-kinase activity	8
	GO:0015279	store-operated calcium channel activity	8
	GO:0009744	response to sucrose stimulus	8
	GO:0008095	inositol 1,4,5-trisphosphate-sensitive calcium-release channel activity	8
	GO:0006556	S-adenosylmethionine biosynthetic process	8
	GO:0005833	hemoglobin complex	8
	GO:0004832	valine-tRNA ligase activity	8
	GO:0001518	voltage-gated sodium channel complex	8

The tree clearly divided the bacterial sequences from the animal-derived ones, albeit with low posterior probabilities. The *L. rubellus* gene is within the eukaryotic clade, but not grouped with the other Lophotrochozoan (mollusc) sequences as would be expected. Instead it is grouped (with 0.83 posterior probability) with two sequences from the deuterostome lancelet (*Branchiostoma floridae*). Inspection of the lancelet sequences suggests these may be poor predictions, and thus that this unexpected association may be a case of long branch attraction. The absence of the mannanase in the other annelid genomes implies that the gene has been lost in *H. robusta* and *C. telata*, but retained in *L. rubellus*.

Figure 5.7: Phylogenetic tree of mannan endo-1,4-β-mannosidase

Tree generated using MrBayes over 1 million generations and rooted with the bacterial clade. Sequences from Bacteria are coloured in turquoise, Arthropoda are in pink and Mollusca in yellow. The scale bar indicates the number of amino acid changes per branch length and the numbers on the nodes indicate the posterior probability scores for each node.

5.2 Bacterial DNA in the *L. rubellus* genome assembly

Sequencing of the earthworm genome provides an opportunity to further investigate the biological interactions of the earthworm and its closest neighbours. Although measures were taken to minimise the amount of non-earthworm DNA that was sequenced (Section 2.3.1), it was inevitable that some would be present. Prior to filtering there was a significant number of genome fragments in the assembly of suspected bacterial origin. These were identified by searching for fragments annotated as exclusively matching bacterial genes and with low sequence coverage (less than 5x). Many gene objects were identified in the pre-filtered assemblies (Table 5.10). These could have derived from several sources:

1. Symbiotic bacteria - bacteria that exist permanently within the earthworm, e.g. *Verminephrobacter*.

2. Passive bacteria - bacteria present within or on the earthworm at the time of DNA extraction that are not symbionts but could still represent a biological interaction (for example gut commensal).

3. Sequencing contamination - bacterial DNA accidentally sequenced because of laboratory contamination.

The first two of these would be biologically interesting, but unfortunately the last two provide false positives that cannot be filtered out without thorough examination. Knapp *et al* [86] list bacterial taxa similar to those in Table 5.10 as diet-related gut microbiota of *L. rubellus* which suggests that the majority of bacterial data present were indeed from commensals.

Table 5.10: Low coverage contigs annotated as being of likely bacteria origin

Taxon of best BLAST match	Number of contigs
Gammaproteobacteria	1664
Alphaproteobacteria	368
Betaproteobacteria	101
Bacteroidetes	50
Bacilli	50
Actinobacteria (class)	24

5.2.1 Verminephrobacter

The genus *Verminephrobacter* (Phylum: Proteobacteria, Class: Betaproteobacteria, Order: Burkholderiales, Family: Comamonadaceae, see section 1.3.2) is an example of a known earthworm symbiont which may be ubiquitous across the earthworms, and is an active area of earthworm research. Before filtering, 37 gene objects were annotated as matching to 32 separate genes of *Verminephrobacter eiseniae* (Table 5.11). This represents the largest set of *L. rubellus Verminephrobacter* sequences documented to date. The *L. rubellus* genomic sequence data were mapped to the *V. eiseniae* genome to identify additional reads for assembly of the *L. rubellus Verminephrobacter*, but the number of reads mapped was very low. In addition, of the 37 contigs containing genes annotated as matching *Verminephrobacter* 32 have an estimated read coverage $\leq$7 suggesting that contigs from the *Verminephrobacter* genome are absent from the assembly due to coverage cutoff issues. The completed *V. eiseniae* genome, available at `http://genome.jgi-psf.org/verei/verei.home.html`, was analysed for clues as to the interaction between the symbionts and their hosts.

Table 5.11: *V. lumbricus* genes

Gene ID	Score	Name	Number
121607016	3e-30	putative inner membrane transmembrane protein	1
121607078	1e-65	GTP-binding protein, HSR1-related	3
121607480	9e-61	heat shock protein	1
121607598	7e-57	hypothetical protein	1
121607716	1e-122	extracellular solute-binding protein	1
121607993	3e-14	dTMP kinase	1
121608323	1e-47	monooxygenase, FAD-binding	1
121608335	8e-29	outer membrane protein	1
121608461	7e-94	acetyl-CoA carboxylase, biotin carboxylase	1
121608822	2e-29	serine/threonine-protein kinase	1
121608937	1e-17	hypothetical protein	1
121609007	1e-152	ABC transporter related	1
121609020	4e-39	beta-lactamase domain-containing protein	1
121609236	1e-46	oxidoreductase domain-containing protein	1
121609299	2e-42	hypothetical protein	1
121609325	4e-50	dihydrodipicolinate synthetase	1
121609523	1e-31	chaperonin GroEL	1
121609819	2e-92	salicylyl-CoA 5-hydroxylase	1
121609860	2e-31	glycine betaine/L-proline ABC transporter, ATPase subunit	2
121609947	2e-32	membrane-bound proton-translocating pyrophosphatase	1
121610172	1e-39	polyhydroxyalkonate synthesis repressor, PhaR	1
121610381	3e-26	binding-protein-dependent transport systems inner membrane component	1
121610507	1e-79	3-ketoacyl-(acyl-carrier-protein) reductase	1
121610576	6e-50	N-formimino-L-glutamate deiminase	1
121610745	2e-43	Tfp pilus assembly protein tip-associated adhesin PilY1-like protein	1
121610836	5e-56	dimethylallyltranstransferase	1
121610897	4e-63	alpha-ketoglutarate decarboxylase	2
121610985	4e-40	peptidase S1 and S6, chymotrypsin/Hap	1
121611305	3e-113	putative flavoprotein involved in K+ transport	1
121611352	1e-62	methylmalonate-semialdehyde dehydrogenase	1
121611467	2e-32	hypothetical protein	2
121611838	2e-86	aspartate kinase	1

If the relationship between earthworms and *Verminephrobacter* is truly symbiotic, one may expect to find some evidence of this within the putative functions assigned to gene products, most notably enzyme predictions, such as cases where the enzymes produced by both species combine to complete an otherwise incomplete pathway. Such instances may represent new pathways forming, or the loss of function from species A due to the provision by species B. To assess this hypothesis, enzymatic content of the *V. eiseniae* gene set was annotated using annot8r. Pathways with unique *V. eiseniae* contributions are in Tables 5.12 and 5.13. Figures 5.8 to 5.12 show predicted enzymes from both *L. rubellus* and *V. eiseniae* mapped onto the complete KEGG metabolic map. Regions which show pathway connections going directly from *L. rubellus* components to *V. eiseniae* components or vice versa suggest a possible area of symbiosis.

One case of this was identified in the Metabolism of Cofactors and Vitamins section (top right of Figure 5.10), notably the Vitamin B6 metabolism pathway. The well studied medicinal leech *Hirudo medicinalis* is known to have a vitamin deficiency which is overcome by provision by endosymbiotic bacteria, possibly *Aeromonas veronii* [63]. Figure 5.13 shows detail of this pathway and indicates that both species contribute to its completion. Of the two central enzymes, pyridoxal phosphatase (EC:3.1.3.74) is found only in *L. rubellus* while pyridoxamine 5'-phosphate oxidase (EC: 1.4.3.5) is only found in the symbiont. Both are required for the pathway to be functional. A very poorly-scoring match to EC:1.4.3.5 was identified in *L. rubellus* and the enzyme is present in *C. telata* and *H. robusta*, suggesting that the enzyme may have been lost in the earthworm.

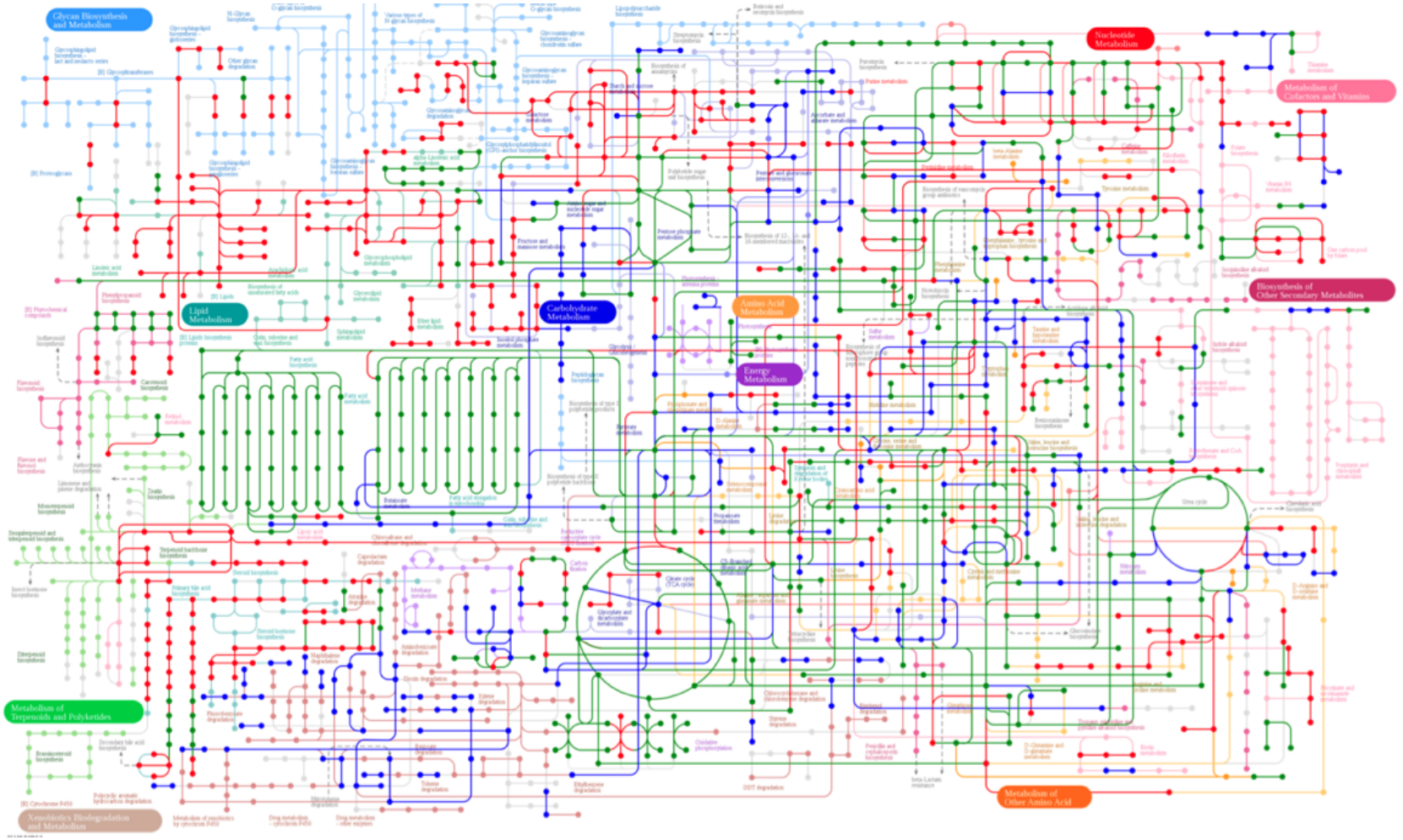

Figure 5.8: Complete KEGG metabolic pathways map for *L. rubellus* and *V. eiseniae*

Enzymes present in both are coloured green, *L. rubellus* only in red, *V. eiseniae* only in blue. Lightly coloured sections are missing completely from the predictions.

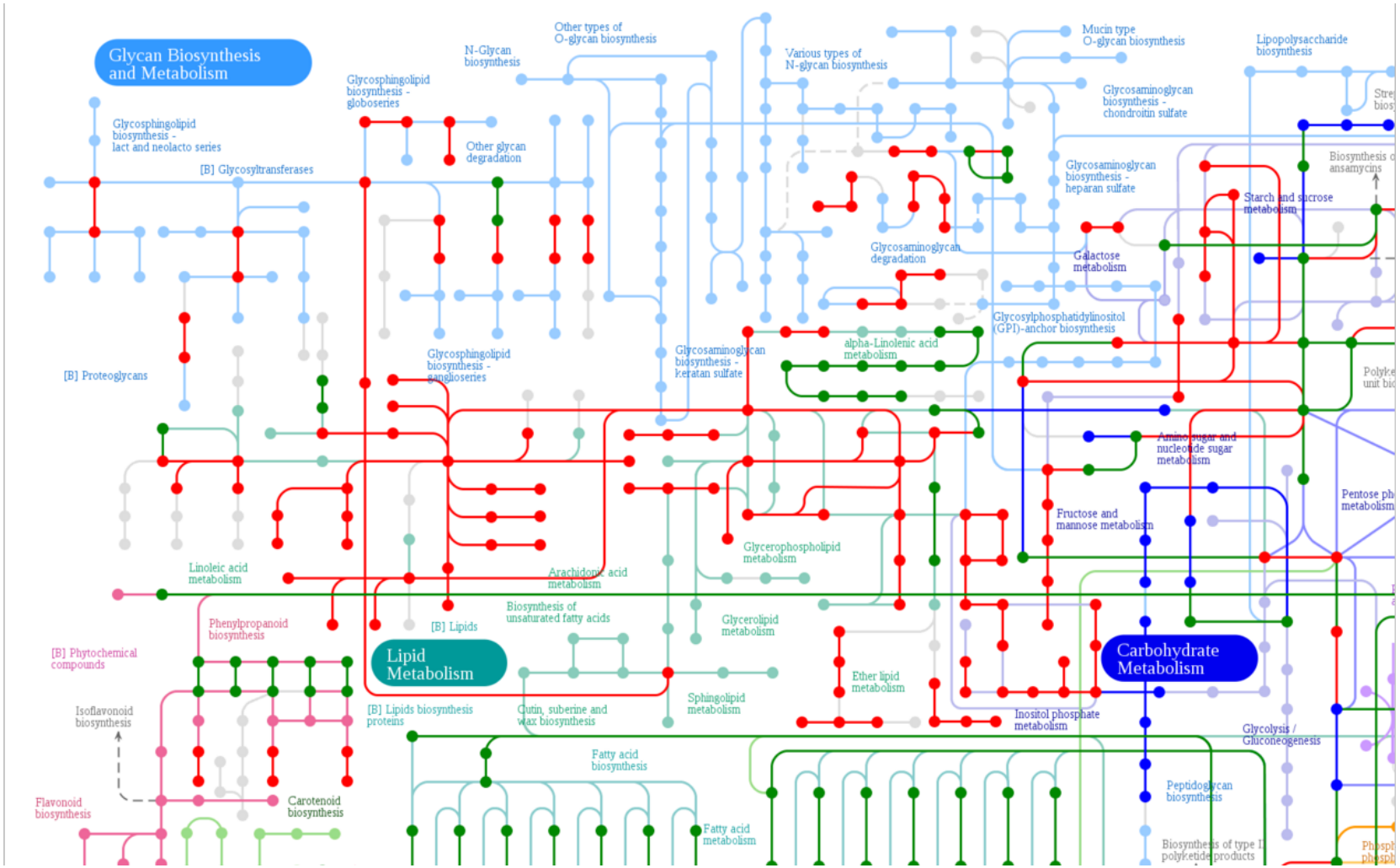

Figure 5.9: Top left segment of complete KEGG metabolic pathways map for *L. rubellus* and *V. eiseniae*

Enzymes present in both are coloured green, *L. rubellus* only in red, *V. eiseniae* only in blue. Lightly coloured sections are missing completely from the predictions.

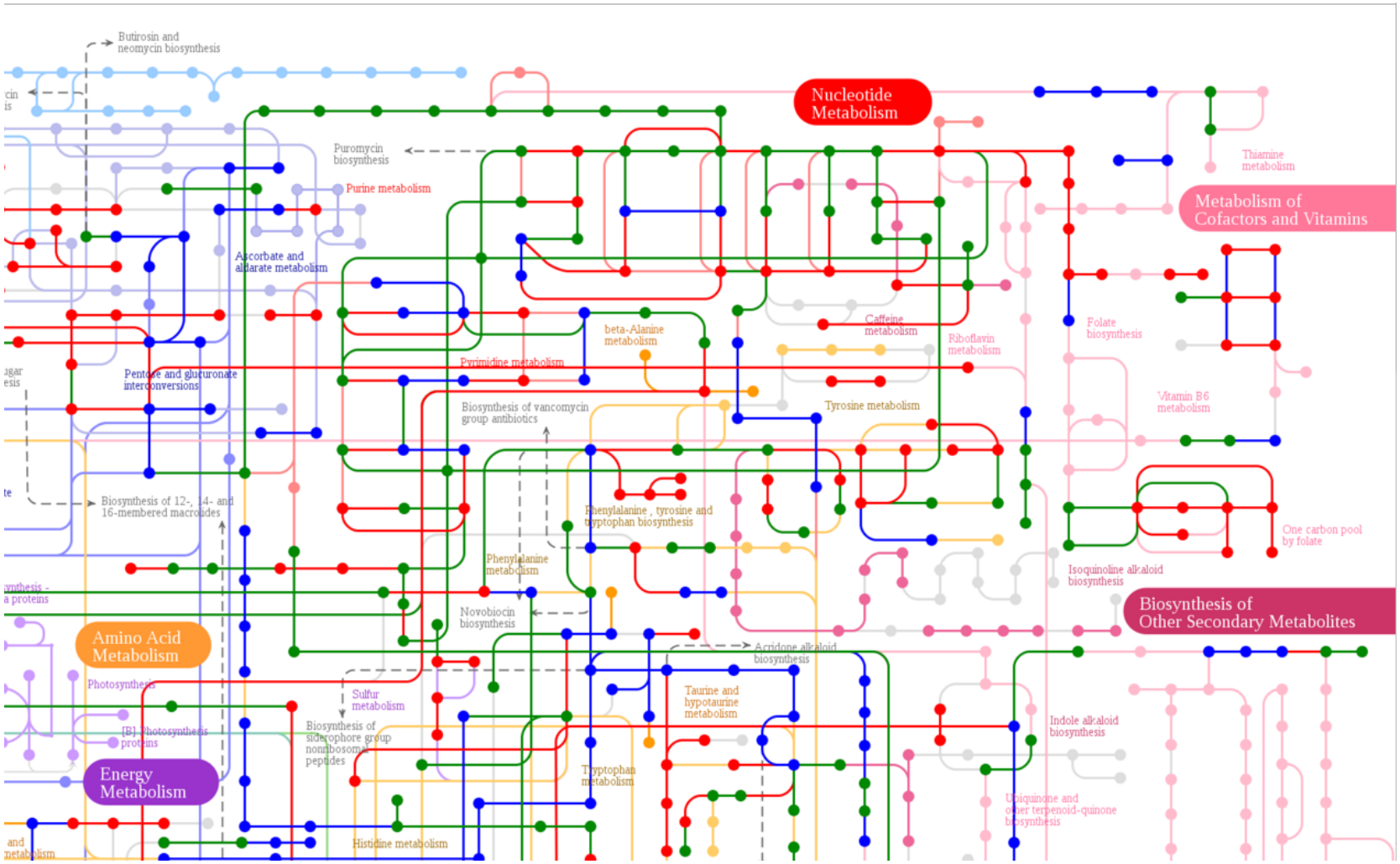

Figure 5.10: Top right segment of complete KEGG metabolic pathways map for *L. rubellus* and *V. eiseniae*

Enzymes present in both are coloured green, *L. rubellus* only in red, *V. eiseniae* only in blue. Lightly coloured sections are missing completely from the predictions.

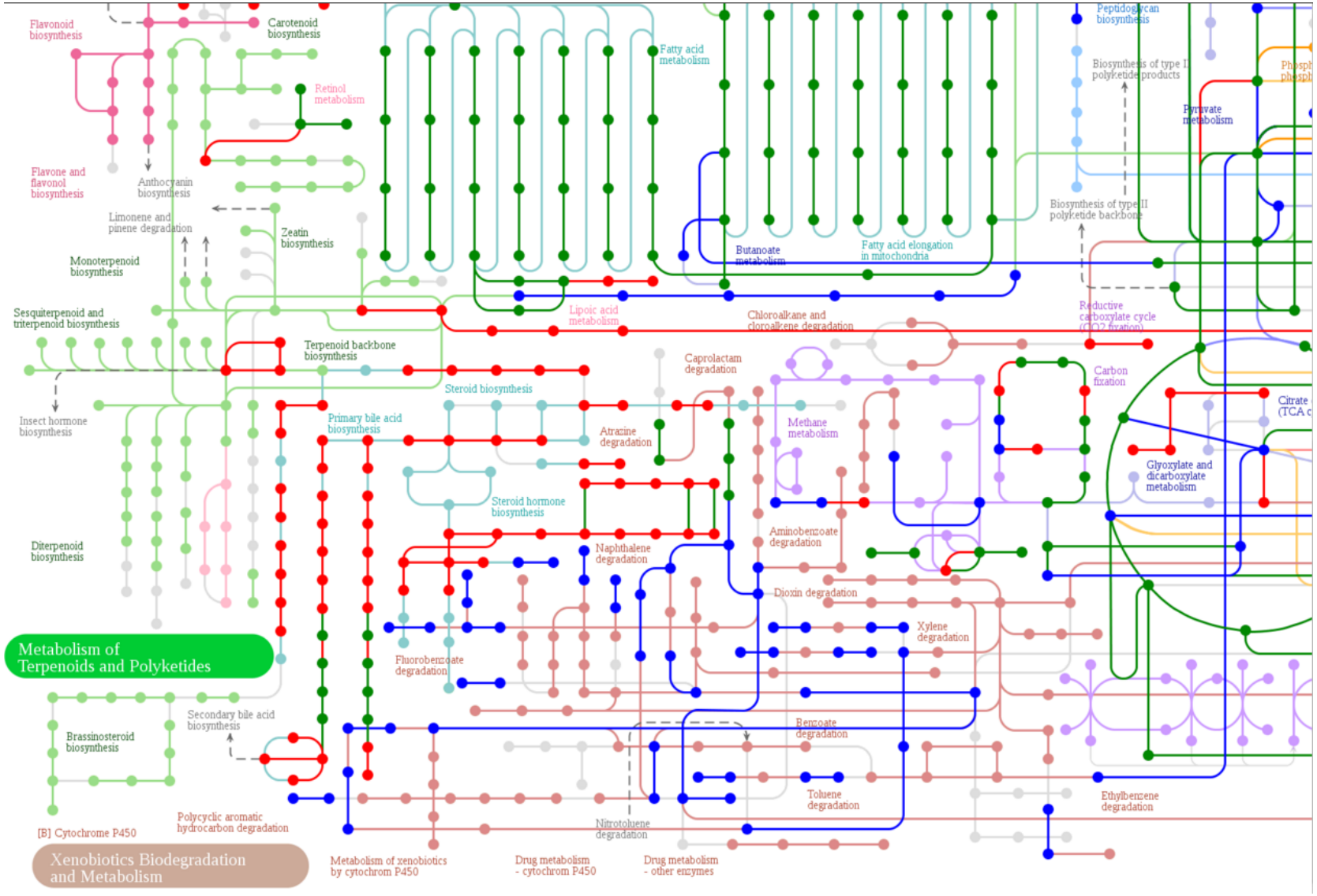

Figure 5.11: Bottom left segment of complete KEGG metabolic pathways map for *L. rubellus* and *V. eiseniae*

Enzymes present in both are coloured green, *L. rubellus* only in red, *V. eiseniae* only in blue. Lightly coloured sections are missing completely from the predictions.

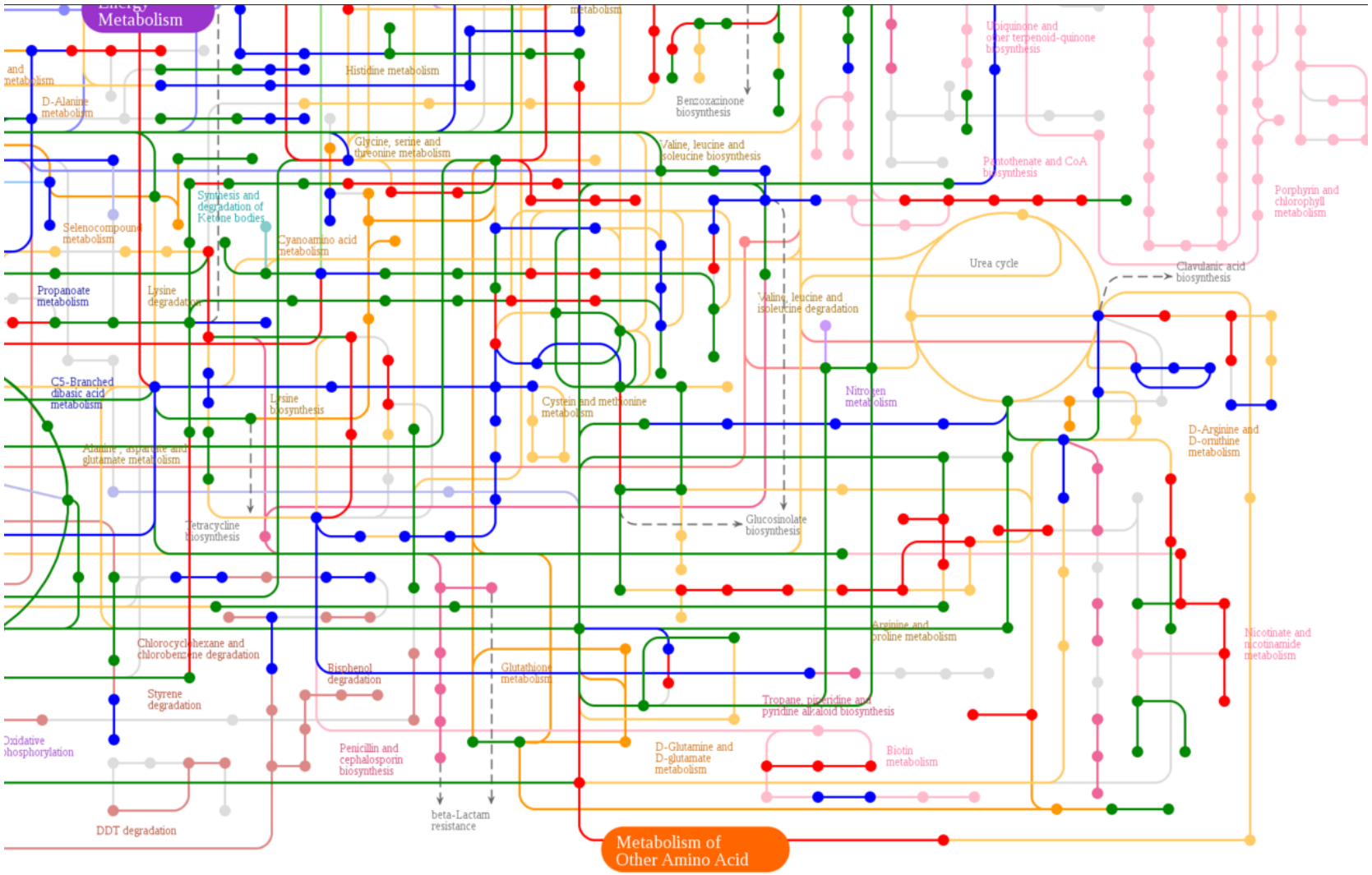

Figure 5.12: Bottom right segment of complete KEGG metabolic pathways map for *L. rubellus* and *V. eiseniae*

Enzymes present in both are coloured green, *L. rubellus* only in red, *V. eiseniae* only in blue. Lightly coloured sections are missing completely from the predictions.

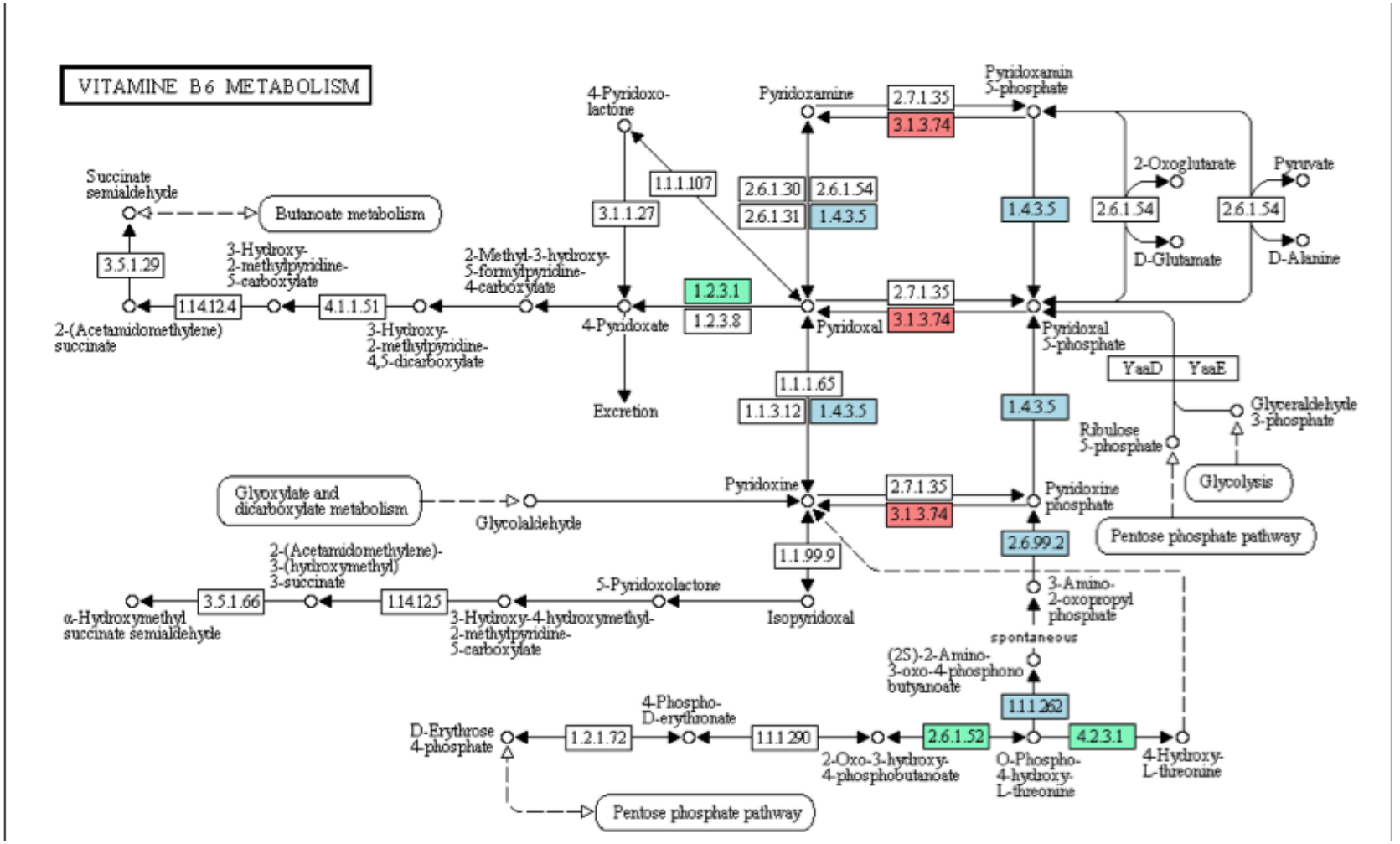

Figure 5.13: Vitamin B6 Metabolism in *L. rubellus* and *V. eiseniae*

Enzymes present in both species are coloured green, *L. rubellus* only in red, *V. eiseniae* only in blue and those missing from all are white.

Table 5.12: Possible symbiotic pathways - part 1

Pathway ID	Pathway Name	Total in all pathways	***L. rubellus / V. eiseniae***			**Other annelids**	
			L. rubellus only	*V. eiseniae* only	*L.rubellus & V. eiseniae*	*C. telata*	*H. robusta*
ec00010	Glycolysis / Gluconeogenesis	41	7	5	13	11	10
ec00030	Pentose phosphate pathway	37	6	9	9	12	11
ec00040	Pentose and glucuronate interconversions	56	8	4	1	11	9
ec00051	Fructose and mannose metabolism	65	14	2	4	15	13
ec00052	Galactose metabolism	37	11	4	2	13	10
ec00053	Ascorbate and aldarate metabolism	44	4	3	1	5	3
ec00062	Fatty acid elongation in mitochondria	8	1	1	3	2	2
ec00071	Fatty acid metabolism	29	3	5	11	8	8
ec00100	Steroid biosynthesis	50	20	5	0	22	22
ec00120	Primary bile acid biosynthesis	19	8	1	3	9	9
ec00130	Terpenoid-quinone biosynthesis	18	1	2	2	2	2
ec00140	C21-Steroid hormone metabolism	15	5	1	0	6	6
ec00190	Oxidative phosphorylation	12	2	2	8	4	3
ec00230	Purine metabolism	101	20	11	25	30	25
ec00240	Pyrimidine metabolism	64	12	4	16	15	12
ec00250	Alanine, aspartate and glutamate metabolism	43	3	4	21	7	7
ec00260	Glycine, serine and threonine metabolism	56	13	10	12	21	16
ec00271	Methionine metabolism	40	6	5	10	11	9
ec00272	Cysteine metabolism	26	1	3	7	4	3
ec00280	Valine, leucine and isoleucine degradation	34	4	3	20	7	7
ec00290	Valine, leucine and isoleucine biosynthesis	18	1	6	6	4	5
ec00300	Lysine biosynthesis	31	1	11	4	10	6
ec00310	Lysine degradation	54	7	4	7	11	11

Table 5.13: Possible symbiotic pathways - part 2

Pathway ID	Pathway Name	Total in all pathways	***L. rubellus / V. eiseniae***			**Other annelids**	
			L. rubellus only	*V. eiseniae* only	*L.rubellus & V. eiseniae*	*C. telata*	*H. robusta*
ec00330	Arginine and proline metabolism	97	11	14	19	24	17
ec00340	Histidine metabolism	37	1	6	7	3	2
ec00350	Tyrosine metabolism	69	7	4	8	11	11
ec00360	Phenylalanine metabolism	46	3	7	7	9	8
ec00361	gamma-Hexachlorocyclohexane degradation	27	2	5	0	6	5
ec00362	Benzoate degradation via hydroxylation	50	1	12	1	10	6
ec00380	Tryptophan metabolism	67	8	3	8	11	10
ec00400	Phenylalanine, tyrosine and tryptophan biosynthesis	33	1	13	5	10	3
ec00410	beta-Alanine metabolism	32	5	2	6	6	6
ec00430	Taurine and hypotaurine metabolism	17	2	5	1	7	5
ec00440	Aminophosphonate metabolism	11	4	2	0	6	4
ec00450	Selenoamino acid metabolism	21	3	1	10	4	4
ec00460	Cyanoamino acid metabolism	20	1	2	4	3	3
ec00471	D-Glutamine and D-glutamate metabolism	13	1	3	1	3	2
ec00480	Glutathione metabolism	40	10	1	8	11	10
ec00500	Starch and sucrose metabolism	71	19	3	2	21	20
ec00520	Amino sugar and nucleotide sugar metabolism	94	20	8	7	26	25
ec00521	Streptomycin biosynthesis	14	3	3	2	5	3
ec00561	Glycerolipid metabolism	36	10	3	2	12	11
ec00562	Inositol phosphate metabolism	40	17	1	1	18	16
ec00564	Glycerophospholipid metabolism	50	24	2	1	25	24
ec00620	Pyruvate metabolism	64	6	11	13	17	13
ec00630	Glyoxylate and dicarboxylate metabolism	58	5	7	7	12	9
ec00640	Propanoate metabolism	47	4	7	12	9	8

Table 5.14: Possible symbiotic pathways - part 3

Pathway ID	Pathway Name	Total in all pathways	***L. rubellus / V. eiseniae***			**Other annelids**	
			L. rubellus only	*V. eiseniae* only	*L.rubellus & V. eiseniae*	*C. telata*	*H. robusta*
ec00643	Styrene degradation	21	2	5	1	7	6
ec00650	Butanoate metabolism	52	5	8	15	11	10
ec00670	One carbon pool by folate	24	5	2	5	7	5
ec00680	Methane metabolism	33	4	7	2	10	9
ec00710	Carbon fixation in photosynthetic organisms	25	6	3	7	9	7
ec00720	Reductive carboxylate cycle (CO2 fixation)	13	1	2	7	3	2
ec00740	Riboflavin metabolism	15	5	1	1	6	4
ec00750	Vitamin B6 metabolism	26	1	4	4	4	2
ec00760	Nicotinate and nicotinamide metabolism	47	9	1	5	10	10
ec00770	Pantothenate and CoA biosynthesis	28	4	5	6	4	4
ec00780	Biotin metabolism	12	1	1		2	2
ec00790	Folate biosynthesis	25	3	2		4	4
ec00830	Retinol metabolism	18	5	1	1	6	5
ec00860	Porphyrin and chlorophyll metabolism	66	7	4	4	10	10
ec00910	Nitrogen metabolism	57	6	9	10	14	13
ec00920	Sulfur metabolism	30	7	2	5	9	7
ec00950	Alkaloid biosynthesis I	38	2	1	4	3	3
ec00960	Alkaloid biosynthesis II	15	2	1		3	3
ec00970	Aminoacyl-tRNA biosynthesis	31	2	1	18	3	3
ec00980	Metabolism of xenobiotics by cytochrome P450	7	5	1	1	6	5
ec00982	Drug metabolism - cytochrome P450	9	5	1	2	6	6
ec00983	Drug metabolism - other enzymes	22	10	1	4	11	9

Pathways were identified where both *L. rubellus* and *V. eiseniae* contributed mutually exclusive enzymes. Sixty-nine pathways were found to match this criteria (Tables 5.12 to 5.14). As previously discussed, the current proposed function of the *Verminephrobacter* is to enhance nitrogen retention [101]. The Nitrogen metabolism pathway (ko00910) had mutually exclusive contributions from both species (Figure 5.14). The dissimilatory reduction of nitrate to nitrite and ammonia on the left hand side of the figure is usually the work of bacteria. *L. rubellus*-only predictions for two nitrate reductase enzymes, EC:1.7.1.1 and EC:1.7.1.2 were likely false positives due to the similarity to cytochrome B5. Two *V. eiseniae*-only enzymes, nitrate reductase (EC:1.7.99.4) and nitrite reductase (EC:1.7.1.4), provide a vital step in the reduction of nitrogen to ammonia but the there was no evidence for the more useful denitrification pathway (the reduction of nitrate to nitrogenous gas and nitrogen). Many studies have identified denitrifying bacteria in the gut of earthworms [40][82][71][107][177] as well as high levels of nitrogenous gas, and the abundance of denitrifying bacteria can be up to three times higher in the guts of earthworms than in the surrounding preingested soil.

A link between *Verminephrobacter* and nitrogen reduction is supported by their location in the nephridia, used for the excretion of nitrogenous waste. It is hypothesised that the reduction of the nephridia in the marine oligochaete *Olavius algarvensis* has led to a symbiotic relationship replacing the entire function of the nephridia [176]. However, from the evidence here, it would appear that products of denitrification are likely to be provided by other bacteria and not *Verminephrobacter*.

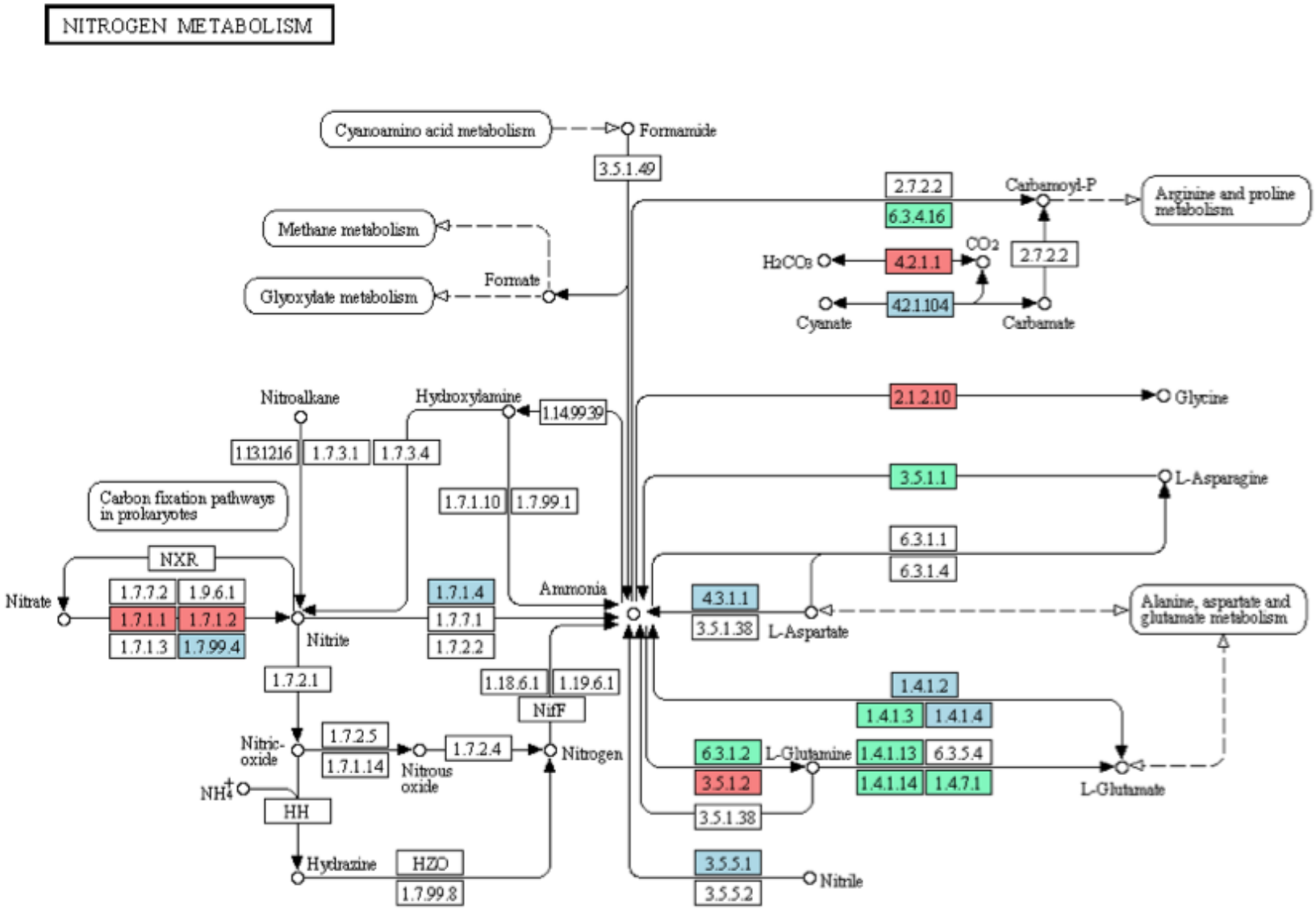

Figure 5.14: Nitrogen Metabolism in *L. rubellus* and *V. eiseniae*

Enzymes present in both species are coloured green, *L. rubellus* only in red, *V. eiseniae* only in blue and those missing from all are white.

There was a substantial contribution to Xenobiotics Biodegradation and Metabolism (Figure 5.11) from the bacteria. The *Verminephrobacter* may help with the degradation of soil toxins including organic herbicides such as atrazine. This process has previously been suggested to be facilitated by bacteria [84] [18].

Tables 5.12 to 5.14 also list the numbers of enzymes for *C. telata* and *H. robusta*. Interestingly, *L. rubellus* has more enzymes predicted than either of the other annelids in 43 of the 69 pathways. This may suggest that *L. rubellus* is more metabolically active than the other annelids with respect to these pathways which is interesting as these pathways have been selected on the basis of a shared contribution with *V. eiseniae*. Conversely, pathways such as ec00362 (Benzoate degradation via hydroxylation) which have far fewer enzymes predicted by *L. rubellus* compared to the other annelids may highlight instances of 'true' symbiosis.

5.3 Conclusions and prospects

In this thesis I have presented assembly, annotation and preliminary analyses of the genome of *L. rubellus*, a temperate earthworm. This work represents the first high quality draft [24] of an earthworm genome. In addition, the first draft of the complete proteome has been produced as both a transcriptome and a set of gene predictions. The transcriptome has also been used to further scaffold the genome as part of a novel approach producing significant improvements in the assembly (`https://github.com/elswob/SCUBAT`). The genome assembly would benefit from additional data in the form of multiple large insert mate-pair libraries or an equivalent (3rd generation sequencing technology) to further scaffold the assembly.

The MAKER2 pipeline has performed well and produced many reliable gene predictions. However, extra gene predictors were also incorporated to improve the gene set. This too would benefit from a more contiguous assembly as the accuracy and quality of annotation is greatly enhanced by genomic contiguity. I have created a new community annotation mechanism which allows consortium members to upload their manual annotations.

This data is currently available to the earthworm community via www.earthworms.org and is being accessed daily.

These annotations have enabled detailed analysis of the proteome including comparative analysis with two other annelids and a symbiont. The two methods of comparative genomics identified key areas which appear to be either highly enhanced or unique to *L. rubellus*, some of which may be related to the origin of the sequenced worm and add to the mounting evidence for using earthworms as bioindicators of soil. Present in significantly higher number compared to the other two annelids were the cytochrome P450 enzyme unspecific monooxygenase (EC:1.14.14.1) which is often linked to the catalysis of xenobiotics, drugs and toxic chemicals, and ATPase transport proteins, in particular Ca^{2+} which have possible links to lead adaptation and metal homeostasis. The enzyme Mannan endo-1,4-β-mannosidase (EC:3.2.1.78) has been retained by *L. rubellus* for lifestyle reasons and it remains to be seen how many other earthworms have continued using this enzyme.

This work will hopefully act as a stepping stone from which many areas of earthworm and annelid biology will develop. It is, however, only a first draft, and will need further work in both assembly and annotation to fully release its potential.

Appendices

Appendix A

PostGreSQL Database Tables

Table A.1: contig

Field Name	Contents
contig_id	The contig/scaffold ID
length	Length of the contig sequence
number_reads	Count of the reads mapped back to the contigs/scaffolds using bowtie version 0.12.5 [90]
coverage	Calculated using equation (1.2) replacing genome length with contig length
seq	The nucleotide contig/scaffold sequence
gc	The GC content of the contig/scaffold

Table A.2: gene_info

Field Name	Contents
id	The unique ID for the gene object (a_ = AUGUSTUS, k_ = MAKER2, p_ = protein2genome, m_ = manual)
source	The source of the gene object as above
contig	The contig/scaffold ID
start	The base on the contig/scaffold the gene starts
stop	The base on the contig/scaffold the gene stops
nuc	The nucleotide sequence of the coding region
pept	The peptide sequence of the coding region
intron	The number of introns in the gene object
coverage	The coverage of the coding region calculated from subsets of bowtie mapping data
rep	0 or 1 to signify absence/presence in the filtered gene set

Table A.3: gene_anno

Field Name	Contents
id	The unique ID for the gene object (a_ = AUGUSTUS, k_ = MAKER2, p_ = protein2genome, m_ = manual)
anno_db	The source of the annotation, e.g. ProfileScan
anno_id	The ID for the annotation, e.g. IPR003961
anno_start	The start base of the annotation on the gene object
anno_stop	The end base of the annotation on the gene object
score	The score for the annotation
descr	A description of the annotation

Table A.4: ncrna

Field Name	Contents
contig	The contig/scaffold ID
id	The unique ID for the ncRNA
number	If more than one per contig/scaffold then a number is assigned
start	The base on the contig/scaffold the ncRNA starts
stop	The base on the contig/scaffold the ncRNA stops
descr	A description of the ncRNA
score	A score for that ncRNA predictions

Table A.5: pathway_map

Field Name	Contents
id	The KEGG pathway ID, e.g. ec00010
ec	Enzyme commission ID, .e.g. 1.1.1.1

Table A.6: pathway_id2name

Field Name	Contents
id	The KEGG pathway ID, e.g. ec00010
name	The name of the pathway, e.g. Glycolysis / Gluconeogenesis

Table A.7: ec2description

Field Name	Contents
ec	The KEGG pathway ID, e.g. ec00010
name	The name of the pathway, e.g. Glycolysis / Gluconeogenesis

Table A.8: interpro_key

Field Name	Contents
dom_id	The domain ID, e.g. G3DSA:1.10.10.10
description	The description of the domain, e.g. Winged helix-turn-helix transcription repressor DNA-binding
database	The database the domain ID is from, e,g. Gene3D
ipr_id	The InterProScan ID, e.g. IPR011991
short_desc	A shorter version of description, e.g. WHTH_trsnscrt_rep_DNA-bd

Appendix B

Combining gene models

```
#!/usr/bin/perl -w

use Pg;
use strict;
use warnings;
use Data::Dumper;


my ($contig,$seq,$dbname,$dbconn,$sqlcom,$dbres);

$dbname = "genome_23_05_11";
unless ($dbname) {
        print "which database do you want to use?  ";
        $dbname=<STDIN>;
}


$dbconn = Pg::connectdb("dbname=$dbname");              #connect to the database

#------------------------------------------------------ process contigs with no allelic contigs
     ------------------------------------------------------------

print "--------------getting non allelic contigs----------------\n";
$sqlcom="select * from proteins where length(nuc) > 80 and allele = 'single' and source != 'e2g' order by contig;";
#print "$sqlcom\n";
$dbres = $dbconn->exec($sqlcom);
my $row=0;
my $rowmax= $dbres->ntuples;
print "There are $rowmax results\n";
my @old_result=();
my %multiple_annotation;
my %single_annotation;
my %annotation;
my $count=1;
my $first_man_count=0;
while ($row<$rowmax){
        my @result = $dbres->fetchrow;
        $annotation{$result[2]}{$count}=[@result];
```

```
        if ($result[1] eq 'manual'){
                $first_man_count++;
        }
        $count++;
        $row++;
}

open ANNO_DUMP,">annotation_dump.txt" or die;
print ANNO_DUMP Dumper( \%annotation );


my $start="";
my $end="";
my $old_end="";
my $gene_count;
my $single_count=0;
my $multiple_count=0;
my $multiple_gene_count=0;
my $total_count=0;
my %final_multi_gene;
my %final_single_gene;
my $early_man_count=0;
print "sorting␣them␣into␣gene␣groups...\n";
for my $key1 (sort keys %annotation){
        #pick out multiple annotations
        if (keys %{$annotation{$key1}}>1){
                #print "--- multiple $key1 ---\n";
                $start="";
                $end="";
                $old_end=1000000;
                $gene_count=1;
                #key part - sort by start point on contig!
                for my $key2 (sort {$annotation{$key1}{$a}[3]<=>$annotation{$key1}{$b}[3]} keys %{$annotation{$key1}}){
                        $start = $annotation{$key1}{$key2}[3];
                        $end = $annotation{$key1}{$key2}[4];
                        if ($start > $old_end){
                                if ($annotation{$key1}{$key2}[1] eq 'manual'){
                                        $early_man_count++;
                                }
                                $gene_count++;
                                $total_count++;
                                #print "--- gene $gene_count ---\n$annotation{$key1}{$key2}[0]\t$annotation{$key1}{
                                      $key2}[1]\t$annotation{$key1}{$key2}[2]\t$annotation{$key1}{$key2}[3]\
                                      t$annotation{$key1}{$key2}[4]\n";
                                $multiple_gene_count++;
                                #add to final gene set
                                $final_multi_gene{$key1}{$gene_count}{$total_count}=[@{$annotation{$key1}{$key2}}];
                        }else{
                                $total_count++;
                                if ($annotation{$key1}{$key2}[1] eq 'manual'){
                                        $early_man_count++;
                                        $gene_count++;
                                }
                                #print "--- gene $gene_count ---\n$annotation{$key1}{$key2}[0]\t$annotation{$key1}{
                                      $key2}[1]\t$annotation{$key1}{$key2}[2]\t$annotation{$key1}{$key2}[3]\
                                      t$annotation{$key1}{$key2}[4]\n";
                                $multiple_count++;
                                #add to final gene set
                                $final_multi_gene{$key1}{$gene_count}{$total_count}=[@{$annotation{$key1}{$key2}}];
```

```
                        }
                        #not all stops go up in order, check for this
                        if ($end > $old_end || $old_end == 1000000){
                                $old_end = $end;
                        }
                }
                #print "\n";
        }else{
                #print "--- single $key1 ---\n";
                for my $key2 (sort {$a<=>$b} keys %{$annotation{$key1}}){
                        if ($annotation{$key1}{$key2}[1] eq 'manual'){
                                $early_man_count++;
                        }
                        $single_count++;
                        $final_single_gene{$key1}{$total_count}=[@{$annotation{$key1}{$key2}}];
                        $total_count++;
                        #print "$annotation{$key1}{$key2}[0]\t$annotation{$key1}{$key2}[1]\t$annotation{$key1}{$key2}[2]\t$annotation{$key1}{$key2}[3]\t$annotation{$key1}{$key2}[4]\n";
                }
        }
}
open DUMP_NA,">dump_non_annelic_final_hash.txt";
print DUMP_NA Dumper( \%final_multi_gene );
print "choosing the best predictions...\n";
my %picker;
open M,">non_allelic_multi_annotation.fa";
open MA,">non_allelic_multi_annotation.aa";
open ALLF,">all_genes.fa";
open ALLA,">all_genes.aa";
open CONT,">non_allelic_contigs.txt";
my $all_gene = 1;
my $count_man=0;
for my $key1 (sort keys %final_multi_gene){
                #print "$key1 -> ";
                for my $key2 (sort {$a<=>$b} keys %{$final_multi_gene{$key1}}){
                        labelbreak: while($key2){
                        %picker=();
                        for my $key3 (sort {$a<=>$b} keys %{$final_multi_gene{$key1}{$key2}}){
                                $picker{$key3}=$final_multi_gene{$key1}{$key2}{$key3}[1];
                        }
                        #print the manual ones separately as all of them need to come out regardless (some of them are nested within one larger prediction and may appear to be one gene)
                        for my $pick_key (keys %picker){
                                if ($picker{$pick_key} eq 'manual'){
                                        $count_man++;
                                        print "non_allelic_$count_man_manual_$final_multi_gene{$key1}{$key2}{$pick_key}[0]\n";
                                        print M ">$all_gene"."_na_multi_$final_multi_gene{$key1}{$key2}{$pick_key}[2]_$final_multi_gene{$key1}{$key2}{$pick_key}[8]_$final_multi_gene{$key1}{$key2}{$pick_key}[0]\n$final_multi_gene{$key1}{$key2}{$pick_key}[6]\n";
                                        print MA ">$all_gene"."_na_multi_$final_multi_gene{$key1}{$key2}{$pick_key}[2]_$final_multi_gene{$key1}{$key2}{$pick_key}[8]_$final_multi_gene{$key1}{$key2}{$pick_key}[0]\n$final_multi_gene{$key1}{$key2}{$pick_key}[7]\n";
                                        print ALLF ">$final_multi_gene{$key1}{$key2}{$pick_key}[0]\n$final_multi_gene{$key1}{$key2}{$pick_key}[6]\n";
                                        print ALLA ">$final_multi_gene{$key1}{$key2}{$pick_key}[0]\n$final_multi_gene{$key1}{$key2}{$pick_key}[7]\n";
                                        print CONT "$final_multi_gene{$key1}{$key2}{$pick_key}[2]\n";
                                        $all_gene++;
```

```
143                                     last labelbreak;
144                                     delete($picker{$pick_key});
145                             }
146                     }for my $pick_key (keys %picker){
147                             if ($picker{$pick_key} eq 'maker'){
148                                     print M ">$all_gene"."_na_multi_$final_multi_gene{$key1}{$key2}{$pick_key}[2]
                                              _$final_multi_gene{$key1}{$key2}{$pick_key}[8]_$final_multi_gene{$key1}{
                                              $key2}{$pick_key}[0]\n$final_multi_gene{$key1}{$key2}{$pick_key}[6]\n";
149                                     print MA ">$all_gene"."_na_multi_$final_multi_gene{$key1}{$key2}{$pick_key}[2]
                                              _$final_multi_gene{$key1}{$key2}{$pick_key}[8]_$final_multi_gene{$key1}{
                                              $key2}{$pick_key}[0]\n$final_multi_gene{$key1}{$key2}{$pick_key}[7]\n";
150                                     print ALLF ">$final_multi_gene{$key1}{$key2}{$pick_key}[0]\n$final_multi_gene{
                                              $key1}{$key2}{$pick_key}[6]\n";
151                                     print ALLA ">$final_multi_gene{$key1}{$key2}{$pick_key}[0]\n$final_multi_gene{
                                              $key1}{$key2}{$pick_key}[7]\n";
152                                     print CONT "$final_multi_gene{$key1}{$key2}{$pick_key}[2]\n";
153                                     $all_gene++;
154                                     last labelbreak;
155                             }
156                     }for my $pick_key (keys %picker){
157                             if ($picker{$pick_key} eq 'augustus'){
158                                     #if ($final_multi_gene{$key1}{$key2}{$pick_key}[2] eq 'contig_9937'){
159                                     #       print "found it -> $key1 - $key2 - $pick_key - $picker{$pick_key} \n";
160                                             #print M ">$all_gene"."_na_multi_$final_multi_gene{$key1}{$key2}{
                                                      $pick_key}[2]_$final_multi_gene{$key1}{$key2}{$pick_key}[8]
                                                      _$final_multi_gene{$key1}{$key2}{$pick_key}[0]\n$final_multi_gene
                                                      {$key1}{$key2}{$pick_key}[6]\n";
161
162                                     #}
163                                     print M ">$all_gene"."_na_multi_$final_multi_gene{$key1}{$key2}{$pick_key}[2]
                                              _$final_multi_gene{$key1}{$key2}{$pick_key}[8]_$final_multi_gene{$key1}{
                                              $key2}{$pick_key}[0]\n$final_multi_gene{$key1}{$key2}{$pick_key}[6]\n";
164                                     print MA ">$all_gene"."_na_multi_$final_multi_gene{$key1}{$key2}{$pick_key}[2]
                                              _$final_multi_gene{$key1}{$key2}{$pick_key}[8]_$final_multi_gene{$key1}{
                                              $key2}{$pick_key}[0]\n$final_multi_gene{$key1}{$key2}{$pick_key}[7]\n";
165                                     print ALLF ">$final_multi_gene{$key1}{$key2}{$pick_key}[0]\n$final_multi_gene{
                                              $key1}{$key2}{$pick_key}[6]\n";
166                                     print ALLA ">$final_multi_gene{$key1}{$key2}{$pick_key}[0]\n$final_multi_gene{
                                              $key1}{$key2}{$pick_key}[7]\n";
167                                     print CONT "$final_multi_gene{$key1}{$key2}{$pick_key}[2]\n";
168                                     $all_gene++;
169                                     last labelbreak;
170                             }
171                     }for my $pick_key (keys %picker){
172                             if ($picker{$pick_key} eq 'p2g'){
173                                     print M ">$all_gene"."_na_multi_$final_multi_gene{$key1}{$key2}{$pick_key}[2]
                                              _$final_multi_gene{$key1}{$key2}{$pick_key}[8]_$final_multi_gene{$key1}{
                                              $key2}{$pick_key}[0]\n$final_multi_gene{$key1}{$key2}{$pick_key}[6]\n";
174                                     print MA ">$all_gene"."_na_multi_$final_multi_gene{$key1}{$key2}{$pick_key}[2]
                                              _$final_multi_gene{$key1}{$key2}{$pick_key}[8]_$final_multi_gene{$key1}{
                                              $key2}{$pick_key}[0]\n$final_multi_gene{$key1}{$key2}{$pick_key}[7]\n";
175                                     print ALLF ">$final_multi_gene{$key1}{$key2}{$pick_key}[0]\n$final_multi_gene{
                                              $key1}{$key2}{$pick_key}[6]\n";
176                                     print ALLA ">$final_multi_gene{$key1}{$key2}{$pick_key}[0]\n$final_multi_gene{
                                              $key1}{$key2}{$pick_key}[7]\n";
177                                     print CONT "$final_multi_gene{$key1}{$key2}{$pick_key}[2]\n";
178                                     $all_gene++;
179                                     last labelbreak;
180                             }
```

```
                        }
                }
        }
}
close M;close MA;
open S,">non_allelic_single_annotation.fa";
open SA,">non_allelic_single_annotation.aa";
$count_man=0;
for my $key1 (sort keys %final_single_gene){
        #print $key1."\n";
        for my $key2 (sort {$a<=>$b} keys %{$final_single_gene{$key1}}){
                if ($final_single_gene{$key1}{$key2}[1] eq 'manual'){
                        $count_man++;
                        print "single non_allelic_manual $count_man $final_single_gene{$key1}{$key2}[0]\n";
                }
                #print "single - $final_single_gene{$key1}{$key2}[0]\t$final_single_gene{$key1}{$key2}[1]\
                      t$final_single_gene{$key1}{$key2}[2]\t$final_single_gene{$key1}{$key2}[3]\t$final_single_gene{
                      $key1}{$key2}[4]\t\n";
                print S ">$all_gene"."_na_single_$final_single_gene{$key1}{$key2}[2]_$final_single_gene{$key1}{$key2
                      }[8]_$final_single_gene{$key1}{$key2}[0]\n$final_single_gene{$key1}{$key2}[6]\n";
                print SA ">$all_gene"."_na_single_$final_single_gene{$key1}{$key2}[2]_$final_single_gene{$key1}{$key2
                      }[8]_$final_single_gene{$key1}{$key2}[0]\n$final_single_gene{$key1}{$key2}[7]\n";
                print ALLF ">$final_single_gene{$key1}{$key2}[0]\n$final_single_gene{$key1}{$key2}[6]\n";
                print ALLA ">$final_single_gene{$key1}{$key2}[0]\n$final_single_gene{$key1}{$key2}[7]\n";
                print CONT "$final_single_gene{$key1}{$key2}[2]\n";
                $all_gene++;
        }
}
close S; close SA; close CONT;

print "$single_count single annotation genes\n";
print "$multiple_count multiple annotation genes\n";
print "$multiple_gene_count multiple gene annotation genes\n";
print "total multi genes = ".(keys %final_multi_gene)."\n";
print "total single genes = ".(keys %final_single_gene)."\n";
print "$first_man_count non - allelic manual genes were read in\n";
print "$early_man_count non - allelic manual genes were processed\n";
$first_man_count=0; $early_man_count=0;


#=cut
#=pod

#--------------- get allelic contig hits--------------------



print "\n--------------- getting the allelic contig data ----------------\n";

print "getting contig lengths...\n";
my %contig_length;
my $sql_len="select contig_id,seq from contig;";
my $dbres_len = $dbconn->exec($sql_len);
my $row_len=0;
my $rowmax_len= $dbres_len->ntuples;

while ($row_len<$rowmax_len){
        my @result_len = $dbres_len->fetchrow;
        $contig_length{$result_len[0]}=length($result_len[1]);
```

```
        $row_len++;
}

my %alleles;
my $sqlcom2="select * from proteins where allele ~ 'contig' and length(nuc) > 80 and source != 'e2g' order by contig;";
my $dbres2 = $dbconn->exec($sqlcom2);
my $row2=0;
my $rowmax2= $dbres2->ntuples;
$count=1;
print "There are $rowmax2 results\n";

while ($row2<$rowmax2){
        #create hash of contig id and allelic contigs as string
        my @result2 = $dbres2->fetchrow;
        if ($result2[1] eq 'manual'){
                $first_man_count++;
        }
        $alleles{$result2[2]}=$result2[5];
        $row2++;
}

print "finding allelic representatives...\n";
my %allele_rep;
my %allele_check;
my $count_groups=1;
my $count_rep=1;
my $select="";
for my $key1 (sort keys %alleles){
        #printf("\r%9d groups processed",$count_groups);
        $count_groups++;
        labelbreak2: while($key1){

        my @split = split(/ /,$alleles{$key1});
        foreach(@split){
                s/\s+//;
                if ( exists $allele_check{$_}){
                        #print "already done!\n";
                        last labelbreak2;
                }
        }
        #find the longest one
        my $length=0;
        foreach(@split){
                s/\s+//;
                if (exists $contig_length{$_}){
                        #print "$_ : length = $contig_length{$_} -";
                        if ($contig_length{$_} > $length){
                                $select = $_;
                                #print "select = $select\n";
                                $allele_check{$_}="";
                                $length = $contig_length{$_};
                        }
                }
        }
        $allele_rep{$select}="";
        #print "chosen = $select\n";
        }
}
print "number of representatives = ".keys(%allele_rep)."\n";
```

```
295 open REP_DUMP, ">allele_rep_dump.txt";
296 print REP_DUMP Dumper(\%allele_rep);
297 $row2=0;
298 my %allele_groups;
299 my $allele_count=0;
300 $dbres2 = $dbconn->exec($sqlcom2);
301 while ($row2<$rowmax2){
302         my @result2 = $dbres2->fetchrow;
303         if (exists $allele_rep{$result2[2]}){
304                 $allele_groups{$result2[2]}{$allele_count}=[@result2];
305         }
306         $row2++;
307         $allele_count++;
308 }
309
310 open DUMP_A,">allele_groups_hash.txt";
311 print DUMP_A Dumper( \%allele_groups );
312
313 print "sorting␣groups␣into␣gene␣groups...\n";
314 $start="";
315 $end="";
316 $old_end="";
317 $gene_count=0;
318 $single_count=0;
319 $multiple_count=0;
320 $multiple_gene_count=0;
321 $total_count=0;
322 %final_multi_gene=();
323 %final_single_gene=();
324 for my $key1 (sort keys %allele_groups){
325         if (keys %{$allele_groups{$key1}}>1){
326                 $count_groups++;
327                 #pick out multiple annotations
328                 #print "--- allele group $key1 ---\n";
329                 $start="";
330                 $end="";
331                 $old_end=1000000;
332                 $gene_count=1;
333                 #key part - sort by start point on contig!
334                 for my $key2 (sort {$allele_groups{$key1}{$a}[3]<=>$allele_groups{$key1}{$b}[3]} keys %{$allele_groups{
                        $key1}}){
335                         $start = $allele_groups{$key1}{$key2}[3];
336                         $end = $allele_groups{$key1}{$key2}[4];
337                         if ($start > $old_end ){
338                                 if ($allele_groups{$key1}{$key2}[1] eq 'manual'){
339                                         $early_man_count++;
340                                 }
341                                 $gene_count++;
342                                 $total_count++;
343                                 #print "--- gene $gene_count ---\n$allele_groups{$key1}{$key2}[0]\t$allele_groups{$key1
                                        }{$key2}[1]\t$allele_groups{$key1}{$key2}[2]\t$allele_groups{$key1}{$key2}[3]\
                                        t$allele_groups{$key1}{$key2}[4]\n";
344                                 $multiple_gene_count++;
345                                 #add to final gene set
346                                 $final_multi_gene{$key1}{$gene_count}{$total_count}=[@{$allele_groups{$key1}{$key2}}];
347                         }else{
348                                 $total_count++;
349                                 if ($allele_groups{$key1}{$key2}[1] eq 'manual'){
350                                         $early_man_count++;
```

```
                                        $gene_count++;
                                }
                                #print "--- gene $gene_count ---\n$allele_groups{$key1}{$key2}[0]\t$allele_groups{$key1
                                      }{$key2}[1]\t$allele_groups{$key1}{$key2}[2]\t$allele_groups{$key1}{$key2}[3]\
                                      t$allele_groups{$key1}{$key2}[4]\n";
                                $multiple_count++;
                                #add to final gene set
                                $final_multi_gene{$key1}{$gene_count}{$total_count}=[@{$allele_groups{$key1}{$key2}}];
                        }
                        #not all stops go up in order, check for this
                        if ($end > $old_end || $old_end == 1000000){
                                $old_end = $end;
                        }
                }
                #print "\n";
        }else{
                #print "--- single $key1 ---\n";
                for my $key2 (sort {$a<=>$b} keys %{$allele_groups{$key1}}){
                        if ($allele_groups{$key1}{$key2}[1] eq 'manual'){
                                $early_man_count++;
                        }
                        $single_count++;
                        $final_single_gene{$key1}{$total_count}=[@{$allele_groups{$key1}{$key2}}];
                        $total_count++;
                        #print "$allele_groups{$key1}{$key2}[0]\t$allele_groups{$key1}{$key2}[1]\t$allele_groups{$key1
                              }{$key2}[2]\t$allele_groups{$key1}{$key2}[3]\t$allele_groups{$key1}{$key2}[4]\n";
                }
        }

}

open DUMP_A,">dump_allelic_final_hash.txt";
print DUMP_A Dumper( \%final_multi_gene );

print "picking␣the␣best␣prediction...\n";
open AM,">allelic_multi_annotation.fa";
open AMA,">allelic_multi_annotation.aa";
$count_man=0;
for my $key1 (sort keys %final_multi_gene){
                #print "$key1 -> ";
                for my $key2 (sort {$a<=>$b} keys %{$final_multi_gene{$key1}}){
                        labelbreak3: while($key2){
                        %picker=();
                        for my $key3 (sort {$a<=>$b} keys %{$final_multi_gene{$key1}{$key2}}){
                                $picker{$key3}=$final_multi_gene{$key1}{$key2}{$key3}[1];
                        }
                        #print the manual ones separately as all of them need to come out regardless (some of them are
                              nested within one larger prediction and may appear to be one gene)
                        for my $pick_key (keys %picker){
                                if ($picker{$pick_key} eq 'manual'){
                                        #my $pick_size = (keys %picker);
                                        #print "picker size1 $key2 = $pick_size\n";
                                        $count_man++;
                                        print "multi␣allelic␣$count_man␣manual␣$final_multi_gene{$key1}{$key2}{
                                              $pick_key}[0]\n";
                                        print AM ">$all_gene"."_a_multi_$final_multi_gene{$key1}{$key2}{$pick_key}[2]
                                              _$final_multi_gene{$key1}{$key2}{$pick_key}[8]_$final_multi_gene{$key1}{
                                              $key2}{$pick_key}[0]\n$final_multi_gene{$key1}{$key2}{$pick_key}[6]\n";
```

```
402                                     print AMA ">$all_gene"."_a_multi_$final_multi_gene{$key1}{$key2}{$pick_key}[2]
                                              _$final_multi_gene{$key1}{$key2}{$pick_key}[8]_$final_multi_gene{$key1}{
                                              $key2}{$pick_key}[0]\n$final_multi_gene{$key1}{$key2}{$pick_key}[7]\n";
403                                     print ALLF ">$final_multi_gene{$key1}{$key2}{$pick_key}[0]\n$final_multi_gene{
                                              $key1}{$key2}{$pick_key}[6]\n";
404                                     print ALLA ">$final_multi_gene{$key1}{$key2}{$pick_key}[0]\n$final_multi_gene{
                                              $key1}{$key2}{$pick_key}[7]\n";
405                                     $all_gene++;
406                                     last labelbreak3;
407                                     delete($picker{$pick_key});
408                             }
409                     }for my $pick_key (keys %picker){
410                             if ($picker{$pick_key} eq 'maker'){
411                                     print AM ">$all_gene"."_a_multi_$final_multi_gene{$key1}{$key2}{$pick_key}[2]
                                              _$final_multi_gene{$key1}{$key2}{$pick_key}[8]_$final_multi_gene{$key1}{
                                              $key2}{$pick_key}[0]\n$final_multi_gene{$key1}{$key2}{$pick_key}[6]\n";
412                                     print AMA ">$all_gene"."_a_multi_$final_multi_gene{$key1}{$key2}{$pick_key}[2]
                                              _$final_multi_gene{$key1}{$key2}{$pick_key}[8]_$final_multi_gene{$key1}{
                                              $key2}{$pick_key}[0]\n$final_multi_gene{$key1}{$key2}{$pick_key}[7]\n";
413                                     print ALLF ">$final_multi_gene{$key1}{$key2}{$pick_key}[0]\n$final_multi_gene{
                                              $key1}{$key2}{$pick_key}[6]\n";
414                                     print ALLA ">$final_multi_gene{$key1}{$key2}{$pick_key}[0]\n$final_multi_gene{
                                              $key1}{$key2}{$pick_key}[7]\n";
415                                     $all_gene++;
416                                     last labelbreak3;
417                             }
418                     }for my $pick_key (keys %picker){
419                             if ($picker{$pick_key} eq 'augustus'){
420                                     print AM ">$all_gene"."_a_multi_$final_multi_gene{$key1}{$key2}{$pick_key}[2]
                                              _$final_multi_gene{$key1}{$key2}{$pick_key}[8]_$final_multi_gene{$key1}{
                                              $key2}{$pick_key}[0]\n$final_multi_gene{$key1}{$key2}{$pick_key}[6]\n";
421                                     print AMA ">$all_gene"."_a_multi_$final_multi_gene{$key1}{$key2}{$pick_key}[2]
                                              _$final_multi_gene{$key1}{$key2}{$pick_key}[8]_$final_multi_gene{$key1}{
                                              $key2}{$pick_key}[0]\n$final_multi_gene{$key1}{$key2}{$pick_key}[7]\n";
422                                     print ALLF ">$final_multi_gene{$key1}{$key2}{$pick_key}[0]\n$final_multi_gene{
                                              $key1}{$key2}{$pick_key}[6]\n";
423                                     print ALLA ">$final_multi_gene{$key1}{$key2}{$pick_key}[0]\n$final_multi_gene{
                                              $key1}{$key2}{$pick_key}[7]\n";
424                                     $all_gene++;
425                                     last labelbreak3;
426                             }
427                     }for my $pick_key (keys %picker){
428                             if ($picker{$pick_key} eq 'p2g'){
429                                     print AM ">$all_gene"."_a_multi_$final_multi_gene{$key1}{$key2}{$pick_key}[2]
                                              _$final_multi_gene{$key1}{$key2}{$pick_key}[8]_$final_multi_gene{$key1}{
                                              $key2}{$pick_key}[0]\n$final_multi_gene{$key1}{$key2}{$pick_key}[6]\n";
430                                     print AMA ">$all_gene"."_a_multi_$final_multi_gene{$key1}{$key2}{$pick_key}[2]
                                              _$final_multi_gene{$key1}{$key2}{$pick_key}[8]_$final_multi_gene{$key1}{
                                              $key2}{$pick_key}[0]\n$final_multi_gene{$key1}{$key2}{$pick_key}[7]\n";
431                                     print ALLF ">$final_multi_gene{$key1}{$key2}{$pick_key}[0]\n$final_multi_gene{
                                              $key1}{$key2}{$pick_key}[6]\n";
432                                     print ALLA ">$final_multi_gene{$key1}{$key2}{$pick_key}[0]\n$final_multi_gene{
                                              $key1}{$key2}{$pick_key}[7]\n";
433                                     $all_gene++;
434                                     last labelbreak3;
435                             }
436                     }
437             }
438     }
```

```
}
close AM;close AMA;
open AS,">allelic_single_annotation.fa";
open ASA,">allelic_single_annotation.aa";
my $count_allelic_single=0;
$count_man=0;
for my $key1 (sort keys %final_single_gene){
        #print $key1."\n";
        for my $key2 (sort {$a<=>$b} keys %{$final_single_gene{$key1}}){
                if ($final_single_gene{$key1}{$key2}[1] eq 'manual'){
                        $count_man++;
                        print "single␣allelic␣manual␣$count_man␣$final_single_gene{$key1}{$key2}[0]\n";
                }
                print "single␣-␣$final_single_gene{$key1}{$key2}[0]\t$final_single_gene{$key1}{$key2}[1]\
                      t$final_single_gene{$key1}{$key2}[2]\t$final_single_gene{$key1}{$key2}[3]\t$final_single_gene{
                      $key1}{$key2}[4]\t\n";
                print "1=>$all_gene"."2=na_single_$final_single_gene{$key1}{$key2}[2]␣3=$final_single_gene{$key1}{$key2
                      }[8]␣4=$final_single_gene{$key1}{$key2}[0]␣5=$final_single_gene{$key1}{$key2}[6]\n";
                print AS ">$all_gene"."_na_single_$final_single_gene{$key1}{$key2}[2]_$final_single_gene{$key1}{$key2
                      }[8]_$final_single_gene{$key1}{$key2}[0]\n$final_single_gene{$key1}{$key2}[6]\n";
                print ASA ">$all_gene"."_na_single_$final_single_gene{$key1}{$key2}[2]_$final_single_gene{$key1}{$key2
                      }[8]_$final_single_gene{$key1}{$key2}[0]\n$final_single_gene{$key1}{$key2}[7]\n";
                print ALLF ">$final_single_gene{$key1}{$key2}[0]\n$final_single_gene{$key1}{$key2}[6]\n";
                print ALLA ">$final_single_gene{$key1}{$key2}[0]\n$final_single_gene{$key1}{$key2}[7]\n";
                $count_allelic_single++;
                $all_gene++;
        }
}
close AS; close ASA; close ALLF; close ALLA;
print "$count_groups␣multiple␣annotation␣groups\n";
print "$multiple_count␣multiple␣annotations␣across␣all␣genes\n";
print "$multiple_gene_count␣multiple␣gene␣annotation␣genes\n";
print "$count_allelic_single␣single␣annotation␣genes\n";
print "total␣multi␣genes␣=␣".(keys %final_multi_gene)."\n";
print "total␣single␣genes␣=␣".(keys %final_single_gene)."\n";


print "\n--------------␣COMBINED␣---------------\n";
print "$first_man_count␣allelic␣manual␣genes␣were␣read␣in\n\n";
print "$early_man_count␣allelic␣manual␣genes␣were␣processed\n";
print "Total␣gene␣objects␣=␣$all_gene\n";
```

Appendix C

CLUSTAL alignment of mannan endo-1,4-β-mannosidase

```
CLUSTAL 2.0.12 multiple sequence alignment

Flammeovirga_yaeyamensis_26186      MKYESKRMSNILKVLTLTLFLYGFNLSLIQAQNTNSEISCSSNDHCPQGY
Branchiostoma_floridae_2608077      --------------------------------------------------
Branchiostoma_floridae_2608148      -------------MSSESGSVNGLLLALAVVAALLSGGEAYGSGAPLSAC
k_10972_ext                         --------------------------------------------------
Mytilus_edulis_90108933             --------------------------------------------------
Mytilus_edulis_47606432             --------------------------------------------------
Haliotis_discus_211908630           --------------------------------------------------
Biomphalaria_glabrata_56462580      --------------------------------------------------
Biomphalaria_glabrata_56462582      --------------------------------------------------
Aplysia_kurodai_317414223           --------------------------------------------------
Haliotis_discus_85658727            --------------------------------------------------
Daphnia_pulex_321460555             --------------------------------------------------
Daphnia_pulex_321460556             --------------------------------------------------
Daphnia_pulex_321460557             --------------------------------------------------
Daphnia_pulex_321450949             --------------------------------------------------
Daphnia_pulex_321465382             --------------------------------------------------
Daphnia_pulex_321450057             --------------------------------------------------
Daphnia_pulex_321465383             --------------------------------------------------
Cryptopygus_antarcticus_157703      --------------------------------------------------
Limnoria_quadripunctata_293629      --------------------------------------------------
Limnoria_quadripunctata_293629      --------------------------------------------------
Limnoria_quadripunctata_293629      --------------------------------------------------
```

```
Callosobruchus_maculatus_31557      --------------------------------------------------
Callosobruchus_maculatus_31557      --------------------------------------------------
Callosobruchus_maculatus_31557      --------------------------------------------------
Callosobruchus_maculatus_31557      --------------------------------------------------
Gastrophysa_viridula_315570658      --------------------------------------------------
uncultured_bacterium_359755046      --------------------------------------------------
Spirochaeta_thermophila_307717      --------------------------------------------------
Saccharophagus_degradans_90021      --------------------------------------------------
Chloroflexus_aurantiacus_16384      --------------------------------------------------
Chloroflexus_aggregans_2198492      --------------------------------------------------
Oscillochloris_trichoides_3097      --------------------------------------------------
Clostridium_papyrosolvens_3262      --------------------------------------------------
Clostridium_cellulolyticum_220      --------------------------------------------------
Clostridium_sp_373945115            --------------------------------------------------
Clostridium_josui_270288704         --------------------------------------------------
Clostridium_acetobutylicum_158      --------------------------------------------------
Fibrobacter_succinogenes_26141      --------------------------------------------------
Chlamydomonas_reinhardtii_1594      --------------------------------------------------


Flammeovirga_yaeyamensis_26186      VCEGWPTYKCVPDGSGGGENPTPGNPIANAGANQTVIDTDGDGVETITLD
Branchiostoma_floridae_2608077      --------------------------------------------------
Branchiostoma_floridae_2608148      TSQRPGHTGTTAQTSTSPYSLTVSSSEYTPGQTLTVQITGADFQGFLIQA
k_10972_ext                         --------------------------------------------------
Mytilus_edulis_90108933             --------------------------------------------------
Mytilus_edulis_47606432             --------------------------------------------------
Haliotis_discus_211908630           --------------------------------------------------
Biomphalaria_glabrata_56462580      --------------------------------------------------
Biomphalaria_glabrata_56462582      --------------------------------------------------
Aplysia_kurodai_317414223           --------------------------------------------------
Haliotis_discus_85658727            --------------------------------------------------
Daphnia_pulex_321460555             --------------------------------------------------
Daphnia_pulex_321460556             --------------------------------------------------
Daphnia_pulex_321460557             --------------------------------------------------
Daphnia_pulex_321450949             --------------------------------------------------
Daphnia_pulex_321465382             --------------------------------------------------
Daphnia_pulex_321450057             --------------------------------------------------
Daphnia_pulex_321465383             --------------------------------------------------
Cryptopygus_antarcticus_157703      --------------------------------------------------
Limnoria_quadripunctata_293629      --------------------------------------------------
Limnoria_quadripunctata_293629      --------------------------------------------------
```

```
Limnoria_quadripunctata_293629      --------------------------------------------------
Callosobruchus_maculatus_31557      --------------------------------------------------
Callosobruchus_maculatus_31557      --------------------------------------------------
Callosobruchus_maculatus_31557      --------------------------------------------------
Callosobruchus_maculatus_31557      --------------------------------------------------
Gastrophysa_viridula_315570658      --------------------------------------------------
uncultured_bacterium_359755046      --------------------------------------------------
Spirochaeta_thermophila_307717      --------------------------------------------------
Saccharophagus_degradans_90021      --------------------------------------------------
Chloroflexus_aurantiacus_16384      ----------------------------------------MARTLH----
Chloroflexus_aggregans_2198492      ----------------------------------------MHPTLR----
Oscillochloris_trichoides_3097      ----------------------MFKYTHYDNPAPTTKELLMYASVRPQSV
Clostridium_papyrosolvens_3262      -------------------------MKKIVTLALTTAMALLAVLPLPAS
Clostridium_cellulolyticum_220      -------------------------MKKIASLVLTTAMVFLAALPLPAS
Clostridium_sp_373945115            -------------------------MKKIVSLTISAAIVFLTALPLPAS
Clostridium_josui_270288704         -------------------------MKKVISLLLTTAMVVLAALPLPVS
Clostridium_acetobutylicum_158      -------------------------MKKLKVISTVVAGVFLS-------
Fibrobacter_succinogenes_26141      --------------------------------------------------
Chlamydomonas_reinhardtii_1594      --------------------------------------------------

Flammeovirga_yaeyamensis_26186      GSGSTAEGAIVSFVWSEDGIELGTGETITENFAVGTHSIELTVTDDQGGT
Branchiostoma_floridae_2608077      --------------MPDAGKTAAIVALAQHAVNIKEYADTLVDR------
Branchiostoma_floridae_2608148      RKVGTTTAVGFFTSLPSGTKSNNCDNAVASGDNTATHSSTAAKR------
k_10972_ext                         --------------------MFRIVVLLLVALGVATGQRFLSVQ------
Mytilus_edulis_90108933             ------------------------------------AAVRLSVS------
Mytilus_edulis_47606432             ----------------------MLLTALAVLFASTGCQARLSVS------
Haliotis_discus_211908630           --------------------MIPCAPVLLLLVLPAVECDRLQIS------
Biomphalaria_glabrata_56462580      -------------------MKTLITGFLVVLCTLKLVCARLAVS------
Biomphalaria_glabrata_56462582      --------------------------------------------------
Aplysia_kurodai_317414223           -------------------MKIACVFLLVALVP--LAHSRLHIQ------
Haliotis_discus_85658727            --------------------MAVSLLVLLACGIAAVQCDRLSVQ------
Daphnia_pulex_321460555             ------------------MKMFKFVGLLLCFWASLS-AGRLTTS------
Daphnia_pulex_321460556             --------------------MFKFVALLLCCWASLS-AGRLTTS------
Daphnia_pulex_321460557             -------------------MLKLSFVLLLGFWASLS-VGRLTTS------
Daphnia_pulex_321450949             --------------------------------------------------
Daphnia_pulex_321465382             ----------------MRLHEFSLIGILFVSALELTRASRLSVS------
Daphnia_pulex_321450057             ----------------MRLHEFSLIGILFVSALELTRASRLSVS------
Daphnia_pulex_321465383             ----------------MRLYEFSLVVIFFVSGLELTRASRLSVS------
Cryptopygus_antarcticus_157703      -----------------MVKLFSFLLLVWVASPAFS-SEFLKAS------
Limnoria_quadripunctata_293629      ------------------MNHHLLEAVFFFGLFTSSFAARLSVS------
```

```
Limnoria_quadripunctata_293629      ------------------MNQMLLQAIFFLGLLSTSFAARLAVS------
Limnoria_quadripunctata_293629      ------------------MNRILVQTVLLLGVLSTSFAARLAVS------
Callosobruchus_maculatus_31557      ----------------MSTIKMKVVLAFLVIFGVHSIDAFLSVR------
Callosobruchus_maculatus_31557      ------------------MIKMKVILAFLVIFGVHSIDAFLTIR------
Callosobruchus_maculatus_31557      ------------------MVKMKAVLAFLVIFGVHSIDAFLSVR------
Callosobruchus_maculatus_31557      --------------------MKIGSAL-LLVVLCLHSIDAFLRVQ------
Gastrophysa_viridula_315570658      ----------------------MKVAVVFVLALGLHSIDAFLKVQ------
uncultured_bacterium_359755046      --------------MKGCRLALFLGLILL--ISLFVVVTDQVEA------
Spirochaeta_thermophila_307717      --------------MRRIYTVLLGALLLAGCVTGSGVPEGGVDP------
Saccharophagus_degradans_90021      -----------MPYFQVPKPSLPLACVLLVMLVSLGACSGAIQSNHGAEA
Chloroflexus_aurantiacus_16384      ---------LALCILLVLITLPHIPTAIAAPANCPVDRSHLIPRFNG---
Chloroflexus_aggregans_2198492      ---------LALIGLLIINSPLLIPTQASTPLSCPVDRSHLIDRLGG---
Oscillochloris_trichoides_3097      LRITTLSSLIFGLILALLPSLTTSPTVVSAAPACPTNQAHLVPFQGG---
Clostridium_papyrosolvens_3262      AEKTYKLGDVDNDTFVSALDLAAVRQHILGLKTLTGEAFKAADVNANGEI
Clostridium_cellulolyticum_220      AATTYKLGDVDNDTLISAIDLAAVQQHILGKKTLTGEAFKAADVNANGEI
Clostridium_sp_373945115            AATTYKLGDVDNDTQISALDLATVKQHLLGIKTLTGDALKAADVNANGEI
Clostridium_josui_270288704         AETTYKLGDVDNDTLVSAVDLAAVQSHILGKKTLTGEAFKAADVNANGEI
Clostridium_acetobutylicum_158      -------------TLVSFSSITKV---------------KAADTTDN---
Fibrobacter_succinogenes_26141      ------------------MNSFKTLIAASLLGTAAFAAPGLKVS------
Chlamydomonas_reinhardtii_1594      --------------------------------------------------


Flammeovirga_yaeyamensis_26186      GSATILVIVQ-------------PFNTTQMNRIAVGDQLVFLSGGNLAWY
Branchiostoma_floridae_2608077      ------------------------------QSHQRINAVLPHQCLHKSKWA
Branchiostoma_floridae_2608148      ------------------------------DLTLTWSAPENQAGQGTIEFV
k_10972_ext                         ------------------------------NGQLTLNGEKVFLSGMNIAWQ
Mytilus_edulis_90108933             ------------------------------GTNLNYNGHHIFLSGANQAWV
Mytilus_edulis_47606432             ------------------------------GTNLNYNGHHIFLSGANQAWV
Haliotis_discus_211908630           ------------------------------GDYFTKDGSRVFLSGVNLAWV
Biomphalaria_glabrata_56462580      ------------------------------GNQFTYNGQRIFLSGGNLPWI
Biomphalaria_glabrata_56462582      --------------------------------------------------
Aplysia_kurodai_317414223           ------------------------------NGHFVLNGQRVFLSGGNLPWM
Haliotis_discus_85658727            ------------------------------GNHFVKGGQKVFLSGANLAWV
Daphnia_pulex_321460555             ------------------------------GTNLYYNGQKVFLSGANIAWN
Daphnia_pulex_321460556             ------------------------------GTNFYYNGQKVFLSGVNIAWN
Daphnia_pulex_321460557             ------------------------------GRDFLYNGQRVFLSGANIAWY
Daphnia_pulex_321450949             --------------------------------------------------
Daphnia_pulex_321465382             ------------------------------GNKLMFNGKSVFLSGVNFAWN
Daphnia_pulex_321450057             ------------------------------GNKLMFNGKSVFLSGVNFAWN
Daphnia_pulex_321465383             ------------------------------GNQLMFNGKSVFLSGVNFAWN
Cryptopygus_antarcticus_157703      ------------------------------GSNFYYGGQKVFLSGVNFAWR
```

```
Limnoria_quadripunctata_293629      ----------------------------GGGLSYGGQKAYLNGANIAWN
Limnoria_quadripunctata_293629      ----------------------------GTSLTYGNDQVYLNGVNIAWN
Limnoria_quadripunctata_293629      ----------------------------GMTLTYGGKQVYLNGENIPWN
Callosobruchus_maculatus_31557      ----------------------------NTSFYYGNDKVFLSGANLAWI
Callosobruchus_maculatus_31557      ----------------------------NNSFYYGEDRVFLSGANIAWI
Callosobruchus_maculatus_31557      ----------------------------NTSFYYGKDKVFLSGANIAWF
Callosobruchus_maculatus_31557      ----------------------------DKKLFYNNDQVFLSGANIAWF
Gastrophysa_viridula_315570658      ----------------------------NNALYYNNDKVFLSGANIAWY
uncultured_bacterium_359755046      ----------------------------EGTLKYMNKDFFASGMNLAWL
Spirochaeta_thermophila_307717      ----------------------------DALLPHNGKRVFLNGMNLAWV
Saccharophagus_degradans_90021      ------------------------ALKTPQGHVVIKGKPVYLSGFNVAWF
Chloroflexus_aurantiacus_16384      ------------------------RWFLLGANVPWLNGGYSADFGTVEEWG
Chloroflexus_aggregans_2198492      ------------------------HWFLVGANVPWLNGGYGADFATVEEWN
Oscillochloris_trichoides_3097      ------------------------RWFLSGVNVPWQSGGYGADFGTVEEWG
Clostridium_papyrosolvens_3262      EALDLSELKQFLLGKITKFSGEGQQQPSGVGITWMDGKTLYPVGVNYAWY
Clostridium_cellulolyticum_220      EALDLAELKQFLLGRITKFSGEGQQQPSGVGITWMDGNTLYPVGVNYAWY
Clostridium_sp_373945115            EALDLSEIKQYLLGKITKFSGEGQQQPSGVGITWMDGNTLYPVGVNYAWY
Clostridium_josui_270288704         EALDLAEIKQFILGRIIKFSGEGQQQPPGLGIIWMDGSTIYPVGVNYAWY
Clostridium_acetobutylicum_158      -------------------------KP---GINWMNGSKYFPLGANYAWD
Fibrobacter_succinogenes_26141      ----------------------------GTDLQYNGKKIFFSGTNLAWS
Chlamydomonas_reinhardtii_1594      --------------------------------------------------

Flammeovirga_yaeyamensis_26186      NFANDFGEKTTN---LDYFEQAIS--------------------------
Branchiostoma_floridae_2608077      LTLEERVCLQLVGDFRLALYRAQATNGR----------------------K
Branchiostoma_floridae_2608148      ATVAQQKATYWMGITSAQLSEAASGGATGATPTGSSQTVTASILARCWFL
k_10972_ext                         NYGADFGNGQYSCCTSSALDDYVR--------------------------
Mytilus_edulis_90108933             NYARDFGHNQYS-KGKSTFESTLS--------------------------
Mytilus_edulis_47606432             NYARDFGHNQYS-KGKSTFESTLS--------------------------
Haliotis_discus_211908630           GYATDFGNNQFA-ARKSSYERFFK--------------------------
Biomphalaria_glabrata_56462580      QYAYDFGDHQWD-SRKGTFENQLT--------------------------
Biomphalaria_glabrata_56462582      --------------------------------------------------
Aplysia_kurodai_317414223           SYAYDFGDGQWQ-RNKNRIEPEFK--------------------------
Haliotis_discus_85658727            QYAYDFGNNHYKGRVQGILEGYIR--------------------------
Daphnia_pulex_321460555             SYGYDFGNGQYAANSKSTLESWLT--------------------------
Daphnia_pulex_321460556             SYGYDFGNGQYAANSKATLESWLT--------------------------
Daphnia_pulex_321460557             SYGYDFGNGVYQSDVKETLETWLT--------------------------
Daphnia_pulex_321450949             --------------------------------------------------
Daphnia_pulex_321465382             SYGNDFGNGKYTANSKTTFEQWLA--------------------------
Daphnia_pulex_321450057             SYGNDFGNGKYTANSKTTFEQWLA--------------------------
Daphnia_pulex_321465383             SYGYDFGNGQYTANSKTTFEQWLA--------------------------
```

```
Cryptopygus_antarcticus_157703      SYGSDFGNGQYASNG-PALKDWIN--------------------------
Limnoria_quadripunctata_293629      SYGYDFGNGNY----DGSIESWMS--------------------------
Limnoria_quadripunctata_293629      SYGYDFGNGNY----DGSIENWMS--------------------------
Limnoria_quadripunctata_293629      NYGYDFGNGVY----DNTIEQWMQ--------------------------
Callosobruchus_maculatus_31557      YFGSDFGSGGY-AKVRSAYESAID--------------------------
Callosobruchus_maculatus_31557      NFAEDFGSGGY-AKVRSSYESAID--------------------------
Callosobruchus_maculatus_31557      NFARDFGSGGY-YQVRSRFETAIN--------------------------
Callosobruchus_maculatus_31557      NFARDFGSGAY-DYVKPRFEQAID--------------------------
Gastrophysa_viridula_315570658      NYGWDFGSGAY-SNVKTNYQQALD--------------------------
uncultured_bacterium_359755046      SFAQDLDR-FYEPRFIRALDEVAA--------------------------
Spirochaeta_thermophila_307717      NFANDLTQ-FDEARFTRAVDDVAS--------------------------
Saccharophagus_degradans_90021      DFARDFGKGVDEKALRKALQQVKD--------------------------
Chloroflexus_aurantiacus_16384      QHTYDAN---------ATRTMFRA--------------------------
Chloroflexus_aggregans_2198492      QHTYDPD---------TTRAMFRA--------------------------
Oscillochloris_trichoides_3097      QHTYSTD---------KTRQMFAA--------------------------
Clostridium_papyrosolvens_3262      NWSYEFSDNNWNYNFSRIKSDLDT--------------------------
Clostridium_cellulolyticum_220      NWSYEFSDNNWNSNFTRIKSDLDT--------------------------
Clostridium_sp_373945115            NWSYEFSDNNWTSNFTRIKSDLDT--------------------------
Clostridium_josui_270288704         NWSYEFLDNNWTYNFTRIKSDLDA--------------------------
Clostridium_acetobutylicum_158      EWDNDFNDNGWTTRFAKIKADFDN--------------------------
Fibrobacter_succinogenes_26141      DYNSDVGASPLDENAWRKAVEGTR--------------------------
Chlamydomonas_reinhardtii_1594      ---------------------------------------------------


Flammeovirga_yaeyamensis_26186      ---------------------------------------------------
Branchiostoma_floridae_2608077      ASTAVLIEKLEEKKK-----------------------------------
Branchiostoma_floridae_2608148      SSSCGYLRRLRTETLGTVYLEYDLLFRCTGDIQQGTPTVQAWCRYGSDFG
k_10972_ext                         -----RIKAEGETHS-----------------------------------
Mytilus_edulis_90108933             ---------------------------------------------------
Mytilus_edulis_47606432             ---------------------------------------------------
Haliotis_discus_211908630           ---------------------------------------------------
Biomphalaria_glabrata_56462580      ---------------------------------------------------
Biomphalaria_glabrata_56462582      ---------------------------------------------------
Aplysia_kurodai_317414223           ---------------------------------------------------
Haliotis_discus_85658727            ---------------------------------------------------
Daphnia_pulex_321460555             ---------------------------------------------------
Daphnia_pulex_321460556             ---------------------------------------------------
Daphnia_pulex_321460557             ---------------------------------------------------
Daphnia_pulex_321450949             ---------------------------------------------------
Daphnia_pulex_321465382             ---------------------------------------------------
Daphnia_pulex_321450057             ---------------------------------------------------
```

```
Daphnia_pulex_321465383          ------------------------------------------------
Cryptopygus_antarcticus_157703   ------------------------------------------------
Limnoria_quadripunctata_293629   ------------------------------------------------
Limnoria_quadripunctata_293629   ------------------------------------------------
Limnoria_quadripunctata_293629   ------------------------------------------------
Callosobruchus_maculatus_31557   ------------------------------------------------
Callosobruchus_maculatus_31557   ------------------------------------------------
Callosobruchus_maculatus_31557   ------------------------------------------------
Callosobruchus_maculatus_31557   ------------------------------------------------
Gastrophysa_viridula_315570658   ------------------------------------------------
uncultured_bacterium_359755046   ------------------------------------------------
Spirochaeta_thermophila_307717   ------------------------------------------------
Saccharophagus_degradans_90021   ------------------------------------------------
Chloroflexus_aurantiacus_16384   ------------------------------------------------
Chloroflexus_aggregans_2198492   ------------------------------------------------
Oscillochloris_trichoides_3097   ------------------------------------------------
Clostridium_papyrosolvens_3262   ------------------------------------------------
Clostridium_cellulolyticum_220   ------------------------------------------------
Clostridium_sp_373945115         ------------------------------------------------
Clostridium_josui_270288704      ------------------------------------------------
Clostridium_acetobutylicum_158   ------------------------------------------------
Fibrobacter_succinogenes_26141   ------------------------------------------------
Chlamydomonas_reinhardtii_1594   ------------------------------------------------


Flammeovirga_yaeyamensis_26186   -------------------------DFASRGGNSMRLWVHINGANNPELD
Branchiostoma_floridae_2608077   --------------SRVARFFTAVMALCICG---LLGGKKDVHEEEHEYI
Branchiostoma_floridae_2608148   GGLYYSYKSSSPCPSSKSKYEQAIMDIANNGGNSLRVWLHVEGQETPVFS
k_10972_ext                      -----------------------------------VRIWLHCDGWYTPSYD
Mytilus_edulis_90108933          -------------------------DMQSHGGNSVRVWLHIEGESTPEFD
Mytilus_edulis_47606432          -------------------------DIQSHGGNSVRVWLHIEGESTPEFD
Haliotis_discus_211908630        -------------------------ELHESGGSSIRIWIHVQGETSPLFD
Biomphalaria_glabrata_56462580   -------------------------QLKNAGGNSIRLWVHIQGESTPAFD
Biomphalaria_glabrata_56462582   ------------------------------------------------
Aplysia_kurodai_317414223        -------------------------KLHDAGGNSMRLWIHIQGETTPAFN
Haliotis_discus_85658727         -------------------------DLSKAGGNSMRVWIHMEGANTPEFD
Daphnia_pulex_321460555          -------------------------QIANSGGNSVRIWLHVEGANTPAFD
Daphnia_pulex_321460556          -------------------------RIDANGGNSVRMWVHVDGKNTPAFD
Daphnia_pulex_321460557          -------------------------MIANSGGNSVRQWVHVEGQNTPAYD
Daphnia_pulex_321450949          ------------------------------------------------
Daphnia_pulex_321465382          -------------------------EVATNGGNSVRVWLHVEGDNTPNYD
```

```
Daphnia_pulex_321450057          -------------------------EVATNGGNSVRVWLHVEGDNTPNYD
Daphnia_pulex_321465383          -------------------------EVATNGGNSVRVWLHVEGDNTPNYD
Cryptopygus_antarcticus_157703   -------------------------KVKASGGNTARVWVHVEGQVSPAFD
Limnoria_quadripunctata_293629   -------------------------DIGSAGGNAARQWVHVEGKSTPQYD
Limnoria_quadripunctata_293629   -------------------------DIGSAGGNTVRMWVQVEGESTPSFD
Limnoria_quadripunctata_293629   -------------------------DIGSAGGNSVRMWVHVEGQNTPSFD
Callosobruchus_maculatus_31557   -------------------------DISSHGGNAMRVWLHADGRYSPKWD
Callosobruchus_maculatus_31557   -------------------------DISSHGGNVIRVWLHADGRWSPKWD
Callosobruchus_maculatus_31557   -------------------------EISSNGGNVIRVWVHTDGQWSPKWD
Callosobruchus_maculatus_31557   -------------------------EISNAGGNVIRVWVHIDGQWSPKWD
Gastrophysa_viridula_315570658   -------------------------EISQAGGNSIRVWVHIDGQWSPKFD
uncultured_bacterium_359755046   ----------------------------AGGNTVRWWLHTNCKMSPMFK
Spirochaeta_thermophila_307717   ----------------------------AGGNVLRWWLHVNGSKTPLFD
Saccharophagus_degradans_90021   ----------------------------SGGNSLRWWMHTDGSQTPEWR
Chloroflexus_aurantiacus_16384   -------------------------LRQQGANTVRWWLFADGRGAPEFN
Chloroflexus_aggregans_2198492   -------------------------LRQKGANTVRWWLFADGRGTPEFD
Oscillochloris_trichoides_3097   -------------------------LKANGINTVRWWVFADGRGAPEFA
Clostridium_papyrosolvens_3262   -------------------------MSTKGIHALRWWVFPDLAYGPLWS
Clostridium_cellulolyticum_220   -------------------------MSSKGINSLRWWVFPDLAYGPLWS
Clostridium_sp_373945115         -------------------------MSTKGIHSLRWWVFPDLAYGPLWS
Clostridium_josui_270288704      -------------------------MSTKGIRSLRWWIFPDLAYGPLWS
Clostridium_acetobutylicum_158   -------------------------MSAQGIHTVRWWVFCNMYASPLFS
Fibrobacter_succinogenes_26141   ----------------------------AAGGNAIRWWLFNNMSQSPTID
Chlamydomonas_reinhardtii_1594   ---------------------------------MYLDVIGHHN----

Flammeovirga_yaeyamensis_26186   AN---GYTSGLEPSMIQDLRDVLDIAYDHNVVLNLCLWSFDMLNTRDYPI
Branchiostoma_floridae_2608077   SW-GSVTQTDSTNQLINDLKSLLSYAKARNVLVFLVLWNGAHHHDM----
Branchiostoma_floridae_2608148   YW-GSVTQTDATDELVNELKDLLVFAKRHNVLVFLVLWNGAHHGNH----
k_10972_ext                      SG-GYVTGTDDQNTMTSELAQFLDVAYDNNLLVFIVLWNGATTPDP----
Mytilus_edulis_90108933          NN-GYVTGIDN--TLISDMRAYLHAAQRHNILIFFTLWNGAVKQST----
Mytilus_edulis_47606432          NN-GYVTGIDN--TLISDMRAYLHAAQRHNILIFFTLWNGAVKQST----
Haliotis_discus_211908630        GN-GYVTGLDSSGTFLSDMNELLGLGQKYNILVFFCLWNGAVKFDK----
Biomphalaria_glabrata_56462580   GN-GYVTAPDHQGTLINDFKDMLDIAQRHNILVFPTLWNAAVDQDN----
Biomphalaria_glabrata_56462582   -------------------------------------------------
Aplysia_kurodai_317414223        DQ-GFVTGPDKQGTMLDDMKDLLDTAKKYNILVFPCLWNAAVNQDS----
Haliotis_discus_85658727         SS-GHVIGMDKGGTMLADLKSMLNYAASHNVLIFLCLWNGAVNQGS----
Daphnia_pulex_321460555          GN-GYVTGPDSTGTMISDMRSFLDFAQSKNILVIFVLWNGAYLTV-----
Daphnia_pulex_321460556          GN-GYVTGLDNTGTMISDLKSFLDFAQSKQLLVVLVLWNGAERPT-----
Daphnia_pulex_321460557          SN-GYVTGPDRTGTIIDDMRSFLDFAQSQNILVIFVLWNGAVLEN-----
Daphnia_pulex_321450949          -------------------------------------------------
```

```
Daphnia_pulex_321465382          AN-GYVLGPDKTGTLISDMKSFLDSAKAKNILVIFVLWNGATLRN-----
Daphnia_pulex_321450057          AN-GYVLGPDKTGTLISDMKSFLDSAKAKNILVIFVLWNGATLRN-----
Daphnia_pulex_321465383          AN-GYVLGPDRTGTLIPDMKLFLDSAKAKNILVIFVLWNGATLRN-----
Cryptopygus_antarcticus_157703   SH-GFVTSTDSKKTLINDLSDLLDYANGQNVFLILVLFNGALQNN-----
Limnoria_quadripunctata_293629   GS-GMVTSCDSSGAFLRDVVSFLDSAQQSDVLVIFTIWNGAVMSN-----
Limnoria_quadripunctata_293629   NR-GMVTACDNTGDFLTDVVTFLDAAQESGVLVIFTVWNGAVMSN-----
Limnoria_quadripunctata_293629   GR-GMVTACDNTGDFLNDVVQFLDSAQQSNVLVMFTVWNGAVMEN-----
Callosobruchus_maculatus_31557   QD-GFATG-EDTQSLIEDLGLMLDYAASKNVFIVLTLWT-LEGTP-----
Callosobruchus_maculatus_31557   KD-GFATG-EDTQSLIDDLGLMLDYAASKNVFVFITLWT-LEGTP-----
Callosobruchus_maculatus_31557   QN-GFATG-EDTQSLIQELGLMLDYAASKNVFVILVLWN-LDVTP-----
Callosobruchus_maculatus_31557   AN-GFATG-EDTPSLINELGQLLDHAAQRNVFVIFTLWD-LNVTP-----
Gastrophysa_viridula_315570658   SE-GYATG-SDTDSLISDLGELLDYAEQKNVFVILCLWN-LAVAP-----
uncultured_bacterium_359755046   D----GKVSGLHRSNIPNLVRALDLAEERGIVLLLSLFSFDMLQDQPGVN
Spirochaeta_thermophila_307717   EN---GMVVGMPEEALINLKRALDISFSRGVGLILCLWSFDMLQPQSGVN
Saccharophagus_degradans_90021   TVKGVRLVAGPGGSLIQDLKTALDIAAEYDVYIVPSIWSFDMLKDNDYRK
Chloroflexus_aurantiacus_16384   ASS-GGAVTGFDATFLPSLASAIQIAAEENIYLVFNLWSFDMLFADSTAT
Chloroflexus_aggregans_2198492   ANN-GGAVTGLDTNFLPGLASAIQIAAEEDIYLVFNLWSFDMLMADSTMY
Oscillochloris_trichoides_3097   ATS-GGAVTGLDANTLPSMADAIKLAQEYNIRIVFNLWSFDMLMPDSNGY
Clostridium_papyrosolvens_3262   GPNEGSLCTGLPEKWVDHMKETCDYAYSKGIKIYWTITSFDCARADDAYD
Clostridium_cellulolyticum_220   GPNEGSLCTGLPEKWVDHMKETCDYAYSKGIKIYWTITSFDCARADDSVD
Clostridium_sp_373945115         GANEGSLCTGLPAKWVDHMKETCDYAYSKGIKIYWTITSFDCARADDSVD
Clostridium_josui_270288704      GPNEGSLCTGLPDKWVDHMKETCDYAYSKGIKIYWTITSFDCAREDDSVD
Clostridium_acetobutylicum_158   SQDGKGVCTGLPDKWTDHMKEAADYAYSKNMKIYFTLTSFDVAKTNNSFY
Fibrobacter_succinogenes_26141   ET--THLVTGPKENTIANMKKALDIAEEYGVMVSMCLFSHNLMEPNQWGL
Chlamydomonas_reinhardtii_1594   ----------------SELPYLLPLVTS----------------------

Flammeovirga_yaeyamensis_26186   E---VSQRARKLLEEEENIDAYINNALIPMVN--------GLKDHEALLS
Branchiostoma_floridae_2608077   ----YWKRLQNLIWDDLKLDTYLEHALKPLAG--------ALKNQRALGG
Branchiostoma_floridae_2608148   ----FWK-VRDLVWDDLKLGTYVEQALRPLAL--------GLKDERALGG
k_10972_ext                      ----QYL---DLIWDESKLQSYIDRALAPMVS--------ALSGKVALGG
Mytilus_edulis_90108933          ----HYR-LNGLMVDTRKLQSYIDHALKPMAN--------ALKNEKALGG
Mytilus_edulis_47606432          ----HYR-LNGLMVDTRKLQSYIDHALKPMAN--------ALKNEKALGG
Haliotis_discus_211908630        ----EYR-MDGLIRDTGKLTSYLQHALIPWVK--------SVKDNPAVGG
Biomphalaria_glabrata_56462580   ----SHR-LDGFIVDWRKLQSYIDKALVPLAS--------AVRGHPALGA
Biomphalaria_glabrata_56462582   --------LDGLIVDERKLQSYIDKVLTPLAT--------AVKGHPALGA
Aplysia_kurodai_317414223        ----HNR-LDGLIKDQHKLQSYIDKALKPIVN--------HVKGHVALGG
Haliotis_discus_85658727         ----HAH-LDGLIRDTNKLQSYINKALIPMVK--------GLAGLPGLGG
Daphnia_pulex_321460555          ------QNTINLFWDDGKLQSYIDNALKPMVS--------ALGDHPALGA
Daphnia_pulex_321460556          ------DNTINLLYDESKLQTYIDNALKPMVD--------ALGNHPALAA
Daphnia_pulex_321460557          ------QNTINLFYDDAKLQSYIDNALKPMVA--------ALGDHPALAA
```

```
Daphnia_pulex_321450949          --------------------------------------------------
Daphnia_pulex_321465382          ------QNSINLYWDNSKLQTYLDKALTPMVK--------ALAAHPALGA
Daphnia_pulex_321450057          ------QNSINLYWDNSKLQTYLDKALTPMVK--------ALAAHPALGA
Daphnia_pulex_321465383          ------QNSINLYWDNSKLQTYLDKALTPMVK--------ALAAHPALGA
Cryptopygus_antarcticus_157703   ------SNVQNLFWDESKLNSYINNALTPMVN--------ALKSKPSLAA
Limnoria_quadripunctata_293629   ------QQYVDMIMDDNKLQSYLDNCLTDFAR--------AVSGHPALGA
Limnoria_quadripunctata_293629   ------QPYIDMILDDDKLQSYLDNCLTDWAK--------AVADHPALGA
Limnoria_quadripunctata_293629   ------QPYIDMVMDDNKIQSYLDNCLTDWVN--------AVKGHPALGS
Callosobruchus_maculatus_31557   ------KPMMHLYYQEDRLQAYLDRVLKPLVA--------GLKDKKALAA
Callosobruchus_maculatus_31557   ------KPMMHLYYQEDRLQSYLDRVLKPLVV--------ALRDKKALAG
Callosobruchus_maculatus_31557   ------QPMLHLYTEDDKLQAYLDRVLKPLVA--------GLKDKKALAA
Callosobruchus_maculatus_31557   ------RQMLHLYSQPDRLQSYLDKVLKPLVA--------ALKDKPALAA
Gastrophysa_viridula_315570658   ------TKMLPLYTDDAKLQSYLEKVLKPMAA--------GLKDKKALAA
uncultured_bacterium_359755046   -----LVNNKNLLEQIDHTQAYIDNALIPMVQ--------AVKDHPALFA
Spirochaeta_thermophila_307717   -----QARNLRLIEDEEVTRSYIENALVPMVR--------MLKRHPGVIA
Saccharophagus_degradans_90021   P---PTQDNYRLLTEDKVLNSYINNALVPMVQ--------ALNYHPQLAA
Chloroflexus_aurantiacus_16384   ARGDHAGGHRDLIVDSAKRASFINNALLPMLRYPVGNSGYTIGTHPNVLA
Chloroflexus_aggregans_2198492   ERGEHAGGHRDLIVDPVKRASFINNALLPMLRYPVGSSGYTIGTHPHVLA
Oscillochloris_trichoides_3097   TRGEHAGGHTDLITDATKRASFINKALLPMLAYPVPGTSYTIGNHPNVLG
Clostridium_papyrosolvens_3262   --------HDDIIDNSTVLQSFLDNAMKPILQ--------TLGTHPGVLG
Clostridium_cellulolyticum_220   --------HDDIIDNPIVLQSFLDNAMKPILQ--------TLGEHPGVLG
Clostridium_sp_373945115         --------HDDIIDNPTVLQSFLDNAMKPILQ--------TLGTHPGVLG
Clostridium_josui_270288704      --------HDDIIDNPTVLQSFLDNAMKPILQ--------ALGTHPGVLG
Clostridium_acetobutylicum_158   --------HGSIIDDPTIRKSYIDNAVTPVVK--------ALGDNPGVMG
Fibrobacter_succinogenes_26141   YNEKLDITANELLFEDAGTKAFIDNVLIPVVK--------AIGNHKALMT
Chlamydomonas_reinhardtii_1594   ------CRLRTVDVSVQRLQRFVSRLAAAVHA------------------

Flammeovirga_yaeyamensis_26186   WEVFNEPEGMSNEFGWDFT--------------------------DHVPM
Branchiostoma_floridae_2608077   WEIMNEPEGSLKIEHHS-DPCYDTIFLQGTGAGWAHLDTSGLPTWTDVPR
Branchiostoma_floridae_2608148   WEIINEPEGSLRVQHDS-DPCYNTDFLAGSGAGWAGNSDG-----NYLPM
k_10972_ext                      WEIMNEPEGIVGAGVSDPNPCFDTQVLGGSGAGWAG---------TFIPM
Mytilus_edulis_90108933          WDIMNEPEGEIKPGESSSEPCFDTRHLSGSGAGWAG---------HLYSA
Mytilus_edulis_47606432          WDIMNEPEGEIKPGESSSEPCFDTRHLSGSGAGWAG---------HLYSA
Haliotis_discus_211908630        WDIMNEPEGLINTQRSSNNPCLNATHLIPGGAGWAG---------RLYNY
Biomphalaria_glabrata_56462580   WDIMNEPEGMINTDISNWDRCYDSTALKNSGAGWAG---------KKYSY
Biomphalaria_glabrata_56462582   WDIMNEPEGMINPDIGNSDRCYDATALKNSGAGWAA---------KKYGY
Aplysia_kurodai_317414223        WDLMNEPEGMMIPDKHNAEKCYDTTALKNSGAGWAG---------NKYLY
Haliotis_discus_85658727         WEVINEPEGVLMPDVTNSDPCFDTTHLKNSGAGWAG---------KLYKY
Daphnia_pulex_321460555          WEIMNEPEGSLLNNQADANACFDTTPLKDTGAGWTN---------LYIPM
Daphnia_pulex_321460556          WEIMNEPEGLLQNNVYNGNPCYDTTPLKDTGAGFAF---------TNIPM
```

```
Daphnia_pulex_321460557          WEIMNEPEGAILLNQASDNPCFDTTPLANTDASWTG---------LTIPM
Daphnia_pulex_321450949          --------------------------------------------------
Daphnia_pulex_321465382          WEIVNEPEALVYNNKADANSCFNTVPMANSGAGWTG---------KWIPM
Daphnia_pulex_321450057          WEIVNEPEALVYNNKADANSCFNTVPMANSGAGWTG---------KWIPM
Daphnia_pulex_321465383          WEIVNEPEGLLYNNKPDSNSCFNTVPIANSGAGWTG---------KWIPM
Cryptopygus_antarcticus_157703   WEVLNEPEGTLQPG-SDQNSCYDTSTLAAQGAGWGG---------KKFPM
Limnoria_quadripunctata_293629   WEAINEPEGSVQVS-TAANACYDTTVIGQSGAGWTG---------TNIPM
Limnoria_quadripunctata_293629   WEAMNEPAGSVHVS-SDANHCYDTTNVGKQGGGWSG---------ANIPM
Limnoria_quadripunctata_293629   WEPINEPEGSVQIT-SDSNPCYDTTIIGQSGAGWTG---------ADIPI
Callosobruchus_maculatus_31557   WDLVNEPMGSLSQTHKDPNPCYDTTHLKDTGAGWAN---------ETIEY
Callosobruchus_maculatus_31557   WDLVNEPMGSISQTQVDPNPCYDTTHLKDSGAGWAG---------KTIDF
Callosobruchus_maculatus_31557   WDLVNEPMGSLSQWQQDPNPCYDTTHLKDTGAGWAG---------TTINY
Callosobruchus_maculatus_31557   WEVVNEPLASITETQRDINPCFDTTHLKYSGAGWSG---------AHLLL
Gastrophysa_viridula_315570658   WDIINEPIGSLTQGLTDSNPCYDTNNLINSGADWTN---------VHLKP
uncultured_bacterium_359755046   WEIFNEPEGMARPFG------WTPVKT----------------------EM
Spirochaeta_thermophila_307717   WEVFNEPEGMLPGGG------WTPRRT----------------------EM
Saccharophagus_degradans_90021   WELFNEPENMTESWFPQQQGFYGGKVP----------------------SL
Chloroflexus_aurantiacus_16384   WDIFNEPEFGISEPPHFTPSGEVAQP----------------------VTL
Chloroflexus_aggregans_2198492   WDIFNEPEFGIDEPPHFTPAHNIAQP----------------------VTL
Oscillochloris_trichoides_3097   WDIFNEPEFGVSDLGAVDP--QISAP----------------------VTL
Clostridium_papyrosolvens_3262   WDIINEPEWIIKKEDNGEP-NNKGEI----------------------FPL
Clostridium_cellulolyticum_220   WDIINEPEWIIKKEDNGEP-NNKGEI----------------------FPL
Clostridium_sp_373945115         WDIINEPEWIINKEDNGEP-NNKGES----------------------FPL
Clostridium_josui_270288704      WDIINEPEWIIKKEDNGEP-NNKGEI----------------------FPL
Clostridium_acetobutylicum_158   WDVINEPEWTISSADGGNP-GDSIKG----------------------WSL
Fibrobacter_succinogenes_26141   WEVFNEPEGMTSECSGWTT--------------------------KKMAL
Chlamydomonas_reinhardtii_1594   ----ADPRALVTVGSHSPPYCSDAQWLVG----------------------

Flammeovirga_yaeyamensis_26186   SVIQRFVNRLSGAIHRADPAA----LVTNGSWSFKANSDIFSGVEKNYYS
Branchiostoma_floridae_2608077   HRVLRFINRQAAAIKAKDPNH----LVTVGSWSEHGQ-----G-VRNLYS
Branchiostoma_floridae_2608148   YRMLRFINRQAAALKEADPNH----LVTVGSWSEKGQ-----G-IRNLYT
k_10972_ext                      ELLQRFINRQSAAIKRADSKA----IVTIGSWSERAQTDAL-G-WRNYYK
Mytilus_edulis_90108933          QEIGRFVNWQAAAIKEVDPGA----MVTVGSWNMKADTDAM-G-FHNLYS
Mytilus_edulis_47606432          QEIGRFVNWQAAAIKEVDPGA----MVTVGSWNMKADTDAM-G-FHNLYS
Haliotis_discus_211908630        EDVQRFINWQVDAIRQTDPGA----LVTLGSWKAQVNTDEY-G-SHNHYS
Biomphalaria_glabrata_56462580   YDTLRFINWQADAIKNVDSGF----LVTVGSWNPKSNTDQF-G-FVDHYS
Biomphalaria_glabrata_56462582   HDIIRFVNWQAAAIKHVDPGF----LVTVGAWNPKSNTDRF-G-FVDHYS
Aplysia_kurodai_317414223        QDILRFLNWQADAIKTTDPGA----LVTMGVWNPKSNTDHF-N-MNNHYS
Haliotis_discus_85658727         DDFLRFINWQAAAIKSADAHT----LVTMGSWNAKSNVNIK-G-YYNHYS
Daphnia_pulex_321460555          QNILKFVNWQADGVKGTNGAA----LVTLGSWSEHAQTDTK-AQSRNYYT
```

```
Daphnia_pulex_321460556            QNILKFVNWQADAVKQRNSAC----LVTIGSWSEHAQTDTK-AQSRNYYT
Daphnia_pulex_321460557            ENNLKFVNWQTHAIKETNSAS----LVTLGSWSEHAQSDAY-EQSRNYYT
Daphnia_pulex_321450949            --NLKFVNWQTHAIKETNSAS----LVTLGSWSEHAQSDAY-EQSRNYYT
Daphnia_pulex_321465382            KQIQLFINWQAAAIKAADPGA----LVTVGTWSQYSQTDVF-SNTRNYYT
Daphnia_pulex_321450057            KQIQLFINWQAAAIKAADPGA----LVTVGTWSQYSQTDVF-SNTRNYYT
Daphnia_pulex_321465383            KQIQLFINWQAAAIKAADPGA----LVTVGTWSQYSQTDVF-SDTRNYYT
Cryptopygus_antarcticus_157703     KQILKTINWISSAIHNADSKA----LVTVGSWSELTQTDSF-G-YRNHYK
Limnoria_quadripunctata_293629     ERFLNLIGKMNQVIRSNDSGG----LCTLGSWAQFSQTDAF-SNTKNHYT
Limnoria_quadripunctata_293629     ERFLILFGKMNQVIRANDPSG----IVTIGAYSQFSTTDAF-SDTTNHYT
Limnoria_quadripunctata_293629     ERFLILIGKMNQLIRELDPQA----ITTQGSWGQWSETDAF-SDTRNHYT
Callosobruchus_maculatus_31557     EKILKLINWHADAIKSVDPKA----LVTSADNGEFTTTTVC-EKCRDHYT
Callosobruchus_maculatus_31557     RLVLKLINWHADAIKSVVPEA----LLSNAENGELLTTNVC-EKCRDHYT
Callosobruchus_maculatus_31557     QNILKLINWHADAIKSVDPKA----LVTNGESGEFTTTTIC-EKCRDHYS
Callosobruchus_maculatus_31557     KDILRFINWHADAIKFVDPKA----LCTIGGAGEWLTTNVS-PVTRDHYT
Gastrophysa_viridula_315570658     KDVLKFINLHADAIKSADPKA----LVTVGESSELTATTIC-EKCRDMYS
uncultured_bacterium_359755046     KYIQQFVNLVTGAIKREAPHN----LVTNGSWNFRVLT-DV-GGMMNYYR
Spirochaeta_thermophila_307717     QYVQRFINLVAGAIHREDPDA----LVTCGSG-MAYQT-DV-GGMINYYR
Saccharophagus_degradans_90021     KQLQKVQALMTAAIHQAALDINQVALVTTGSKSMGKYNSDI-AGGINLYR
Chloroflexus_aurantiacus_16384     AQMQRFIAEISGAIHRNSNQL-----TTVGSASMKWNSSTALGASGNFWR
Chloroflexus_aggregans_2198492     AQMQQFIAEIAGTIHRNSNQL-----TTVGSASMKWNSTGALGASGNFWN
Oscillochloris_trichoides_3097     VQMQRFIAEISGAIHRNSNQL-----TTVGSAAMRWNSDRSLGATGNVWK
Clostridium_papyrosolvens_3262     SAMRNYIKTTCEFIHQYAKQP-----VSFGSANMKW-----LGAQYDLWD
Clostridium_cellulolyticum_220     AAMRNYIKTTCDFIHQYAKQP-----VSFGSANMKW-----LGAQYDLWD
Clostridium_sp_373945115           SAMRNYIKTTCEFIHQYAKQP-----VSFGSANMKW-----LGAQYDLWD
Clostridium_josui_270288704        SAMRNYIKTTCDFIHQYAKQP-----VSFGSANMKW-----LGAQYDLWD
Clostridium_acetobutylicum_158     STLRSFVKDVVDCIHQYAKQP-----VSVGSASLKW-----LGEQYDFWS
Fibrobacter_succinogenes_26141     AKIQKFTNKVAAAIHTTNPEL----LVSTGSVNIKY----------QKHWN
Chlamydomonas_reinhardtii_1594     -NISEFETRPRNLFSDAELRR--------------------AFVLRNGSL
                                                  .                                .

Flammeovirga_yaeyamensis_26186     DAELIAAG---GDAEGILDYYQVHYYSWAGTTYS-----PFVHPASHWEL
Branchiostoma_floridae_2608077     DSCLQQAG---GLTSGVLDFYQIHTYSHNGNYGS---QAPFVVTDASHYP
Branchiostoma_floridae_2608148     DDCLRKAGD-YSYRTGVLDFYQIHTYSKSGSYGS---QAPFRVIMMIMCA
k_10972_ext                        DNCLIDAG---GDSLGVIDLQQMHTYSWEGAYTS---SSPLNVHNSAYNL
Mytilus_edulis_90108933            DHCLVKAG---GKQSGTLSFYQVHTYDWQNHFGN---ESPFKHSFSNFRL
Mytilus_edulis_47606432            DHCLVKAG---GKQSGTLSFYQVHTYDWQNHFGN---ESPFKHSFSNFRL
Haliotis_discus_211908630          DHCLTQAG---GKAQGVLQFYTVHSYGKR--FDN---LSPFKHQKSDYKL
Biomphalaria_glabrata_56462580     DNCL-VKL---GKPNGKLDFYQFHTYSYQGNFDN---VSPFKHSAGDYGT
Biomphalaria_glabrata_56462582     DACL-LKG---GKPNGKLDFYQVHSYSYQGNFDN---VSPFKHSAGDFGT
Aplysia_kurodai_317414223          DHCLRLAG---GKQKGVFDFYQFHSYSWQGKWDE---VAPFTHQASDYGL
Haliotis_discus_85658727           DACLIKAG---GKKQGVLDFFQIHSYDWQGKFDE---VSPFTMAASVYHM
```

```
Daphnia_pulex_321460555          DSCLVAAG---GKANGKLDFYQMHTYAFNGQWGP---DAPFKVSSSSYGL
Daphnia_pulex_321460556          DSCLVGAG---GKAAGKLDFYQMHTYDYNGQWNS---DAPFTVSASSYGL
Daphnia_pulex_321460557          DACLLAAG---GRSLGTLDFYQFHTYTYTGQWDP---SEPFKVTATSYKL
Daphnia_pulex_321450949          DACLLAAG---GRSLGTLDFYQFHTYTYTGQWDP---SEPFKVTATSYKL
Daphnia_pulex_321465382          DACLIAAG---GKTLGKLDFYQIHTYTP---FSA---SAPFKVAASAYGL
Daphnia_pulex_321450057          DACLIAAG---GKTLGKLDFYQIHTYTP---FSA---SAPFKV-------
Daphnia_pulex_321465383          DACLVAAG---GKTLGKLDFYQIHTYTP---FSA---SAPLKVAASAYGL
Cryptopygus_antarcticus_157703   DSCLTGAG---GKSNGIINFYQMHTYSHSGKWNQ---NAPFKVNRWAYNV
Limnoria_quadripunctata_293629   NQCLNGAG---G-SGSQLDFYQMHSYDWSGSWSP---NAPFTVQASDYND
Limnoria_quadripunctata_293629   DECLNGAA---G-SGSELDFYQVHTYDWQGSWPP---HGPFTLQASDFEL
Limnoria_quadripunctata_293629   DTCLNGAA---G-SGSQIDFYQMHAYDWNGEWSP---NAPFTVKASDYKV
Callosobruchus_maculatus_31557   DECLIGAG---GRAKGTIDFYALHSYTWEGRYQP---TSPFKHNFDFYNS
Callosobruchus_maculatus_31557   DECLIGAG---GRANGTIDFYAMHSYTWEGRFAP---TSPFLHNFDFYKS
Callosobruchus_maculatus_31557   DECLIGAG---GRAKGTIDFYAMHSYTWEGRYQP---TSPFKHNFDFYKK
Callosobruchus_maculatus_31557   DACLIAAG---GRQLGTLDMVMVHTYTFQGRFVSD--TCPFKKRFLDYHT
Gastrophysa_viridula_315570658   DSCLVGAG---GKALGTIDFYQLHSYTWNGAFST---SSPFKNAAAAFKS
uncultured_bacterium_359755046   DDRLIEAGG---DTLGVLDFYQVHFYP-VHFDES---TSPFHKPASYWEL
Spirochaeta_thermophila_307717   DDRLVAAGG---DPEGTLDFYSVHFYP-QHMDES---VSPFHHPASYWQL
Saccharophagus_degradans_90021   DDRMIAAAGG--NPLATLDFYAPHYYNNESKHGA---WSPFHHHVDYWQV
Chloroflexus_aurantiacus_16384   DAALTAYD-----PQGYLDFYQIHYYGWMNGDEQY--WSYSPLFNDWYEA
Chloroflexus_aggregans_2198492   DTALTAYD-----PQGYLDFYQIHYYGWMNGDETY--WSYSPLFNDWYEG
Oscillochloris_trichoides_3097   DAALTPYD-----AKGYLDFYQIHYYGWMNGDGVY--WSYSPTLIDWATA
Clostridium_papyrosolvens_3262   GLG--------------LDFYDFHWYDWATP---Y--FNPVTTPASSLK-
Clostridium_cellulolyticum_220   GLG--------------LDFYDFHWYDWATP---Y--FNPVTTPASSLK-
Clostridium_sp_373945115         GLG--------------LDFYDFHWYDWATP---Y--FNPVTTPASSLK-
Clostridium_josui_270288704      GLG--------------LDFYDFHWYDWATP---Y--FNPVTTPASSLK-
Clostridium_acetobutylicum_158   GLG--------------LDFYDFHWYDWATP---Y--FNPLKTPVSQLKA
Fibrobacter_succinogenes_26141   DAALIEAG---GEANGTLDFFQTHYYPYWQDNSVS--PFVNTAAQMATKY
Chlamydomonas_reinhardtii_1594   GRGWDAAG------GGTLDFYAPHGYPYWGHDSITRLISPFHVPAQQYQL
                                 .               :.    * *

Flammeovirga_yaeyamensis_26186   DK---PLMIGEFYVEN-----------QPGSIAKEDLFPILYNNGYAGAW
Branchiostoma_floridae_2608077   ELSGKPIVIGEFSQAR------------GAGMTITEQFSRAYSHGFAGAW
Branchiostoma_floridae_2608148   RLSLYNSMLRLTAP-----------------------------------
k_10972_ext                      NK---PNILGEFSQSG------------GDGRSIQEQFDWAYTQGYCGAW
Mytilus_edulis_90108933          KK---PMVIGEFNQEH------------GAGMSSESMFEWAYTKGYSGAW
Mytilus_edulis_47606432          KK---PMVIGEFNQEH------------GAGMSSESMFEWAYTKGYSGAW
Haliotis_discus_211908630        NK---PLMVGEFASKN------------GGGMAIESMFQYAYGHGYCGAW
Biomphalaria_glabrata_56462580   GK---PIVVGEFWEQD------------GGGMNIDQLFDYVYNHGYAGAW
Biomphalaria_glabrata_56462582   GK---PIVVGEFWEQD------------GGGMNINQLFEYVYNHGYAGAW
Aplysia_kurodai_317414223        HK---PIVVGEFWEQD------------GGGMTITQMFNYVYNHGYAGAW
```

```
Haliotis_discus_85658727          DK---PIVIGEFRESQ------------GAGMTIQEMFNHAYNTGYSGAW
Daphnia_pulex_321460555           N---KPLVIGEFASVC------------AQNEGIQNLFQYGYTNGYQGVW
Daphnia_pulex_321460556           N---KPLVVGEFASVC------------SQNSDIGNMFQYVYDNGYQGAW
Daphnia_pulex_321460557           D---KPLVIGEFATVC------------GGPESSPTLFQYSYDNGYQGVW
Daphnia_pulex_321450949           D---KPLVIGEFATVC------------GGPESSPTLFQYSYDNGYQGVW
Daphnia_pulex_321465382           N---KPVTIGEFSASC------------SGGMTIQQMYNYAYGNGYQAAW
Daphnia_pulex_321450057           -------------------------------GLLIIIVFK-----------
Daphnia_pulex_321465383           N---KPVTIGEFSASC------------SGGMTIQQMYNYAYGNGYQAAW
Cryptopygus_antarcticus_157703    ND--KPLLIGEFASVC------------SQNEGIQNLYKYAYNNGYNGAL
Limnoria_quadripunctata_293629    N---KPILLGEYAGSC------------GAGTALADLHEYAYENGYVGGL
Limnoria_quadripunctata_293629    D---KPFLIGEYSGDC------------GAGNTLSELNTYAYENGYVGGL
Limnoria_quadripunctata_293629    D---KPILLGEYAGVC------------AAGTSLEDLNIYAYENGYVGGF
Callosobruchus_maculatus_31557    K---KPYLMEEFSTTN------------SESHSPSWNYHHIYEGGFGGIL
Callosobruchus_maculatus_31557    K---KPILMQEFSTTI------------TESHNASWNYRHIYEGDYVGIM
Callosobruchus_maculatus_31557    N---KPFVVEEFSTTN------------SESHSPVWNYHHIYEGGFGGIL
Callosobruchus_maculatus_31557    T---KPMVIEEFSTAC------------NECHDAVANYRYLYDSGYSGAL
Gastrophysa_viridula_315570658    D---KPIVVGEFATCC------------SELQDSAKNYQYLYNSGFSGAL
uncultured_bacterium_359755046    DK---PILIGEFPAYGVLAKSGQR-FRPRTELNAEEAWVYALENGYAGAL
Spirochaeta_thermophila_307717    DK---PIVVAEFPAKG-IREIGFG-FRPKTSLTTEEAYLWLIENGYAGAL
Saccharophagus_degradans_90021    TK---PVVIGEFHANETLDVLN-------DPVKAEDLCSRLIDNGYAGGW
Chloroflexus_aurantiacus_16384    G-FDKPVVVGELPANA---------GGT--NRTPAQLLTELHANCYAGAW
Chloroflexus_aggregans_2198492    R-FDKPVVIGEVPANA---------GGT--NRTPTQLIAELHANCYAGVW
Oscillochloris_trichoides_3097    G-FDKPTVIGEFPANA---------GET--GYTPAGLLEKLHSNCYGGAW
Clostridium_papyrosolvens_3262    --LDKPVIIGEMMPDT---------QGSSLKMTHKQVLDAIYKNGYAGYM
Clostridium_cellulolyticum_220    --LDKPVIIGEMMPDT---------QSSSLKMTHKQVLDAIYKNGYAGYM
Clostridium_sp_373945115          --LDKPVIIGEMMPDT---------QGSSLKMTHKQVLDAIYKNGYAGYM
Clostridium_josui_270288704       --LDKPVIIGEMMPDT---------LSSSLKMTHKQVLDAIYKNGYAGYM
Clostridium_acetobutylicum_158    K-FDKPVIIGEMMPDT---------QNSSLKMSHKQVLDGLVNNGYSGYM
Fibrobacter_succinogenes_26141    SYDSKPMIIGEFPASGWAGDTYRSNFAAKTEITTEECYRKAFDGGYAGAL
Chlamydomonas_reinhardtii_1594    D---KPALVGEFWDQVS----------DSESLTAKHWKDLWRKNGYMGEP


Flammeovirga_yaeyamensis_26186    GWQYRTYDDGD--NTDEMHRDDMLDGIKTIDGYPEIAIDPDRNHRPTLIG
Branchiostoma_floridae_2608077    SWHYLADR-AD--DTATDASATQLIGLRELRFKNDQT--RGGCVRINLNG
Branchiostoma_floridae_2608148    --------------------------------------------------
k_10972_ext                       SWQANAG------GEHADSFATQALGLNHLRGRNDQN--TGGRVDIDLQ-
Mytilus_edulis_90108933           TWSRTD-----------VSWNNQLRGMQHLKSRTDH-----GQVQFGL--
Mytilus_edulis_47606432           TWSRTD-----------VSWNNQLRGIQHLKSRTDH-----GQVQFGL--
Haliotis_discus_211908630         SWSATDN------YE-GDAWETQKRGVASIRNNSDAS--K-GTVHFTL--
Biomphalaria_glabrata_56462580    SWDLMAH------GD------NQRGGISHIKNYNWN-----GQIGINL--
Biomphalaria_glabrata_56462582    SWDLQAH------GA------NQRGGISHIKGLTSN-----GVIPINV--
```

```
Aplysia_kurodai_317414223          SWHLVQR------GD------NQRKGITNIKDKTSN-----GKIPISL--
Haliotis_discus_85658727           TWAITDD------WTPKDTWAHQQVGITTVANRHDH-----GLVKFTL--
Daphnia_pulex_321460555            SWQYNAG------GECSDTQATQDSGMNQLKGQN----GAGGAVNFPVGF
Daphnia_pulex_321460556            SWHYLEQ------GKCTDSQEAQNIGMTRIKDQT-----ANGAVTFPL--
Daphnia_pulex_321460557            SWSYNGGP-TG--STCCDNQTTQDSGMLQLKGQNN---GIGGAVNFPIVP
Daphnia_pulex_321450949            SWSYNGGP-TG--STCCDNQTTQDSGMLQLKGQN----GAGGAVNFPIVP
Daphnia_pulex_321465382            GWQYAGG-------YCSDSRATFDSGMLQLKGKS----GSGGLVNFPVV-
Daphnia_pulex_321450057            --------------------------------------------------
Daphnia_pulex_321465383            GWQYAGG-------YCSDSRATFDSGMLQLKGKS----GSGGLVNFPVV-
Cryptopygus_antarcticus_157703     TWQFNSG------GDCSDTYSNQMYGMQALKGQNDQSGGKGGMVSVNIN-
Limnoria_quadripunctata_293629     SWHWAAT------GDCSDSRAVQRSALGRLAGRT-----DNGVVDINVG-
Limnoria_quadripunctata_293629     SWHWAAT------GDCSDTRAVQRQALGTLVDRT-----DNGVVDFIVG-
Limnoria_quadripunctata_293629     AWCWLT-------GTCSDSRQEQRQALGALSGRT-----DYGTVDFIVG-
Callosobruchus_maculatus_31557     SWQYNQW------GKWVDSKESMFEGMASIRNLT-----SNGKIDIKL--
Callosobruchus_maculatus_31557     SWQYNQW------GKWVDTKESMFEGMGAIRNLT-----SHGKINIKL--
Callosobruchus_maculatus_31557     SWQYNEE------GKWVDSKQSMFEGMSSIRNLT-----SNGKIDIKL--
Callosobruchus_maculatus_31557     AFQYNGP------GQCVDDHPVMFAGMSAIRNLN-----YNGRIDIRL--
Gastrophysa_viridula_315570658     SWQYNEG------GNCADPKSVIDQGMSAIKDYT-----YNGNVHVTL--
uncultured_bacterium_359755046     GWTWTNHD-GN--GGVKDAEPGMKKVLELAPERVVIDQDVMNESE-----
Spirochaeta_thermophila_307717     SWTWTGHD-GF--GNIYDAAPGISAVAMRYPEYARLNREGLDLSPKVVKP
Saccharophagus_degradans_90021     SWQWNEH----------------VEHLMHCQERAAIR--------------
Chloroflexus_aurantiacus_16384     VWPYFNVNDGT--GQWSDAQAAVRSLSNAVPHEVQILR-------------
Chloroflexus_aggregans_2198492     VWPYFNVSDGT--GQWSDAASAVQAIADTVPAEVRLT--------------
Oscillochloris_trichoides_3097     AWSYENV-DGA--GGWNDIAAAYKAFNTTYAREVNITTGGTPNPTATPVV
Clostridium_papyrosolvens_3262     LWSWNDG--AF--DCKPYVGNNFIDFAAEHPDVVK----------------
Clostridium_cellulolyticum_220     LWSWNDG--AF--DCKPYVGNNFIDFAAEHPDVVR----------------
Clostridium_sp_373945115           LWSWNDG--AF--DCKPYVGNNFIDFAVEHPDVVR----------------
Clostridium_josui_270288704        LWSWNDG--AF--DCKPYIENNFIDFAAEHTDVVR----------------
Clostridium_acetobutylicum_158     LWAWTDA--SV--NCVGKTAPDFDEFKTEHPELPIDMP---------TMP
Fibrobacter_succinogenes_26141     AWQYIGDKTEANFGGYSYTIDPALKAMTALAATEEASIKIKDVDISGSTG
Chlamydomonas_reinhardtii_1594     -WAGAHD-------------RPPQRALC-----------------------


Flammeovirga_yaeyamensis_26186     EINNFRIDKNSNPLVNYTDLNTVFSDPEGTALQFSVETSPQNVVIPSVNN
Branchiostoma_floridae_2608077     STNYCEKKPPCRQLYQYFYG--------------------------------
Branchiostoma_floridae_2608148     --------------------------------------------------
k_10972_ext                        --------------------------------------------------
Mytilus_edulis_90108933            --------------------------------------------------
Mytilus_edulis_47606432            --------------------------------------------------
Haliotis_discus_211908630          --------------------------------------------------
Biomphalaria_glabrata_56462580     --------------------------------------------------
```

```
Biomphalaria_glabrata_56462582      --------------------------------------------------
Aplysia_kurodai_317414223           --------------------------------------------------
Haliotis_discus_85658727            --------------------------------------------------
Daphnia_pulex_321460555             --------------------------------------------------
Daphnia_pulex_321460556             --------------------------------------------------
Daphnia_pulex_321460557             --------------------------------------------------
Daphnia_pulex_321450949             --------------------------------------------------
Daphnia_pulex_321465382             --------------------------------------------------
Daphnia_pulex_321450057             --------------------------------------------------
Daphnia_pulex_321465383             --------------------------------------------------
Cryptopygus_antarcticus_157703      --------------------------------------------------
Limnoria_quadripunctata_293629      --------------------------------------------------
Limnoria_quadripunctata_293629      --------------------------------------------------
Limnoria_quadripunctata_293629      --------------------------------------------------
Callosobruchus_maculatus_31557      --------------------------------------------------
Callosobruchus_maculatus_31557      --------------------------------------------------
Callosobruchus_maculatus_31557      --------------------------------------------------
Callosobruchus_maculatus_31557      --------------------------------------------------
Gastrophysa_viridula_315570658      --------------------------------------------------
uncultured_bacterium_359755046      --------------------------------------------------
Spirochaeta_thermophila_307717      ISHLILSVNEVGKAEEIDLREVFEDQEDGDDLSYEVKKVGDPALVEVSLT
Saccharophagus_degradans_90021      --------------------------------------------------
Chloroflexus_aurantiacus_16384      ------------SPSTLPVRVYIPLTTR----------------------
Chloroflexus_aggregans_2198492      --------------ATILSPRVYIPLALRQQP------------------
Oscillochloris_trichoides_3097      PTATPAPVPPTATPAPVPPTATPVPPTPTPAPATTTQIYTDNLAAGWVNW
Clostridium_papyrosolvens_3262      --------------------------------------------------
Clostridium_cellulolyticum_220      --------------------------------------------------
Clostridium_sp_373945115            --------------------------------------------------
Clostridium_josui_270288704         --------------------------------------------------
Clostridium_acetobutylicum_158      GLKGDVNGDGVINGRDLMVLRQYLAGQSVNINKANTDVNGDGVVNGRDLM
Fibrobacter_succinogenes_26141      GNGMMAVTYGADNGQVEYQNKGGWDLSGATTFTWTAKNNGKTDADIYLIF
Chlamydomonas_reinhardtii_1594      --------------------------------------------------


Flammeovirga_yaeyamensis_26186      GFLSLAFAEEATGDVIVSITATDSGTPTTLRSYDFIVSVKEPGTGNLALY
Branchiostoma_floridae_2608077      --------------------------------------------------
Branchiostoma_floridae_2608148      --------------------------------------------------
k_10972_ext                         --------------------------------------------------
Mytilus_edulis_90108933             --------------------------------------------------
Mytilus_edulis_47606432             --------------------------------------------------
Haliotis_discus_211908630           --------------------------------------------------
```

```
Biomphalaria_glabrata_56462580      --------------------------------------------------
Biomphalaria_glabrata_56462582      --------------------------------------------------
Aplysia_kurodai_317414223           --------------------------------------------------
Haliotis_discus_85658727            --------------------------------------------------
Daphnia_pulex_321460555             --------------------------------------------------
Daphnia_pulex_321460556             --------------------------------------------------
Daphnia_pulex_321460557             --------------------------------------------------
Daphnia_pulex_321450949             --------------------------------------------------
Daphnia_pulex_321465382             --------------------------------------------------
Daphnia_pulex_321450057             --------------------------------------------------
Daphnia_pulex_321465383             --------------------------------------------------
Cryptopygus_antarcticus_157703      --------------------------------------------------
Limnoria_quadripunctata_293629      --------------------------------------------------
Limnoria_quadripunctata_293629      --------------------------------------------------
Limnoria_quadripunctata_293629      --------------------------------------------------
Callosobruchus_maculatus_31557      --------------------------------------------------
Callosobruchus_maculatus_31557      --------------------------------------------------
Callosobruchus_maculatus_31557      --------------------------------------------------
Callosobruchus_maculatus_31557      --------------------------------------------------
Gastrophysa_viridula_315570658      --------------------------------------------------
uncultured_bacterium_359755046      --------------------------------------------------
Spirochaeta_thermophila_307717      KDGKALVRLKEARVGSSDVLIAAKDSGMNESGLYFTVHALDPDRGNIALF
Saccharophagus_degradans_90021      --------------------------------------------------
Chloroflexus_aurantiacus_16384      --------------------------------------------------
Chloroflexus_aggregans_2198492      --------------------------------------------------
Oscillochloris_trichoides_3097      SWDSTVNFGSTAKKKVGTRSIAVTYNRAWAGLYLHTDQALNTQGYTKVRF
Clostridium_papyrosolvens_3262      --------------------------------------------------
Clostridium_cellulolyticum_220      --------------------------------------------------
Clostridium_sp_373945115            --------------------------------------------------
Clostridium_josui_270288704         --------------------------------------------------
Clostridium_acetobutylicum_158      EIVKIVMSK-----------------------------------------
Fibrobacter_succinogenes_26141      KLTDSWTWTETDGSCKVPAGEKVTCSIDISSFADRNKTLSITLANYASGY
Chlamydomonas_reinhardtii_1594      --------------------------------------------------


Flammeovirga_yaeyamensis_26186      QSTSASSVEEGPNTAASVNDGDQLTRWSSLYVDPSWIAIAFDQEYSVNEV
Branchiostoma_floridae_2608077      --------------------------------------------------
Branchiostoma_floridae_2608148      --------------------------------------------------
k_10972_ext                         --------------------------------------------------
Mytilus_edulis_90108933             --------------------------------------------------
Mytilus_edulis_47606432             --------------------------------------------------
```

```
Haliotis_discus_211908630          --------------------------------------------------
Biomphalaria_glabrata_56462580     --------------------------------------------------
Biomphalaria_glabrata_56462582     --------------------------------------------------
Aplysia_kurodai_317414223          --------------------------------------------------
Haliotis_discus_85658727           --------------------------------------------------
Daphnia_pulex_321460555            --------------------------------------------------
Daphnia_pulex_321460556            --------------------------------------------------
Daphnia_pulex_321460557            --------------------------------------------------
Daphnia_pulex_321450949            --------------------------------------------------
Daphnia_pulex_321465382            --------------------------------------------------
Daphnia_pulex_321450057            --------------------------------------------------
Daphnia_pulex_321465383            --------------------------------------------------
Cryptopygus_antarcticus_157703     --------------------------------------------------
Limnoria_quadripunctata_293629     --------------------------------------------------
Limnoria_quadripunctata_293629     --------------------------------------------------
Limnoria_quadripunctata_293629     --------------------------------------------------
Callosobruchus_maculatus_31557     --------------------------------------------------
Callosobruchus_maculatus_31557     --------------------------------------------------
Callosobruchus_maculatus_31557     --------------------------------------------------
Callosobruchus_maculatus_31557     --------------------------------------------------
Gastrophysa_viridula_315570658     --------------------------------------------------
uncultured_bacterium_359755046     --------------------------------------------------
Spirochaeta_thermophila_307717     KPVEASSV-EDPNLPEYVNDGTLKTRWSSLYEDDEYIQIDLQGRFRIERI
Saccharophagus_degradans_90021     --------------------------------------------------
Chloroflexus_aurantiacus_16384     --------------------------------------------------
Chloroflexus_aggregans_2198492     --------------------------------------------------
Oscillochloris_trichoides_3097     WAHGAGGKQKLKLWVRKADGTTSPTVALPTLTSSWMQIEVPLSQLGNPAN
Clostridium_papyrosolvens_3262     --------------------------------------------------
Clostridium_cellulolyticum_220     --------------------------------------------------
Clostridium_sp_373945115           --------------------------------------------------
Clostridium_josui_270288704        --------------------------------------------------
Clostridium_acetobutylicum_158     --------------------------------------------------
Fibrobacter_succinogenes_26141     TGTVIYDDIKAGDLTLFDFNTDKYDAFKRGYENTEEMIPEIKIVFDENYV
Chlamydomonas_reinhardtii_1594     --------------------------------------------------


Flammeovirga_yaeyamensis_26186     RLYWEGSYSSQYEIQVSQDGESWNTVYTNNDGRGGEDIITFPAVNAKHIR
Branchiostoma_floridae_2608077     --------------------------------------------------
Branchiostoma_floridae_2608148     --------------------------------------------------
k_10972_ext                        --------------------------------------------------
Mytilus_edulis_90108933            --------------------------------------------------
```

```
Mytilus_edulis_47606432          ------------------------------------------------
Haliotis_discus_211908630        ------------------------------------------------
Biomphalaria_glabrata_56462580   ------------------------------------------------
Biomphalaria_glabrata_56462582   ------------------------------------------------
Aplysia_kurodai_317414223        ------------------------------------------------
Haliotis_discus_85658727         ------------------------------------------------
Daphnia_pulex_321460555          ------------------------------------------------
Daphnia_pulex_321460556          ------------------------------------------------
Daphnia_pulex_321460557          ------------------------------------------------
Daphnia_pulex_321450949          ------------------------------------------------
Daphnia_pulex_321465382          ------------------------------------------------
Daphnia_pulex_321450057          ------------------------------------------------
Daphnia_pulex_321465383          ------------------------------------------------
Cryptopygus_antarcticus_157703   ------------------------------------------------
Limnoria_quadripunctata_293629   ------------------------------------------------
Limnoria_quadripunctata_293629   ------------------------------------------------
Limnoria_quadripunctata_293629   ------------------------------------------------
Callosobruchus_maculatus_31557   ------------------------------------------------
Callosobruchus_maculatus_31557   ------------------------------------------------
Callosobruchus_maculatus_31557   ------------------------------------------------
Callosobruchus_maculatus_31557   ------------------------------------------------
Gastrophysa_viridula_315570658   ------------------------------------------------
uncultured_bacterium_359755046   ------------------------------------------------
Spirochaeta_thermophila_307717   VLHWEVAYGKDYDILGSLDGKAWFPIVQVRGGDGDVDELSFDPVEVSYVR
Saccharophagus_degradans_90021   ------------------------------------------------
Chloroflexus_aurantiacus_16384   ------------------------------------------------
Chloroflexus_aggregans_2198492   ------------------------------------------------
Oscillochloris_trichoides_3097   LSDLVIQDAAGRAQAVFYIDQIELVP------------------------
Clostridium_papyrosolvens_3262   ------------------------------------------------
Clostridium_cellulolyticum_220   ------------------------------------------------
Clostridium_sp_373945115         ------------------------------------------------
Clostridium_josui_270288704      ------------------------------------------------
Clostridium_acetobutylicum_158   ------------------------------------------------
Fibrobacter_succinogenes_26141   YGKTSISKSKLAATSKFSINGDKITLNTKAKGQVSVDVFGMNG-RIVATL
Chlamydomonas_reinhardtii_1594   ------------------------------------------------


Flammeovirga_yaeyamensis_26186   MYGKQRAIEWGHSIYEFEVYGDVPKSISANAGEDQFVSDSDGDGVETITL
Branchiostoma_floridae_2608077   ------------------------------------------------
Branchiostoma_floridae_2608148   ------------------------------------------------
k_10972_ext                      ------------------------------------------------
```

```
Mytilus_edulis_90108933          ------------------------------------------------
Mytilus_edulis_47606432          ------------------------------------------------
Haliotis_discus_211908630        ------------------------------------------------
Biomphalaria_glabrata_56462580   ------------------------------------------------
Biomphalaria_glabrata_56462582   ------------------------------------------------
Aplysia_kurodai_317414223        ------------------------------------------------
Haliotis_discus_85658727         ------------------------------------------------
Daphnia_pulex_321460555          ------------------------------------------------
Daphnia_pulex_321460556          ------------------------------------------------
Daphnia_pulex_321460557          ------------------------------------------------
Daphnia_pulex_321450949          ------------------------------------------------
Daphnia_pulex_321465382          ------------------------------------------------
Daphnia_pulex_321450057          ------------------------------------------------
Daphnia_pulex_321465383          ------------------------------------------------
Cryptopygus_antarcticus_157703   ------------------------------------------------
Limnoria_quadripunctata_293629   ------------------------------------------------
Limnoria_quadripunctata_293629   ------------------------------------------------
Limnoria_quadripunctata_293629   ------------------------------------------------
Callosobruchus_maculatus_31557   ------------------------------------------------
Callosobruchus_maculatus_31557   ------------------------------------------------
Callosobruchus_maculatus_31557   ------------------------------------------------
Callosobruchus_maculatus_31557   ------------------------------------------------
Gastrophysa_viridula_315570658   ------------------------------------------------
uncultured_bacterium_359755046   ------------------------------------------------
Spirochaeta_thermophila_307717   MHGITRGTEWGFSLWEMEVYGERVE------------------------
Saccharophagus_degradans_90021   ------------------------------------------------
Chloroflexus_aurantiacus_16384   ------------------------------------------------
Chloroflexus_aggregans_2198492   ------------------------------------------------
Oscillochloris_trichoides_3097   ------------------------------------------------
Clostridium_papyrosolvens_3262   ------------------------------------------------
Clostridium_cellulolyticum_220   ------------------------------------------------
Clostridium_sp_373945115         ------------------------------------------------
Clostridium_josui_270288704      ------------------------------------------------
Clostridium_acetobutylicum_158   ------------------------------------------------
Fibrobacter_succinogenes_26141   FNGMLGAGNYVFSLADMPKGQYIIRMKGAGITTTQPVIVK----------
Chlamydomonas_reinhardtii_1594   ------------------------------------------------


Flammeovirga_yaeyamensis_26186   DGTASTDSEQTITSYEWSENGVSLASGATANVPLGVGIHTITLTVTNNLG
Branchiostoma_floridae_2608077   ------------------------------------------------
Branchiostoma_floridae_2608148   ------------------------------------------------
```

```
k_10972_ext                          --------------------------------------------------
Mytilus_edulis_90108933              --------------------------------------------------
Mytilus_edulis_47606432              --------------------------------------------------
Haliotis_discus_211908630            --------------------------------------------------
Biomphalaria_glabrata_56462580       --------------------------------------------------
Biomphalaria_glabrata_56462582       --------------------------------------------------
Aplysia_kurodai_317414223            --------------------------------------------------
Haliotis_discus_85658727             --------------------------------------------------
Daphnia_pulex_321460555              --------------------------------------------------
Daphnia_pulex_321460556              --------------------------------------------------
Daphnia_pulex_321460557              --------------------------------------------------
Daphnia_pulex_321450949              --------------------------------------------------
Daphnia_pulex_321465382              --------------------------------------------------
Daphnia_pulex_321450057              --------------------------------------------------
Daphnia_pulex_321465383              --------------------------------------------------
Cryptopygus_antarcticus_157703       --------------------------------------------------
Limnoria_quadripunctata_293629       --------------------------------------------------
Limnoria_quadripunctata_293629       --------------------------------------------------
Limnoria_quadripunctata_293629       --------------------------------------------------
Callosobruchus_maculatus_31557       --------------------------------------------------
Callosobruchus_maculatus_31557       --------------------------------------------------
Callosobruchus_maculatus_31557       --------------------------------------------------
Callosobruchus_maculatus_31557       --------------------------------------------------
Gastrophysa_viridula_315570658       --------------------------------------------------
uncultured_bacterium_359755046       --------------------------------------------------
Spirochaeta_thermophila_307717       --------------------------------------------------
Saccharophagus_degradans_90021       --------------------------------------------------
Chloroflexus_aurantiacus_16384       --------------------------------------------------
Chloroflexus_aggregans_2198492       --------------------------------------------------
Oscillochloris_trichoides_3097       --------------------------------------------------
Clostridium_papyrosolvens_3262       --------------------------------------------------
Clostridium_cellulolyticum_220       --------------------------------------------------
Clostridium_sp_373945115             --------------------------------------------------
Clostridium_josui_270288704          --------------------------------------------------
Clostridium_acetobutylicum_158       --------------------------------------------------
Fibrobacter_succinogenes_26141       --------------------------------------------------
Chlamydomonas_reinhardtii_1594       --------------------------------------------------


Flammeovirga_yaeyamensis_26186       VSSTDNVRIIVDDGSFGPQIFEAENATLSSVTVASDATASEGAYVNMEGN
Branchiostoma_floridae_2608077       --------------------------------------------------
```

```
Branchiostoma_floridae_2608148      --------------------------------------------------
k_10972_ext                         --------------------------------------------------
Mytilus_edulis_90108933             --------------------------------------------------
Mytilus_edulis_47606432             --------------------------------------------------
Haliotis_discus_211908630           --------------------------------------------------
Biomphalaria_glabrata_56462580      --------------------------------------------------
Biomphalaria_glabrata_56462582      --------------------------------------------------
Aplysia_kurodai_317414223           --------------------------------------------------
Haliotis_discus_85658727            --------------------------------------------------
Daphnia_pulex_321460555             --------------------------------------------------
Daphnia_pulex_321460556             --------------------------------------------------
Daphnia_pulex_321460557             --------------------------------------------------
Daphnia_pulex_321450949             --------------------------------------------------
Daphnia_pulex_321465382             --------------------------------------------------
Daphnia_pulex_321450057             --------------------------------------------------
Daphnia_pulex_321465383             --------------------------------------------------
Cryptopygus_antarcticus_157703      --------------------------------------------------
Limnoria_quadripunctata_293629      --------------------------------------------------
Limnoria_quadripunctata_293629      --------------------------------------------------
Limnoria_quadripunctata_293629      --------------------------------------------------
Callosobruchus_maculatus_31557      --------------------------------------------------
Callosobruchus_maculatus_31557      --------------------------------------------------
Callosobruchus_maculatus_31557      --------------------------------------------------
Callosobruchus_maculatus_31557      --------------------------------------------------
Gastrophysa_viridula_315570658      --------------------------------------------------
uncultured_bacterium_359755046      --------------------------------------------------
Spirochaeta_thermophila_307717      --------------------------------------------------
Saccharophagus_degradans_90021      --------------------------------------------------
Chloroflexus_aurantiacus_16384      --------------------------------------------------
Chloroflexus_aggregans_2198492      --------------------------------------------------
Oscillochloris_trichoides_3097      --------------------------------------------------
Clostridium_papyrosolvens_3262      --------------------------------------------------
Clostridium_cellulolyticum_220      --------------------------------------------------
Clostridium_sp_373945115            --------------------------------------------------
Clostridium_josui_270288704         --------------------------------------------------
Clostridium_acetobutylicum_158      --------------------------------------------------
Fibrobacter_succinogenes_26141      --------------------------------------------------
Chlamydomonas_reinhardtii_1594      --------------------------------------------------


Flammeovirga_yaeyamensis_26186      GTISWTFNAVSAGTQTIKIGYLLPYGSKDQYVSLNGNSLGSIPFDGPINT
```

```
Branchiostoma_floridae_2608077      ------------------------------------------------
Branchiostoma_floridae_2608148      ------------------------------------------------
k_10972_ext                         ------------------------------------------------
Mytilus_edulis_90108933             ------------------------------------------------
Mytilus_edulis_47606432             ------------------------------------------------
Haliotis_discus_211908630           ------------------------------------------------
Biomphalaria_glabrata_56462580      ------------------------------------------------
Biomphalaria_glabrata_56462582      ------------------------------------------------
Aplysia_kurodai_317414223           ------------------------------------------------
Haliotis_discus_85658727            ------------------------------------------------
Daphnia_pulex_321460555             ------------------------------------------------
Daphnia_pulex_321460556             ------------------------------------------------
Daphnia_pulex_321460557             ------------------------------------------------
Daphnia_pulex_321450949             ------------------------------------------------
Daphnia_pulex_321465382             ------------------------------------------------
Daphnia_pulex_321450057             ------------------------------------------------
Daphnia_pulex_321465383             ------------------------------------------------
Cryptopygus_antarcticus_157703      ------------------------------------------------
Limnoria_quadripunctata_293629      ------------------------------------------------
Limnoria_quadripunctata_293629      ------------------------------------------------
Limnoria_quadripunctata_293629      ------------------------------------------------
Callosobruchus_maculatus_31557      ------------------------------------------------
Callosobruchus_maculatus_31557      ------------------------------------------------
Callosobruchus_maculatus_31557      ------------------------------------------------
Callosobruchus_maculatus_31557      ------------------------------------------------
Gastrophysa_viridula_315570658      ------------------------------------------------
uncultured_bacterium_359755046      ------------------------------------------------
Spirochaeta_thermophila_307717      ------------------------------------------------
Saccharophagus_degradans_90021      ------------------------------------------------
Chloroflexus_aurantiacus_16384      ------------------------------------------------
Chloroflexus_aggregans_2198492      ------------------------------------------------
Oscillochloris_trichoides_3097      ------------------------------------------------
Clostridium_papyrosolvens_3262      ------------------------------------------------
Clostridium_cellulolyticum_220      ------------------------------------------------
Clostridium_sp_373945115            ------------------------------------------------
Clostridium_josui_270288704         ------------------------------------------------
Clostridium_acetobutylicum_158      ------------------------------------------------
Fibrobacter_succinogenes_26141      ------------------------------------------------
Chlamydomonas_reinhardtii_1594      ------------------------------------------------
```

```
Flammeovirga_yaeyamensis_26186      WLEKSLTVDLVEGNNTLTITKHWGYMYFDYLSTGGVSNTRVVAIEPSETL
Branchiostoma_floridae_2608077      --------------------------------------------------
Branchiostoma_floridae_2608148      --------------------------------------------------
k_10972_ext                         --------------------------------------------------
Mytilus_edulis_90108933             --------------------------------------------------
Mytilus_edulis_47606432             --------------------------------------------------
Haliotis_discus_211908630           --------------------------------------------------
Biomphalaria_glabrata_56462580      --------------------------------------------------
Biomphalaria_glabrata_56462582      --------------------------------------------------
Aplysia_kurodai_317414223           --------------------------------------------------
Haliotis_discus_85658727            --------------------------------------------------
Daphnia_pulex_321460555             --------------------------------------------------
Daphnia_pulex_321460556             --------------------------------------------------
Daphnia_pulex_321460557             --------------------------------------------------
Daphnia_pulex_321450949             --------------------------------------------------
Daphnia_pulex_321465382             --------------------------------------------------
Daphnia_pulex_321450057             --------------------------------------------------
Daphnia_pulex_321465383             --------------------------------------------------
Cryptopygus_antarcticus_157703      --------------------------------------------------
Limnoria_quadripunctata_293629      --------------------------------------------------
Limnoria_quadripunctata_293629      --------------------------------------------------
Limnoria_quadripunctata_293629      --------------------------------------------------
Callosobruchus_maculatus_31557      --------------------------------------------------
Callosobruchus_maculatus_31557      --------------------------------------------------
Callosobruchus_maculatus_31557      --------------------------------------------------
Callosobruchus_maculatus_31557      --------------------------------------------------
Gastrophysa_viridula_315570658      --------------------------------------------------
uncultured_bacterium_359755046      --------------------------------------------------
Spirochaeta_thermophila_307717      --------------------------------------------------
Saccharophagus_degradans_90021      --------------------------------------------------
Chloroflexus_aurantiacus_16384      --------------------------------------------------
Chloroflexus_aggregans_2198492      --------------------------------------------------
Oscillochloris_trichoides_3097      --------------------------------------------------
Clostridium_papyrosolvens_3262      --------------------------------------------------
Clostridium_cellulolyticum_220      --------------------------------------------------
Clostridium_sp_373945115            --------------------------------------------------
Clostridium_josui_270288704         --------------------------------------------------
Clostridium_acetobutylicum_158      --------------------------------------------------
Fibrobacter_succinogenes_26141      --------------------------------------------------
Chlamydomonas_reinhardtii_1594      --------------------------------------------------
```

```
Flammeovirga_yaeyamensis_26186      AEVVLYPNPATTTFTVKADHFQSLTIYDMKGVALISSNLKNVDISELGSG
Branchiostoma_floridae_2608077      --------------------------------------------------
Branchiostoma_floridae_2608148      --------------------------------------------------
k_10972_ext                         --------------------------------------------------
Mytilus_edulis_90108933             --------------------------------------------------
Mytilus_edulis_47606432             --------------------------------------------------
Haliotis_discus_211908630           --------------------------------------------------
Biomphalaria_glabrata_56462580      --------------------------------------------------
Biomphalaria_glabrata_56462582      --------------------------------------------------
Aplysia_kurodai_317414223           --------------------------------------------------
Haliotis_discus_85658727            --------------------------------------------------
Daphnia_pulex_321460555             --------------------------------------------------
Daphnia_pulex_321460556             --------------------------------------------------
Daphnia_pulex_321460557             --------------------------------------------------
Daphnia_pulex_321450949             --------------------------------------------------
Daphnia_pulex_321465382             --------------------------------------------------
Daphnia_pulex_321450057             --------------------------------------------------
Daphnia_pulex_321465383             --------------------------------------------------
Cryptopygus_antarcticus_157703      --------------------------------------------------
Limnoria_quadripunctata_293629      --------------------------------------------------
Limnoria_quadripunctata_293629      --------------------------------------------------
Limnoria_quadripunctata_293629      --------------------------------------------------
Callosobruchus_maculatus_31557      --------------------------------------------------
Callosobruchus_maculatus_31557      --------------------------------------------------
Callosobruchus_maculatus_31557      --------------------------------------------------
Callosobruchus_maculatus_31557      --------------------------------------------------
Gastrophysa_viridula_315570658      --------------------------------------------------
uncultured_bacterium_359755046      --------------------------------------------------
Spirochaeta_thermophila_307717      --------------------------------------------------
Saccharophagus_degradans_90021      --------------------------------------------------
Chloroflexus_aurantiacus_16384      --------------------------------------------------
Chloroflexus_aggregans_2198492      --------------------------------------------------
Oscillochloris_trichoides_3097      --------------------------------------------------
Clostridium_papyrosolvens_3262      --------------------------------------------------
Clostridium_cellulolyticum_220      --------------------------------------------------
Clostridium_sp_373945115            --------------------------------------------------
Clostridium_josui_270288704         --------------------------------------------------
Clostridium_acetobutylicum_158      --------------------------------------------------
Fibrobacter_succinogenes_26141      --------------------------------------------------
Chlamydomonas_reinhardtii_1594      --------------------------------------------------
```

```
Flammeovirga_yaeyamensis_26186      IYLVKVYTLNGVKNLRLSVR
Branchiostoma_floridae_2608077      --------------------
Branchiostoma_floridae_2608148      --------------------
k_10972_ext                         --------------------
Mytilus_edulis_90108933             --------------------
Mytilus_edulis_47606432             --------------------
Haliotis_discus_211908630           --------------------
Biomphalaria_glabrata_56462580      --------------------
Biomphalaria_glabrata_56462582      --------------------
Aplysia_kurodai_317414223           --------------------
Haliotis_discus_85658727            --------------------
Daphnia_pulex_321460555             --------------------
Daphnia_pulex_321460556             --------------------
Daphnia_pulex_321460557             --------------------
Daphnia_pulex_321450949             --------------------
Daphnia_pulex_321465382             --------------------
Daphnia_pulex_321450057             --------------------
Daphnia_pulex_321465383             --------------------
Cryptopygus_antarcticus_157703      --------------------
Limnoria_quadripunctata_293629      --------------------
Limnoria_quadripunctata_293629      --------------------
Limnoria_quadripunctata_293629      --------------------
Callosobruchus_maculatus_31557      --------------------
Callosobruchus_maculatus_31557      --------------------
Callosobruchus_maculatus_31557      --------------------
Callosobruchus_maculatus_31557      --------------------
Gastrophysa_viridula_315570658      --------------------
uncultured_bacterium_359755046      --------------------
Spirochaeta_thermophila_307717      --------------------
Saccharophagus_degradans_90021      --------------------
Chloroflexus_aurantiacus_16384      --------------------
Chloroflexus_aggregans_2198492      --------------------
Oscillochloris_trichoides_3097      --------------------
Clostridium_papyrosolvens_3262      --------------------
Clostridium_cellulolyticum_220      --------------------
Clostridium_sp_373945115            --------------------
Clostridium_josui_270288704         --------------------
Clostridium_acetobutylicum_158      --------------------
Fibrobacter_succinogenes_26141      --------------------
```

```
Chlamydomonas_reinhardtii_1594      --------------------
```

Bibliography

[1] Genome sequence of the nematode c. elegans: a platform for investigating biology. *Science (New York, N.Y.)*, 282(5396):2012–2018, Dec. 1998. PMID: 9851916.

[2] G. Abrusan, N. Grundmann, L. DeMester, and W. Makalowski. TEclass–a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics*, 25(10):1329–1330, May 2009.

[3] M. D. Adams, S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, P. W. Li, R. A. Hoskins, R. F. Galle, R. A. George, S. E. Lewis, S. Richards, M. Ashburner, S. N. Henderson, G. G. Sutton, J. R. Wortman, M. D. Yandell, Q. Zhang, L. X. Chen, R. C. Brandon, Y.-H. C. Rogers, R. G. Blazej, M. Champe, B. D. Pfeiffer, K. H. Wan, C. Doyle, E. G. Baxter, G. Helt, C. R. Nelson, G. L. Gabor, Miklos, J. F. Abril, A. Agbayani, H.-J. An, C. Andrews-Pfannkoch, D. Baldwin, R. M. Ballew, A. Basu, J. Baxendale, L. Bayraktaroglu, E. M. Beasley, K. Y. Beeson, P. V. Benos, B. P. Berman, D. Bhandari, S. Bolshakov, D. Borkova, M. R. Botchan, J. Bouck, P. Brokstein, P. Brottier, K. C. Burtis, D. A. Busam, H. Butler, E. Cadieu, A. Center, I. Chandra, J. M. Cherry, S. Cawley, C. Dahlke, L. B. Davenport, P. Davies, B. d. Pablos, A. Delcher, Z. Deng, A. D. Mays, I. Dew, S. M. Dietz, K. Dodson, L. E. Doup, M. Downes, S. Dugan-Rocha, B. C. Dunkov, P. Dunn, K. J. Durbin, C. C. Evangelista, C. Ferraz, S. Ferriera, W. Fleischmann, C. Fosler, A. E. Gabrielian, N. S. Garg, W. M. Gelbart, K. Glasser, A. Glodek, F. Gong, J. H. Gorrell, Z. Gu, P. Guan, M. Harris, N. L. Harris, D. Harvey, T. J. Heiman, J. R. Hernandez, J. Houck, D. Hostin, K. A. Houston, T. J. How-

land, M.-H. Wei, C. Ibegwam, M. Jalali, F. Kalush, G. H. Karpen, Z. Ke, J. A. Kennison, K. A. Ketchum, B. E. Kimmel, C. D. Kodira, C. Kraft, S. Kravitz, D. Kulp, Z. Lai, P. Lasko, Y. Lei, A. A. Levitsky, J. Li, Z. Li, Y. Liang, X. Lin, X. Liu, B. Mattei, T. C. McIntosh, M. P. McLeod, D. McPherson, G. Merkulov, N. V. Milshina, C. Mobarry, J. Morris, A. Moshrefi, S. M. Mount, M. Moy, B. Murphy, L. Murphy, D. M. Muzny, D. L. Nelson, D. R. Nelson, K. A. Nelson, K. Nixon, D. R. Nusskern, J. M. Pacleb, M. Palazzolo, G. S. Pittman, S. Pan, J. Pollard, V. Puri, M. G. Reese, K. Reinert, K. Remington, R. D. C. Saunders, F. Scheeler, H. Shen, B. C. Shue, I. Sidén-Kiamos, M. Simpson, M. P. Skupski, T. Smith, E. Spier, A. C. Spradling, M. Stapleton, R. Strong, E. Sun, R. Svirskas, C. Tector, R. Turner, E. Venter, A. H. Wang, X. Wang, Z.-Y. Wang, D. A. Wassarman, G. M. Weinstock, J. Weissenbach, S. M. Williams, T. Woodage, K. C. Worley, D. Wu, S. Yang, Q. A. Yao, J. Ye, R.-F. Yeh, J. S. Zaveri, M. Zhan, G. Zhang, Q. Zhao, L. Zheng, X. H. Zheng, F. N. Zhong, W. Zhong, X. Zhou, S. Zhu, X. Zhu, H. O. Smith, R. A. Gibbs, E. W. Myers, G. M. Rubin, and J. C. Venter. The genome sequence of drosophila melanogaster. *Science*, 287(5461):2185–2195, Mar. 2000.

[4] D. Aird, M. G. Ross, W. Chen, M. Danielsson, T. Fennell, C. Russ, D. B. Jaffe, C. Nusbaum, and A. Gnirke. Analyzing and minimizing PCR amplification bias in illumina sequencing libraries. *Genome Biology*, 12(2):R18, 2011. PMID: 21338519.

[5] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, Sept. 1997. PMID: 9254694.

[6] J. Andre, R. A. King, S. R. Stürzenbaum, P. Kille, M. E. Hodson, and A. J. Morgan. Molecular genetic differentiation in earthworms inhabiting a heterogeneous pb-polluted landscape. *Environmental Pollution (Barking, Essex: 1987)*, 158(3):883–890, Mar. 2010. PMID: 19818541.

[7] D. Arendt, A. S. Denes, G. Jékely, and K. Tessmar-Raible. The evolution of nervous system centralization. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1496):1523 –1528, Apr. 2008.

[8] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genetics*, 25(1):25–29, May 2000. PMID: 10802651.

[9] A. Bairoch. The universal protein resource (UniProt). *Nucleic Acids Research*, 33(Database issue):D154–D159, Dec. 2004.

[10] A. Bairoch and R. Apweiler. The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Research*, 24(1):21 –25, Jan. 1996.

[11] Z. Bao and S. R. Eddy. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Research*, 12(8):1269–1276, Aug. 2002. PMID: 12176934 PMCID: 186642.

[12] S. Batzoglou, D. B. Jaffe, K. Stanley, J. Butler, S. Gnerre, E. Mauceli, B. Berger, J. P. Mesirov, and E. S. Lander. ARACHNE: a Whole-Genome shotgun assembler. *Genome Res.*, 12(1):177–189, Jan. 2002.

[13] A. Beloqui, T. Y. Nechitaylo, N. López-Cortés, A. Ghazi, M. Guazzaroni, J. Polaina, A. W. Strittmatter, O. Reva, A. Waliczek, M. M. Yakimov, O. V. Golyshina, M. Ferrer, and P. N. Golyshin. Diversity of glycosyl hydrolases from Cellulose-Depleting communities enriched from casts of two earthworm species. *Applied and Environmental Microbiology*, 76(17):5934 –5946, Sept. 2010.

[14] A. E. Bely. Distribution of segment regeneration ability in the annelida. *Integrative and Comparative Biology*, 46(4):508 –518, June 2006.

[15] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. GenBank. *Nucleic Acids Research*, 39(Database issue):D32–37, Jan. 2011. PMID: 21071399.

[16] G. Benson. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, 27(2):573–580, Jan. 1999. PMID: 9862982.

[17] M. Berriman, B. Haas, P. Loverde, A. Wilson, G. Dillon, G. Cerqueira, S. Mashiyama, B. Al-Lazikani, L. Andrade, P. Ashton, M. Aslett, D. Bartholomeu, G. Blandin, C. Caffrey, A. Coghlan, R. Coulson, T. Day, A. Delcher, R. Demarco, A. Djikeng, T. Eyre, J. Gamble, E. Ghedin, Y. Gu, C. Hertz-Fowler, H. Hirai, Y. Hirai, R. Houston, A. Ivens, D. Johnston, D. Lacerda, C. Macedo, P. Mcveigh, Z. Ning, G. Oliveira, J. Overington, J. Parkhill, M. Pertea, R. Pierce, A. Protasio, M. Quail, M. Rajandream, J. Rogers, M. Sajid, S. Salzberg, M. Stanke, A. Tivey, O. White, D. Williams, J. Wortman, W. Wu, M. Zamanian, A. Zerlotini, C. Fraser-Liggett, B. Barrell, and N. El-Sayed. The genome of the blood fluke schistosoma mansoni. *Nature*, 460(7253):358, 352, July 2009.

[18] F. Binet, A. Kersanté, C. Munier-Lamy, R. Le Bayon, M. Belgy, and M. J. Shipitalo. Lumbricid macrofauna alter atrazine mineralization and sorption in a silt loam soil. *Soil Biology and Biochemistry*, 38(6):1255–1263, June 2006.

[19] S. Boisvert, F. Laviolette, and J. Corbeil. Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, 17(11):1519–1533, Nov. 2010. PMID: 20958248.

[20] S. Boyer and S. D. Wratten. The potential of earthworms to restore ecosystem services after opencast mining - a review. *Basic and Applied Ecology*, 11(3):196–203, May 2010.

[21] R. Breathnach, C. Benoist, K. O'Hare, F. Gannon, and P. Chambon. Ovalbumin gene: evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries. *Proceedings of the National Academy of Sciences*, 75(10):4853 –4857, Oct. 1978.

[22] B. Brejova, T. Vinar, Y. Chen, S. Wang, G. Zhao, D. G. Brown, M. Li, and Y. Zhou. Finding genes in schistosoma japonicum: annotating novel genomes with help of extrinsic evidence. *Nucl. Acids Res.*, 37(7):e52, Apr. 2009.

[23] J. Butler, I. MacCallum, M. Kleber, I. A. Shlyakhter, M. K. Belmonte, E. S. Lander, C. Nusbaum, and D. B. Jaffe. ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Res.*, 18(5):810–820, May 2008.

[24] P. S. G. Chain, D. V. Grafham, R. S. Fulton, M. G. Fitzgerald, J. Hostetler, D. Muzny, J. Ali, B. Birren, D. C. Bruce, C. Buhay, J. R. Cole, Y. Ding, S. Dugan, D. Field, G. M. Garrity, R. Gibbs, T. Graves, C. S. Han, S. H. Harrison, S. Highlander, P. Hugenholtz, H. M. Khouri, C. D. Kodira, E. Kolker, N. C. Kyrpides, D. Lang, A. Lapidus, S. A. Malfatti, V. Markowitz, T. Metha, K. E. Nelson, J. Parkhill, S. Pitluck, X. Qin, T. D. Read, J. Schmutz, S. Sozhamannan, P. Sterk, R. L. Strausberg, G. Sutton, N. R. Thomson, J. M. Tiedje, G. Weinstock, A. Wollam, and J. C. Detter. Genomics. genome project standards in a new era of sequencing. *Science (New York, N.Y.)*, 326(5950):236–237, Oct. 2009. PMID: 19815760.

[25] J. H. Cho, C. B. Park, Y. G. Yoon, and S. C. Kim. Lumbricin i, a novel proline-rich antimicrobial peptide from the earthworm: purification, cDNA cloning and molecular characterization. *Biochimica Et Biophysica Acta*, 1408(1):67–76, Oct. 1998. PMID: 9784609.

[26] S. Cho, M. S. Lee, E. S. Tak, E. Lee, K. S. Koh, C. H. Ahn, and S. C. Park. Gene expression profile in the anterior regeneration of the earthworm using expressed sequence tags. *Bioscience, Biotechnology, and Biochemistry*, 73(1):29–34, Jan. 2009. PMID: 19129665.

[27] M. S. Cline, M. Smoot, E. Cerami, A. Kuchinsky, N. Landys, C. Workman, R. Christmas, I. Avila-Campilo, M. Creech, B. Gross, K. Hanspers, R. Isserlin, R. Kelley, S. Killcoyne, S. Lotia, S. Maere, J. Morris, K. Ono, V. Pavlovic, A. R. Pico, A. Vailaya, P. Wang, A. Adler, B. R. Conklin, L. Hood, M. Kuiper, C. Sander, I. Schmulevich, B. Schwikowski, G. J. Warner, T. Ideker, and G. D. Bader. Integration of biological networks and gene expression data using cytoscape. *Nature Protocols*, 2(10):2366–2382, 2007. PMID: 17947979.

[28] P. J. A. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice. The sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38(6):1767 –1771, Apr. 2010.

[29] P. E. C. Compeau, P. A. Pevzner, and G. Tesler. How to apply de bruijn graphs to genome assembly. *Nat Biotech*, 29(11):987–991, Nov. 2011.

[30] A. Conesa and S. Götz. Blast2GO: a comprehensive suite for functional analysis in plant genomics. *International Journal of Plant Genomics*, 2008, 2008. PMID: 18483572 PMCID: 2375974.

[31] A. Conesa, S. Götz, J. M. García-Gómez, J. Terol, M. Talón, and M. Robles. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics (Oxford, England)*, 21(18):3674–3676, Sept. 2005. PMID: 16081474.

[32] V. Curwen, E. Eyras, T. D. Andrews, L. Clarke, E. Mongin, S. M. J. Searle, and M. Clamp. The ensembl automatic gene annotation system. *Genome Research*, 14(5):942–950, May 2004. PMID: 15123590.

[33] C. Darwin. On the formation of mould. *Proc. Geol. Soc. Lond.*, 2:574–576, 1838.

[34] C. Darwin. *The Origin of Species*. Oxford University Press, 1859.

[35] C. Darwin. *The formation of vegetable mould through the action of worms, with some observations on their habits*. John Murray, 1881.

[36] J. Daub, P. P. Gardner, J. Tate, D. Ramsköld, M. Manske, W. G. Scott, Z. Weinberg, S. Griffiths-Jones, and A. Bateman. The RNA WikiProject: community annotation of RNA families. *RNA*, 14(12):2462–2464, Dec. 2008. PMID: 18945806 PMCID: 2590952.

[37] A. Davison and M. Blaxter. Ancient origin of glycosyl hydrolase family 9 cellulase genes. *Molecular Biology and Evolution*, 22(5):1273 –1284, May 2005.

[38] N. G. de Bruijn. A combinatorial problem. *Koninklijke Nederlandse Akademie v. Wetenschappen*, 49:758–764, 1946.

[39] A. S. Denes, G. Jékely, P. R. H. Steinmetz, F. Raible, H. Snyman, B. Prud'homme, D. E. K. Ferrier, G. Balavoine, and D. Arendt. Molecular architecture of annelid nerve cord supports common origin of nervous system centralization in bilateria. *Cell*, 129(2):277–288, Apr. 2007. PMID: 17448990.

[40] P. S. Depkat-Jakob, M. Hilgarth, M. A. Horn, and H. L. Drake. Effect of earthworm feeding guilds on ingested dissimilatory nitrate reducers and denitrifiers in the alimentary canal of the earthworm. *Applied and Environmental Microbiology*, 76(18):6205–6214, Sept. 2010. PMID: 20656855 PMCID: 2937516.

[41] C. Dieterich, S. W. Clifton, L. N. Schuster, A. Chinwalla, K. Delehaunty, I. Dinkelacker, L. Fulton, R. Fulton, J. Godfrey, P. Minx, M. Mitreva, W. Roeseler, H. Tian, H. Witte, S. Yang, R. K. Wilson, and R. J. Sommer. The pristionchus pacificus genome provides a unique perspective on nematode lifestyle and parasitism. *Nat Genet*, 40(10):1193–1198, Oct. 2008.

[42] D. A. Earl, K. Bradnam, J. St. John, A. Darling, D. Lin, J. Faas, H. O. K. Yu, B. Vince, D. R. Zerbino, M. Diekhans, N. Nguyen, P. Nuwantha, A. W. Sung, Z. Ning, M. Haimel, J. T. Simpson, N. A. Fronseca, İ. Birol, T. R. Docking, I. Y. Ho, D. S. Rokhsar, R. Chikhi, D. Lavenier, G. Chapuis, D. Naquin, N. Maillet, M. C. Schatz, D. R. Kelly, A. M. Phillippy, S. Koren, S. Yang, W. Wu, W. Chou, A. Srivas-

tava, T. I. Shaw, J. G. Ruby, P. Skewes-Cox, M. Betegon, M. T. Dimon, V. Solovyev, P. Kosarev, D. Vorobyev, R. Ramirez-Gonzalez, R. Leggett, D. MacLean, F. Xia, R. Luo, Z. L, Y. Xie, B. Liu, S. Gnerre, I. MacCallum, D. Przybylski, F. J. Ribeiro, S. Yin, T. Sharpe, G. Hall, P. J. Kersey, R. Durbin, S. D. Jackman, J. A. Chapman, X. Huang, J. L. DeRisi, M. Caccamo, Y. Li, D. B. Jaffe, R. Green, D. Haussler, I. Korf, and B. Paten. Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Research*, 2011.

[43] E. Eden, R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, 10:48, Feb. 2009. PMID: 19192299 PMCID: 2644678.

[44] R. C. Edgar. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460 –2461, Oct. 2010.

[45] R. C. Edgar and E. W. Myers. PILER: identification and classification of genomic repeats. *Bioinformatics*, 21(Suppl 1):i152–i158, June 2005.

[46] C. A. Edwards. *Earthworm ecology*. CRC Press, Mar. 2004.

[47] C. A. Edwards and P. J. Bohlen. *Biology and ecology of earthworms*. Springer, 1996.

[48] K. Eilbeck, S. E. Lewis, C. J. Mungall, M. Yandell, L. Stein, R. Durbin, and M. Ashburner. The sequence ontology: a tool for the unification of genome annotations. *Genome Biology*, 6(5):R44, 2005. PMID: 15892872.

[49] C. G. Elsik, A. J. Mackey, J. T. Reese, N. V. Milshina, D. S. Roos, and G. M. Weinstock. Creating a honey bee consensus gene set. *Genome Biology*, 8(1):R13, 2007. PMID: 17241472.

[50] C. C. Englbrecht, H. Schoof, and S. Böhm. Conservation, diversification and expansion of C2H2 zinc finger proteins in the arabidopsis thaliana genome. *BMC genomics*, 5(1):39, July 2004. PMID: 15236668.

[51] A. J. Enright, S. Van Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7):1575–1584, Apr. 2002. PMID: 11917018.

[52] Fischer and L. Koszorus. Sublethal effects, accumulation capacities and elimination rates of arsenic, mercury and selenium in the manure worm, eisenia foetida (Oligochaeta, lumbricidae). *Pedobiologia*, 36(3):172–178, 1992.

[53] R. Fleischmann, M. Adams, O. White, R. Clayton, E. Kirkness, A. Kerlavage, C. Bult, J. Tomb, B. Dougherty, J. Merrick, and e. al. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science*, 269(5223):496 –512, July 1995.

[54] T. Flutre, E. Duprat, C. Feuillet, and H. Quesneville. Considering transposable element diversification in de novo annotation approaches. *PLoS ONE*, 6(1):e16526, Jan. 2011.

[55] P. P. Gardner, J. Daub, J. G. Tate, E. P. Nawrocki, D. L. Kolbe, S. Lindgreen, A. C. Wilkinson, R. D. Finn, S. Griffiths-Jones, S. R. Eddy, and A. Bateman. Rfam: updates to the RNA families database. *Nucleic Acids Research*, 37(Database issue):D136–140, Jan. 2009. PMID: 18953034.

[56] V. Garg, R. Gupta, and P. Kaushik. Vermicomposting of solid textile mill sludge spiked with cow dung and horse dung: a pilot-scale study. *International Journal of Environment and Pollution*, 38(4):385 – 396, 2009.

[57] D. Gerlach, M. Wolf, T. Dandekar, T. Müller, A. Pokorny, and S. Rahmann. Deep metazoan phylogeny. *In Silico Biology*, 7(2):151–154, 2007. PMID: 17688440.

[58] E. Ghedin, S. Wang, D. Spiro, E. Caler, Q. Zhao, J. Crabtree, J. E. Allen, A. L. Delcher, D. B. Guiliano, D. Miranda-Saavedra, S. V. Angiuoli, T. Creasy, P. Amedeo, B. Haas, N. M. El-Sayed, J. R. Wortman, T. Feldblyum, L. Tallon, M. Schatz, M. Shumway, H. Koo, S. L. Salzberg, S. Schobel, M. Pertea, M. Pop,

O. White, G. J. Barton, C. K. S. Carlow, M. J. Crawford, J. Daub, M. W. Dimmic, C. F. Estes, J. M. Foster, M. Ganatra, W. F. Gregory, N. M. Johnson, J. Jin, R. Komuniecki, I. Korf, S. Kumar, S. Laney, B. Li, W. Li, T. H. Lindblom, S. Lustigman, D. Ma, C. V. Maina, D. M. A. Martin, J. P. McCarter, L. McReynolds, M. Mitreva, T. B. Nutman, J. Parkinson, J. M. Peregrin-Alvarez, C. Poole, Q. Ren, L. Saunders, A. E. Sluder, K. Smith, M. Stanke, T. R. Unnasch, J. Ware, A. D. Wei, G. Weil, D. J. Williams, Y. Zhang, S. A. Williams, C. Fraser-Liggett, B. Slatko, M. L. Blaxter, and A. L. Scott. Draft genome of the filarial nematode parasite brugia malayi. *Science*, 317(5845):1756–1760, Sept. 2007.

[59] T. C. Glenn. Field guide to nextgeneration DNA sequencers. *Molecular Ecology Resources*, 11(5):759–769, Sept. 2011.

[60] S. Gnerre, I. MacCallum, D. Przybylski, F. J. Ribeiro, J. N. Burton, B. J. Walker, T. Sharpe, G. Hall, T. P. Shea, S. Sykes, A. M. Berlin, D. Aird, M. Costello, R. Daza, L. Williams, R. Nicol, A. Gnirke, C. Nusbaum, E. S. Lander, and D. B. Jaffe. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences*, Dec. 2010.

[61] A. Goffeau, B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, and S. G. Oliver. Life with 6000 genes. *Science (New York, N.Y.)*, 274(5287):546, 563–567, Oct. 1996. PMID: 8849441.

[62] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotech*, 29(7):644–652, July 2011.

[63] J. Graf. Molecular requirements for the colonization of hirudo medicinalis by aeromonas veronii. In J. Overmann, editor, *Molecular Basis of Symbiosis*, volume 41, pages 291–303. Springer-Verlag, Berlin/Heidelberg.

[64] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. R. Eddy. Rfam: an RNA family database. *Nucleic Acids Research*, 31(1):439–441, Jan. 2003. PMID: 12520045.

[65] S. Griffiths-Jones, S. Moxon, M. Marshall, A. Khanna, S. R. Eddy, and A. Bateman. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research*, 33(Database issue):D121–124, Jan. 2005. PMID: 15608160.

[66] D. R. Gustafsson, D. A. Price, and C. Erséus. Genetic variation in the popular lab worm lumbriculus variegatus (Annelida: clitellata: Lumbriculidae) reveals cryptic speciation. *Molecular Phylogenetics and Evolution*, 51(2):182–189, May 2009.

[67] B. J. Haas, S. L. Salzberg, W. Zhu, M. Pertea, J. E. Allen, J. Orvis, O. White, C. R. Buell, and J. R. Wortman. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biology*, 9(1):R7, 2008.

[68] O. Harismendy, P. C. Ng, R. L. Strausberg, X. Wang, T. B. Stockwell, K. Y. Beeson, N. J. Schork, S. S. Murray, E. J. Topol, S. Levy, and K. A. Frazer. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology*, 10(3):R32, 2009. PMID: 19327155 PMCID: 2691003.

[69] D. Hole, A. Perkins, J. Wilson, I. Alexander, P. Grice, and A. Evans. Does organic farming benefit biodiversity? *Biological Conservation*, 122(1):113–130, Mar. 2005.

[70] C. Holt and M. Yandell. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, 12(1):491, Dec. 2011. PMID: 22192575.

[71] M. A. Horn, R. Mertel, M. Gehre, M. Kästner, and H. L. Drake. In vivo emission of dinitrogen by earthworms via denitrifying bacteria in the gut. *Applied and Environmental Microbiology*, 72(2):1013–1018, Feb. 2006. PMID: 16461643.

[72] X. Huang and A. Madan. CAP3: a DNA sequence assembly program. *Genome Research*, 9(9):868 –877, 1999.

[73] J. P. Huelsenbeck and F. Ronquist. MRBAYES: bayesian inference of phylogenetic trees. *Bioinformatics (Oxford, England)*, 17(8):754–755, Aug. 2001. PMID: 11524383.

[74] T. Hulsen, J. de Vlieg, and W. Alkema. BioVenn – a web application for the comparison and visualization of biological lists using area-proportional venn diagrams. *BMC Genomics*, 9:488, Oct. 2008. PMID: 18925949 PMCID: 2584113.

[75] S. S. Hung, J. Wasmuth, C. Sanford, and J. Parkinson. DETECT–a density estimation tool for enzyme ClassificaTion and its application to plasmodium falciparum. *Bioinformatics*, 26(14):1690–1698, July 2010.

[76] S. M. Huse, J. A. Huber, H. G. Morrison, M. L. Sogin, and D. Welch. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology*, 8(7):R143, 2007.

[77] U. Irmler. Changes in earthworm populations during conversion from conventional to organic farming. *Agriculture, Ecosystems & Environment*, 135(3):194–198, Jan. 2010.

[78] S. W. James, D. Porco, T. Decaëns, B. Richard, R. Rougerie, and C. Erséus. DNA barcoding reveals cryptic diversity in lumbricus terrestris l., 1758 (Clitellata): resurrection of l. herculeus (Savigny, 1826). *PLoS ONE*, 5(12):e15629, Dec. 2010.

[79] J. Jurka. Repeats in genomic DNA: mining and meaning. *Current Opinion in Structural Biology*, 8(3):333–337, June 1998. PMID: 9666329.

[80] J. Jurka, P. Klonowski, V. Dagman, and P. Pelton. CENSOR–a program for identification and elimination of repetitive elements from DNA sequences. *Computers & Chemistry*, 20(1):119–121, Mar. 1996. PMID: 8867843.

[81] W. Kao, A. H. Chan, and Y. S. Song. ECHO: a reference-free short-read error correction algorithm. *Genome Research*, 21(7):1181–1192, July 2011. PMID: 21482625.

[82] G. R. Karsten and H. L. Drake. Denitrifying bacteria in the earthworm gastrointestinal tract and in vivo emission of nitrous oxide (N(inf2)O) by earthworms. *Applied and Environmental Microbiology*, 63(5):1878–1882, May 1997. PMID: 16535603.

[83] W. J. Kent. BLAT–the BLAST-like alignment tool. *Genome Research*, 12(4):656–664, Apr. 2002. PMID: 11932250.

[84] A. Kersanté, F. MartinLaurent, G. Soulas, and F. Binet. Interactions of earthworms with atrazinedegrading bacteria in an agricultural soil. *FEMS Microbiology Ecology*, 57(2):192–205, Aug. 2006.

[85] R. A. King, A. L. Tibble, and W. O. C. Symondson. Opening a can of worms: unprecedented sympatric cryptic diversity within british lumbricid earthworms. *Molecular Ecology*, 17(21):4684–4698, Nov. 2008. PMID: 18992008.

[86] B. Knapp, S. Podmirseg, J. Seeber, E. Meyer, and H. Insam. Diet-related composition of the gut microbiota of lumbricus rubellus as revealed by a molecular fingerprinting technique and cloning. *Soil Biology and Biochemistry*, 41(11):2299–2307, Nov. 2009.

[87] I. Korf. Gene finding in novel genomes. *BMC Bioinformatics*, 5(1):59, 2004.

[88] C. J. Krebs, L. K. Larkins, S. M. Khan, and D. M. Robins. Expansion and diversification of KRAB zinc-finger genes within a cluster including regulator of sex-limitation 1 and 2. *Genomics*, 85(6):752–761, June 2005.

[89] C. J. Langdon, A. A. Meharg, J. Feldmann, T. Balgar, J. Charnock, M. Farquhar, T. G. Piearce, K. T. Semple, and J. Cotter-Howells. Arsenic-speciation in arsenate-resistant and non-resistant populations of the earthworm, lumbricus rubellus. *J. Environ. Monit.*, 4(4):603–608, May 2002.

[90] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, 2009. PMID: 19261174.

[91] E. Lerat. Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. *Heredity*, 104(6):520–533, June 2010.

[92] S. Lewis, S. Searle, N. Harris, M. Gibson, V. Iyer, J. Richter, C. Wiel, L. Bayraktaroglu, E. Birney, M. Crosby, J. Kaminker, B. Matthews, S. Prochnik, C. Smith, J. Tupy, G. Rubin, S. Misra, C. Mungall, and M. Clamp. Apollo: a sequence annotation editor. *Genome Biology*, 3(12), 2002.

[93] J. Li, M. M. Riehle, Y. Zhang, J. Xu, F. Oduol, S. M. Gomez, K. Eiglmeier, B. M. Ueberheide, J. Shabanowitz, D. F. Hunt, J. M. C. Ribeiro, and K. D. Vernick. Anopheles gambiae genome reannotation through synthesis of ab initio and comparative gene prediction algorithms. *Genome Biology*, 7(3):R24, 2006. PMID: 16569258.

[94] L. Li, C. J. Stoeckert, and D. S. Roos. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13(9):2178 –2189, 2003.

[95] R. Li, W. Fan, G. Tian, H. Zhu, L. He, J. Cai, Q. Huang, Q. Cai, B. Li, Y. Bai, Z. Zhang, Y. Zhang, W. Wang, J. Li, F. Wei, H. Li, M. Jian, J. Li, Z. Zhang, R. Nielsen, D. Li, W. Gu, Z. Yang, Z. Xuan, O. A. Ryder, F. C. Leung, Y. Zhou, J. Cao, X. Sun, Y. Fu, X. Fang, X. Guo, B. Wang, R. Hou, F. Shen, B. Mu, P. Ni, R. Lin, W. Qian, G. Wang, C. Yu, W. Nie, J. Wang, Z. Wu, H. Liang, J. Min, Q. Wu,

S. Cheng, J. Ruan, M. Wang, Z. Shi, M. Wen, B. Liu, X. Ren, H. Zheng, D. Dong, K. Cook, G. Shan, H. Zhang, C. Kosiol, X. Xie, Z. Lu, H. Zheng, Y. Li, C. C. Steiner, T. T. Lam, S. Lin, Q. Zhang, G. Li, J. Tian, T. Gong, H. Liu, D. Zhang, L. Fang, C. Ye, J. Zhang, W. Hu, A. Xu, Y. Ren, G. Zhang, M. W. Bruford, Q. Li, L. Ma, Y. Guo, N. An, Y. Hu, Y. Zheng, Y. Shi, Z. Li, Q. Liu, Y. Chen, J. Zhao, N. Qu, S. Zhao, F. Tian, X. Wang, H. Wang, L. Xu, X. Liu, T. Vinar, Y. Wang, T. Lam, S. Yiu, S. Liu, H. Zhang, D. Li, Y. Huang, X. Wang, G. Yang, Z. Jiang, J. Wang, N. Qin, L. Li, J. Li, L. Bolund, K. Kristiansen, G. K. Wong, M. Olson, X. Zhang, S. Li, H. Yang, J. Wang, and J. Wang. The sequence and de novo assembly of the giant panda genome. *Nature*, advance online publication, Dec. 2009.

[96] R. Li, H. Zhu, J. Ruan, W. Qian, X. Fang, Z. Shi, Y. Li, S. Li, G. Shan, K. Kristiansen, S. Li, H. Yang, J. Wang, and J. Wang. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, Dec. 2009.

[97] W. Li and A. Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658 –1659, July 2006.

[98] B. Linard, J. D. Thompson, O. Poch, and O. Lecompte. OrthoInspector: comprehensive orthology analysis and visual exploration. *BMC Bioinformatics*, 12:11, 2011. PMID: 21219603.

[99] C. Lloyd. *What on Earth Evolved?: 100 Species That Changed the World*. Bloomsbury Publishing PLC, Oct. 2009.

[100] M. B. Lund, S. K. Davidson, M. Holmstrup, S. James, K. U. Kjeldsen, D. A. Stahl, and A. Schramm. Diversity and host specificity of the verminephrobacter–earthworm symbiosis. *Environmental Microbiology*, 12(8):2142–2151, Aug. 2010.

[101] M. B. Lund, M. Holmstrup, B. A. Lomstein, C. Damgaard, and A. Schramm. Beneficial effect of verminephrobacter nephridial symbionts on the fitness of the

earthworm aporrectodea tuberculata. *Applied and Environmental Microbiology*, 76(14):4738 –4743, July 2010.

[102] P. Maeder, A. Fliessbach, D. Dubois, L. Gunst, P. Fried, and U. Niggli. Soil fertility and biodiversity in organic farming. *Science*, 296(5573):1694 –1697, May 2002.

[103] S. Maere, K. Heymans, and M. Kuiper. BiNGO: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics (Oxford, England)*, 21(16):3448–3449, Aug. 2005. PMID: 15972284.

[104] N. O. Mainoo, S. Barrington, J. K. Whalen, and L. Sampedro. Pilot-scale vermicomposting of pineapple wastes with earthworms native to accra, ghana. *Bioresource Technology*, 100(23):5872–5875, Dec. 2009.

[105] G. Marçais and C. Kingsford. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, Jan. 2011.

[106] F. Mariño and A. Morgan. The time-course of metal (Ca, cd, cu, pb, zn) accumulation from a contaminated soil by three populations of the earthworm, lumbricus rubellus. *Applied Soil Ecology*, 12(2):169–177, May 1999.

[107] C. Matthies, A. Griesshammer, M. Schmittroth, and H. L. Drake. Evidence for involvement of gut-associated denitrifying bacteria in emission of nitrous oxide (N(2)O) by earthworms obtained from garden and forest soils. *Applied and Environmental Microbiology*, 65(8):3599–3604, Aug. 1999. PMID: 10427055.

[108] T. Maximilian J. A single origin of the central nervous system? *Cell*, 129(2):237–239, Apr. 2007.

[109] A. A. Meharg, R. F. Shore, and K. I. o. T. E. Broadgate. *Edaphic factors affecting the toxicity and accumulation of arsenate in the earthworm Lumbricus terrestris*. June 1998.

[110] R. Melgar, E. Benitez, and R. Nogales. Bioconversion of wastes from olive oil industries by vermicomposting process using the epigeic earthworm eisenia andrei. *Journal of Environmental Science and Health, Part B*, 44(5):488–495, June 2009.

[111] I. Milne, M. Bayer, L. Cardle, P. Shaw, G. Stephen, F. Wright, and D. Marshall. Tablet—next generation sequence assembly visualization. *Bioinformatics*, 26(3):401 –402, Feb. 2010.

[112] J. Morgan and A. Morgan. Calcium-lead interactions involving earthworms. part 1: The effect of exogenous calcium on lead accumulation by earthworms under field and laboratory conditions. *Environmental Pollution*, 54(1):41–53, 1988.

[113] J. Morgan and A. Morgan. Calcium-lead interactions involving earthworms. part 2: The effect of accumulated lead on endogenous calcium in lumbricus rubellus. *Environmental Pollution*, 55(1):41–54, 1988.

[114] O. Morozova and M. A. Marra. Applications of next-generation sequencing technologies in functional genomics. *Genomics*, Aug. 2008. PMID: 18703132.

[115] A. Mortazavi, E. M. Schwarz, B. Williams, L. Schaeffer, I. Antoshechkin, B. J. Wold, and P. W. Sternberg. Scaffolding a caenorhabditis nematode genome with RNA-seq. *Genome Research*, 20(12):1740–1747, Dec. 2010. PMID: 20980554.

[116] I. M.P. Metal accumulation by the earthworms lumbricus rubellus, dendrobaena veneta and eiseniella tetraedra living in heavy metal polluted sites. *Environmental Pollution (1970)*, 19(3):201–206, July 1979.

[117] J. C. Mullikin and Z. Ning. The phusion assembler. *Genome Research*, 13(1):81–90, Jan. 2003. PMID: 12529309.

[118] C. J. Mungall, D. B. Emmert, and T. F. Consortium. A chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, 23(13):i337 –i346, July 2007.

[119] E. W. Myers, G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, C. M. Mobarry, K. H. J. Reinert, K. A. Remington, E. L. Anson, R. A. Bolanos, H. Chou, C. M. Jordan, A. L. Halpern, S. Lonardi, E. M. Beasley, R. C. Brandon, L. Chen, P. J. Dunn, Z. Lai, Y. Liang, D. R. Nusskern, M. Zhan, Q. Zhang, X. Zheng, G. M. Rubin, M. D. Adams, and J. C. Venter. A Whole-Genome assembly of drosophila. *Science*, 287(5461):2196–2204, Mar. 2000.

[120] G. Narzisi and B. Mishra. Comparing de novo genome assembly: the long and short of it. *PloS One*, 6(4):e19175, 2011. PMID: 21559467.

[121] V. Nene, J. R. Wortman, D. Lawson, B. Haas, C. Kodira, Z. J. Tu, B. Loftus, Z. Xi, K. Megy, M. Grabherr, Q. Ren, E. M. Zdobnov, N. F. Lobo, K. S. Campbell, S. E. Brown, M. F. Bonaldo, J. Zhu, S. P. Sinkins, D. G. Hogenkamp, P. Amedeo, P. Arensburger, P. W. Atkinson, S. Bidwell, J. Biedler, E. Birney, R. V. Bruggner, J. Costas, M. R. Coy, J. Crabtree, M. Crawford, B. deBruyn, D. DeCaprio, K. Eiglmeier, E. Eisenstadt, H. El-Dorry, W. M. Gelbart, S. L. Gomes, M. Hammond, L. I. Hannick, J. R. Hogan, M. H. Holmes, D. Jaffe, J. S. Johnston, R. C. Kennedy, H. Koo, S. Kravitz, E. V. Kriventseva, D. Kulp, K. LaButti, E. Lee, S. Li, D. D. Lovin, C. Mao, E. Mauceli, C. F. M. Menck, J. R. Miller, P. Montgomery, A. Mori, A. L. Nascimento, H. F. Naveira, C. Nusbaum, S. O'Leary, J. Orvis, M. Pertea, H. Quesneville, K. R. Reidenbach, Y. Rogers, C. W. Roth, J. R. Schneider, M. Schatz, M. Shumway, M. Stanke, E. O. Stinson, J. M. C. Tubio, J. P. VanZee, S. Verjovski-Almeida, D. Werner, O. White, S. Wyder, Q. Zeng, Q. Zhao, Y. Zhao, C. A. Hill, A. S. Raikhel, M. B. Soares, D. L. Knudson, N. H. Lee, J. Galagan, S. L. Salzberg, I. T. Paulsen, G. Dimopoulos, F. H. Collins, B. Birren, C. M. Fraser-Liggett, and D. W. Severson. Genome sequence of aedes aegypti, a major arbovirus vector. *Science*, 316(5832):1718 –1723, June 2007.

[122] M. Novo, A. Almodóvar, R. Fernández, D. Trigo, and D. J. Díaz Cosín. Cryptic speciation of hormogastrid earthworms revealed by mitochondrial and nuclear data. *Molecular Phylogenetics and Evolution*, 56(1):507–512, July 2010.

[123] E. Olchawa, M. Bzowska, S. R. Stürzenbaum, A. J. Morgan, and B. Plytycz. Heavy metals affect the coelomocyte-bacteria balance in earthworms: environmental interactions between abiotic and biotic stressors. *Environmental Pollution (Barking, Essex: 1987)*, 142(2):373–81, July 2006. PMID: 16309804.

[124] S. Ootsuka, N. Saga, K.-i. Suzuki, A. Inoue, and T. Ojima. Isolation and cloning of an endo-beta-1,4-mannanase from pacific abalone haliotis discus hannai. *Journal of Biotechnology*, 125(2):269–280, Sept. 2006. PMID: 16621092.

[125] G. Ostlund, T. Schmitt, K. Forslund, T. Köstler, D. N. Messina, S. Roopra, O. Frings, and E. L. L. Sonnhammer. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Research*, 38(Database issue):D196–203, Jan. 2010. PMID: 19892828.

[126] J. Owen, B. A. Hedley, C. Svendsen, J. Wren, M. J. Jonker, P. K. Hankard, L. J. Lister, S. R. Stürzenbaum, A. J. Morgan, D. J. Spurgeon, M. L. Blaxter, and P. Kille. Transcriptome profiling of developmental and xenobiotic responses in a keystone soil animal, the oligochaete annelid lumbricus rubellus. *BMC Genomics*, 9:266, 2008. PMC2440553.

[127] A. Pain, U. Bohme, A. E. Berry, K. Mungall, R. D. Finn, A. P. Jackson, T. Mourier, J. Mistry, E. M. Pasini, M. A. Aslett, S. Balasubrammaniam, K. Borgwardt, K. Brooks, C. Carret, T. J. Carver, I. Cherevach, T. Chillingworth, T. G. Clark, M. R. Galinski, N. Hall, D. Harper, D. Harris, H. Hauser, A. Ivens, C. S. Janssen, T. Keane, N. Larke, S. Lapp, M. Marti, S. Moule, I. M. Meyer, D. Ormond, N. Peters, M. Sanders, S. Sanders, T. J. Sargeant, M. Simmonds, F. Smith, R. Squares, S. Thurston, A. R. Tivey, D. Walker, B. White, E. Zuiderwijk, C. Churcher, M. A. Quail, A. F. Cowman, C. M. R. Turner, M. A. Rajandream, C. H. M. Kocken, A. W. Thomas, C. I. Newbold, B. G. Barrell, and M. Berriman. The genome of the simian and human malaria parasite plasmodium knowlesi. *Nature*, 455(7214):799–803, Oct. 2008.

[128] J. Parkinson, A. Anthony, J. Wasmuth, R. Schmid, A. Hedley, and M. Blaxter. PartiGene—constructing partial genomes. *Bioinformatics*, 20(9):1398 –1404, June 2004.

[129] 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.

[130] P. A. Pevzner, H. Tang, and M. S. Waterman. An eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America*, 98(17):9748–9753, Aug. 2001. PMID: 11504945.

[131] E. Picardi and G. Pesole. Computational methods for ab initio and comparative gene finding. *Methods in Molecular Biology (Clifton, N.J.)*, 609:269–284, 2010. PMID: 20221925.

[132] A. L. Price, N. C. Jones, and P. A. Pevzner. De novo identification of repeat families in large genomes. *Bioinformatics*, 21(suppl_1):i351–358, June 2005.

[133] E. Quevillon, V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler, and R. Lopez. InterProScan: protein domains identifier. *Nucleic Acids Research*, 33(Web Server):W116–W120, July 2005.

[134] O. Riba-Grognuz, L. Keller, L. Falquet, I. Xenarios, and Y. Wurm. Visualization and quality assessment of de novo genome assemblies. *Bioinformatics*, 27(24):3425 –3426, Dec. 2011.

[135] S. Richards, R. A. Gibbs, G. M. Weinstock, S. J. Brown, R. Denell, R. W. Beeman, R. Gibbs, R. W. Beeman, S. J. Brown, G. Bucher, M. Friedrich, C. J. P. Grimmelikhuijzen, M. Klingler, M. Lorenzen, S. Richards, S. Roth, R. Schröder, D. Tautz, E. M. Zdobnov, D. Muzny, R. A. Gibbs, G. M. Weinstock, T. Attaway, S. Bell, C. J. Buhay, M. N. Chandrabose, D. Chavez, K. P. Clerk-Blankenburg, A. Cree, M. Dao, C. Davis, J. Chacko, H. Dinh, S. Dugan-Rocha, G. Fowler, T. T. Garner, J. Garnes, A. Gnirke, A. Hawes, J. Hernandez, S. Hines, M. Holder, J. Hume, S. N. Jhangiani,

V. Joshi, Z. M. Khan, L. Jackson, C. Kovar, A. Kowis, S. Lee, L. R. Lewis, J. Margolis, M. Morgan, L. V. Nazareth, N. Nguyen, G. Okwuonu, D. Parker, S. Richards, S.-J. Ruiz, J. Santibanez, J. Savard, S. E. Scherer, B. Schneider, E. Sodergren, D. Tautz, S. Vattahil, D. Villasana, C. S. White, R. Wright, Y. Park, R. W. Beeman, J. Lord, B. Oppert, M. Lorenzen, S. Brown, L. Wang, J. Savard, D. Tautz, S. Richards, G. Weinstock, R. A. Gibbs, Y. Liu, K. Worley, G. Weinstock, C. G. Elsik, J. T. Reese, E. Elhaik, G. Landan, D. Graur, P. Arensburger, P. Atkinson, R. W. Beeman, J. Beidler, S. J. Brown, J. P. Demuth, D. W. Drury, Y.-Z. Du, H. Fujiwara, M. Lorenzen, V. Maselli, M. Osanai, Y. Park, H. M. Robertson, Z. Tu, J.-j. Wang, S. Wang, S. Richards, H. Song, L. Zhang, E. Sodergren, D. Werner, M. Stanke, B. Morgenstern, V. Solovyev, P. Kosarev, G. Brown, H.-C. Chen, O. Ermolaeva, W. Hlavina, Y. Kapustin, B. Kiryutin, P. Kitts, D. Maglott, K. Pruitt, V. Sapojnikov, A. Souvorov, A. J. Mackey, R. M. Waterhouse, S. Wyder, E. M. Zdobnov, E. M. Zdobnov, S. Wyder, E. V. Kriventseva, T. Kadowaki, P. Bork, M. Aranda, R. Bao, A. Beermann, N. Berns, R. Bolognesi, F. Bonneton, D. Bopp, S. J. Brown, G. Bucher, T. Butts, A. Chaumot, R. E. Denell, D. E. K. Ferrier, M. Friedrich, C. M. Gordon, M. Jindra, M. Klingler, Q. Lan, H. M. G. Lattorff, V. Laudet, C. von Levetsow, Z. Liu, R. Lutz, J. A. Lynch, R. N. da Fonseca, N. Posnien, R. Reuter, S. Roth, J. Savard, J. B. Schinko, C. Schmitt, M. Schoppmeier, R. Schröder, T. D. Shippy, F. Simonnet, H. Marques-Souza, D. Tautz, Y. Tomoyasu, J. Trauner, M. Van der Zee, M. Vervoort, N. Wittkopp, E. A. Wimmer, X. Yang, A. K. Jones, D. B. Sattelle, P. R. Ebert, D. Nelson, J. G. Scott, R. W. Beeman, S. Muthukrishnan, K. J. Kramer, Y. Arakane, R. W. Beeman, Q. Zhu, D. Hogenkamp, R. Dixit, B. Oppert, H. Jiang, Z. Zou, J. Marshall, E. Elpidina, K. Vinokurov, C. Oppert, Z. Zou, J. Evans, Z. Lu, P. Zhao, N. Sumathipala, B. Altincicek, A. Vilcinskas, M. Williams, D. Hultmark, C. Hetru, H. Jiang, C. J. P. Grimmelikhuijzen, F. Hauser, G. Cazzamali, M. Williamson, Y. Park, B. Li, Y. Tanaka, R. Predel, S. Neupert, J. Schachtner, P. Verleyen, F. Raible, P. Bork, M. Friedrich, K. K. O. Walden, H. M. Robertson, S. Angeli, S. Forêt, G. Bucher, S. Schuetz, R. Maleszka, E. A. Wimmer,

R. W. Beeman, M. Lorenzen, Y. Tomoyasu, S. C. Miller, D. Grossmann, and G. Bucher. The genome of the model beetle and pest tribolium castaneum. *Nature*, 452(7190):949–955, Apr. 2008. PMID: 18362917.

[136] G. Robertson, J. Schein, R. Chiu, R. Corbett, M. Field, S. D. Jackman, K. Mungall, S. Lee, H. M. Okada, J. Q. Qian, M. Griffith, A. Raymond, N. Thiessen, T. Cezard, Y. S. Butterfield, R. Newsome, S. K. Chan, R. She, R. Varhol, B. Kamoh, A. Prabhu, A. Tam, Y. Zhao, R. A. Moore, M. Hirst, M. A. Marra, S. J. M. Jones, P. A. Hoodless, and I. Birol. De novo assembly and analysis of RNA-seq data. *Nat Meth*, 7(11):909–912, Nov. 2010.

[137] K. Rutherford, J. Parkhill, J. Crook, T. Horsnell, P. Rice, M. Rajandream, and B. Barrell. Artemis: sequence visualization and annotation. *Bioinformatics*, 16:944–945, Oct. 2000.

[138] F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, C. A. Fiddes, C. A. Hutchison, P. M. Slocombe, and M. Smith. Nucleotide sequence of bacteriophage phi x174 DNA. *Nature*, 265(5596):687–695, Feb. 1977. PMID: 870828.

[139] P. Sangwan, C. Kaushik, and V. Garg. Vermicomposting of sugar industry waste (press mud) mixed with cow dung employing an epigeic earthworm eisenia fetida. *Waste Management & Research*, 28(1):71 –75, Jan. 2010.

[140] R. Schmid and M. L. Blaxter. annot8r: GO, EC and KEGG annotation of EST datasets. *BMC Bioinformatics*, 9:180, 2008. PMID: 18400082.

[141] A. Schramm, S. K. Davidson, J. A. Dodsworth, H. L. Drake, D. A. Stahl, and N. Dubilier. Acidovorax-like symbionts in the nephridia of earthworms. *Environmental Microbiology*, 5(9):804–809, Sept. 2003. PMID: 12919416.

[142] J. Schröder, H. Schröder, S. J. Puglisi, R. Sinha, and B. Schmidt. SHREC: a short-read error correction method. *Bioinformatics (Oxford, England)*, 25(17):2157–2163, Sept. 2009. PMID: 19542152.

[143] M. H. Schulz, D. R. Zerbino, M. Vingron, and E. Birney. Oases: Robust de novo RNA-Seq assembly across the dynamic range of expression levels. *Bioinformatics*, 28(8):1086–1092, Apr. 2012.

[144] S. Schwartz, J. Silva, D. Burstein, T. Pupko, E. Eyras, and G. Ast. Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. *Genome Research*, 18(1):88 –103, Jan. 2008.

[145] O. Simakov, F. Marletaz, S.-J. Cho, E. Edsinger-Gonzales, P. Havlak, U. Hellsten, D.-H. Kuo, T. Larsson, J. Lv, D. Arendt, R. Savage, K. Osoegawa, P. d. Jong, J. Grimwood, J. A. Chapman, H. Shapiro, A. Aerts, R. P. Otillar, A. Y. Terry, J. L. Boore, I. V. Grigoriev, D. R. Lindberg, E. C. Seaver, D. A. Weisblat, N. H. Putnam, and D. S. Rokhsar. Insights into bilaterian evolution from three spiralian genomes. *Nature*, 2012.

[146] J. T. Simpson and R. Durbin. Efficient de novo assembly of large genomes using compressed data structures. *Genome Research*, Dec. 2011.

[147] 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. PMID: 19251739.

[148] R. W. Sims, B. M. Gerard, L. S. of London., Estuarine, C. S. Association., and F. S. C. G. Britain). *Earthworms : notes for the identification of British species*. Published for the Linnean Society of London and the Estuarine and Coastal Sciences Association by Field Studies Council, Shrewsbury, 1999.

[149] T. Sizmur, M. J. Watts, G. D. Brown, B. Palumbo-Roe, and M. E. Hodson. Impact of gut passage and mucus secretion by the earthworm lumbricus terrestris on mobility and speciation of arsenic in contaminated soil. *Journal of Hazardous Materials*, 197(0):169–175, Dec. 2011.

[150] G. S. C. Slater and E. Birney. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6:31, 2005. PMID: 15713233.

[151] H. R. Smit, AFA. Repeatmodeler open-1.0, 2008-2010.

[152] H. R. Smit, AFA and P. Green. Repeatmasker open-3.0, 1996-2010.

[153] E. A. Sperling, J. Vinther, V. N. Moy, B. M. Wheeler, M. Sémon, D. E. G. Briggs, and K. J. Peterson. MicroRNAs resolve an apparent conflict between annelid systematics and their fossil record. *Proceedings. Biological Sciences / The Royal Society*, Sept. 2009. PMID: 19755470.

[154] D. J. Spurgeon, C. Svendsen, L. J. Lister, P. K. Hankard, and P. Kille. Earthworm responses to cd and cu under fluctuating environmental conditions: a comparison with results from laboratory exposures. *Environmental Pollution (Barking, Essex: 1987)*, 136(3):443–52, Aug. 2005. PMID: 15862398.

[155] D. J. Spurgeon, C. Svendsen, J. M. Weeks, P. K. Hankard, H. E. Stubberud, and J. E. Kammenga. Quantifying copper and cadmium impacts on intrinsic rate of population increase in the terrestrial oligochaete lumbricus rubellus. *Environmental Toxicology and Chemistry / SETAC*, 22(7):1465–1472, July 2003. PMID: 12836970.

[156] J. E. Stajich, D. Block, K. Boulez, S. E. Brenner, S. A. Chervitz, C. Dagdigian, G. Fuellen, J. G. R. Gilbert, I. Korf, H. Lapp, H. Lehväslaiho, C. Matsalla, C. J. Mungall, B. I. Osborne, M. R. Pocock, P. Schattner, M. Senger, L. D. Stein, E. Stupka, M. D. Wilkinson, and E. Birney. The bioperl toolkit: Perl modules for the life sciences. *Genome Research*, 12(10):1611–1618, Oct. 2002. PMID: 12368254.

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

[158] S. R. Starkenburg, F. W. Larimer, L. Y. Stein, M. G. Klotz, P. S. G. Chain, L. A. Sayavedra-Soto, A. T. Poret-Peterson, M. E. Gentry, D. J. Arp, B. Ward, and P. J.

Bottomley. Complete genome sequence of nitrobacter hamburgensis x14 and comparative genomic analysis of species within the genus nitrobacter. *Applied and Environmental Microbiology*, 74(9):2852 –2863, May 2008.

[159] L. D. Stein, C. Mungall, S. Shu, M. Caudy, M. Mangone, A. Day, E. Nickerson, J. E. Stajich, T. W. Harris, A. Arva, and S. Lewis. The generic genome browser: a building block for a model organism system database. *Genome Research*, 12(10):1599–1610, Oct. 2002. PMID: 12368253.

[160] T. H. Struck, C. Paul, N. Hill, S. Hartmann, C. Hosel, M. Kube, B. Lieb, A. Meyer, R. Tiedemann, G. Purschke, and C. Bleidorn. Phylogenomic analyses unravel annelid evolution. *Nature*, 471(7336):95–98, Mar. 2011.

[161] S. R. Stürzenbaum, J. Andre, P. Kille, and A. J. Morgan. Earthworm genomes, genes and proteins: the (re)discovery of darwin's worms. *Proceedings. Biological Sciences / The Royal Society*, 276(1658):789–97, Mar. 2009. PMID: 19129111.

[162] S. R. Stürzenbaum, O. Georgiev, A. J. Morgan, and P. Kille. Cadmium detoxification in earthworms: from genes to cells. *Environmental Science & Technology*, 38(23):6283–9, Dec. 2004. PMID: 15597883.

[163] G. Suen, C. Teiling, L. Li, C. Holt, E. Abouheif, E. Bornberg-Bauer, P. Bouffard, E. J. Caldera, E. Cash, A. Cavanaugh, O. Denas, E. Elhaik, M. Favé, J. Gadau, J. D. Gibson, D. Graur, K. J. Grubbs, D. E. Hagen, T. T. Harkins, M. Helmkampf, H. Hu, B. R. Johnson, J. Kim, S. E. Marsh, J. A. Moeller, M. C. Muñoz-Torres, M. C. Murphy, M. C. Naughton, S. Nigam, R. Overson, R. Rajakumar, J. T. Reese, J. J. Scott, C. R. Smith, S. Tao, N. D. Tsutsui, L. Viljakainen, L. Wissler, M. D. Yandell, F. Zimmer, J. Taylor, S. C. Slater, S. W. Clifton, W. C. Warren, C. G. Elsik, C. D. Smith, G. M. Weinstock, N. M. Gerardo, and C. R. Currie. The genome sequence of the Leaf-Cutter ant atta cephalotes reveals insights into its obligate symbiotic lifestyle. *PLoS Genet*, 7(2):e1002007, Feb. 2011.

[164] M. Sugimoto and N. Nakajima. Molecular cloning, sequencing, and expression of cDNA encoding serine protease with fibrinolytic activity from earthworm. *Bioscience, Biotechnology, and Biochemistry*, 65(7):1575–1580, July 2001. PMID: 11515541.

[165] H. D. Tadepally, G. Burger, and M. Aubry. Evolution of C2H2-zinc finger genes and subfamilies in mammals: Species-specific duplication and loss of clusters, genes and effector domains. *BMC Evolutionary Biology*, 8(1):176, June 2008.

[166] M. Tahtouh, F. Croq, J. Vizioli, P. Sautiere, C. Van Camp, M. Salzet, M. R. Daha, J. Pestel, and C. Lefebvre. Evidence for a novel chemotactic c1q domain-containing factor in the leech nerve cord. *Molecular Immunology*, 46(4):523–531, Feb. 2009. PMID: 18952286.

[167] J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, 22(22):4673–4680, Nov. 1994. PMID: 7984417.

[168] G. T.R. and H. P.D.N. Genome size estimates for some oligochaete annelids. *Canadian Journal of Zoology*, 80(8):1485–1489, 2002.

[169] J. Trisina, F. Sunardi, M. T. Suhartono, and R. R. Tjandrawinata. DLBS1033, a protein extract from lumbricus rubellus, possesses antithrombotic and thrombolytic activities. 2011, 2011. PMID: 21403877 PMCID: 3051164.

[170] I. J. Tsai, T. D. Otto, and M. Berriman. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biology*, 11(4):R41, 2010. PMID: 20388197.

[171] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen,

M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. D. Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Y. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. C. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigo, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. Chiang, M. Coyne, C. Dahlke, A. D. Mays, M. Dombroski,

M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. The sequence of the human genome. *Science*, 291(5507):1304–1351, Feb. 2001.

[172] E. M. Volkov, L. F. Nurullin, I. Svandová, E. E. Nikolsky, and F. Vyskocil. Participation of electrogenic Na+-K+-ATPase in the membrane potential of earthworm body wall muscles. *Physiological Research / Academia Scientiarum Bohemoslovaca*, 49(4):481–484, 2000. PMID: 11072810.

[173] C. Wang, Z. Sun, Y. Liu, X. Zhang, and G. Xu. A novel antimicrobial vermipeptide family from earthworm eisenia fetida. *European Journal of Soil Biology*, 43(Supplement 1):S127–S134, Nov. 2007.

[174] H. Watanabe and G. Tokuda. Cellulolytic systems in insects. *Annual Review of Entomology*, 55:609–632, 2010. PMID: 19754245.

[175] D. A. Wheeler, M. Srinivasan, M. Egholm, Y. Shen, L. Chen, A. McGuire, W. He, Y. Chen, V. Makhijani, G. T. Roth, X. Gomes, K. Tartaro, F. Niazi, C. L. Turcotte, G. P. Irzyk, J. R. Lupski, C. Chinault, X.-z. Song, Y. Liu, Y. Yuan, L. Nazareth, X. Qin, D. M. Muzny, M. Margulies, G. M. Weinstock, R. A. Gibbs, and J. M. Rothberg. The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452(7189):872–876, Apr. 2008.

[176] T. Woyke, H. Teeling, N. N. Ivanova, M. Huntemann, M. Richter, F. O. Gloeckner, D. Boffelli, I. J. Anderson, K. W. Barry, H. J. Shapiro, E. Szeto, N. C. Kyrpides, M. Mussmann, R. Amann, C. Bergin, C. Ruehland, E. M. Rubin, and N. Dubilier.

Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature*, 443(7114):950–955, Oct. 2006.

[177] P. K. Wüst, M. A. Horn, G. Henderson, P. H. Janssen, B. H. A. Rehm, and H. L. Drake. Gut-associated denitrification and in vivo emission of nitrous oxide by the earthworm families megascolecidae and lumbricidae in new zealand. *Applied and Environmental Microbiology*, 75(11):3430–3436, June 2009. PMID: 19346358.

[178] Y. Xu, X. Wang, J. Yang, J. Vaynberg, and J. Qin. PASA–a program for automated protein NMR backbone signal assignment by pattern-filtering approach. *Journal of biomolecular NMR*, 34(1):41–56, Jan. 2006. PMID: 16505963.

[179] X. Yang, S. P. Chockalingam, and S. Aluru. A survey of error-correction methods for next-generation sequencing. *Briefings in Bioinformatics*, Apr. 2012.

[180] X. Yang, K. S. Dorman, and S. Aluru. Reptile: representative tiling for short read error correction. *Bioinformatics*, 26(20):2526 –2533, Oct. 2010.

[181] G. W. Yeates, V. A. Orchard, T. W. Speir, J. L. Hunt, and M. C. C. Hermans. Impact of pasture contamination by copper, chromium, arsenic timber preservative on soil biological activity. *Biology and Fertility of Soils*, 18(3):200–208, 1994.

[182] U. A. Zahura, M. M. Rahman, A. Inoue, H. Tanaka, and T. Ojima. An endo-beta-1,4-mannanase, AkMan, from the common sea hare aplysia kurodai. *Comparative Biochemistry and Physiology. Part B, Biochemistry & Molecular Biology*, 157(1):137–143, Sept. 2010. PMID: 20639136.

[183] B. R. Zeeberg, W. Feng, G. Wang, M. D. Wang, A. T. Fojo, M. Sunshine, S. Narasimhan, D. W. Kane, W. C. Reinhold, S. Lababidi, K. J. Bussey, J. Riss, J. C. Barrett, and J. N. Weinstein. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biology*, 4(4):R28, 2003. PMID: 12702209.

[184] D. Zerbino and E. Birney. Velvet: Algorithms for de novo short read assembly using de bruijn graphs. *Genome Res.*, page gr.074492.107, Mar. 2008.

[185] D. Zerbino, G. McEwen, E. Margulies, and E. Birney. Pebble and rock band: Heuristic resolution of repeats and scaffolding in the velvet Short-Read de novo assembler. *PLoS ONE*, 4(12):e8407, Dec. 2009.

[186] W. Zhang, J. Chen, Y. Yang, Y. Tang, J. Shang, and B. Shen. A practical comparison of de novo genome assembly software tools for Next-Generation sequencing technologies. *PLoS ONE*, 6(3):e17915, Mar. 2011.

[187] Y. Zhou, H. Zheng, Y. Chen, L. Zhang, K. Wang, J. Guo, Z. Huang, B. Zhang, W. Huang, K. Jin, T. Dou, M. Hasegawa, L. Wang, Y. Zhang, J. Zhou, L. Tao, Z. Cao, Y. Li, T. Vinar, B. Brejova, D. Brown, M. Li, D. J. Miller, D. Blair, Y. Zhong, Z. Chen, F. Liu, W. Hu, Z.-Q. Wang, Q.-H. Zhang, H.-D. Song, S. Chen, X. Xu, B. Xu, C. Ju, Y. Huang, P. J. Brindley, D. P. McManus, Z. Feng, Z.-G. Han, G. Lu, S. Ren, Y. Wang, W. Gu, H. Kang, J. Chen, X. Chen, S. Chen, L. Wang, J. Yan, B. Wang, X. Lv, L. Jin, B. Wang, S. Pu, X. Zhang, W. Zhang, Q. Hu, G. Zhu, J. Wang, J. Yu, J. Wang, H. Yang, Z. Ning, M. Beriman, C.-L. Wei, Y. Ruan, G. Zhao, S. Wang, F. Liu, Y. Zhou, Z.-Q. Wang, G. Lu, H. Zheng, P. J. Brindley, D. P. McManus, D. Blair, Q.-h. Zhang, Y. Zhong, S. Wang, Z.-G. Han, Z. Chen, S. Wang, Z.-G. Han, Z. Chen, Y. Zhou, H. Zheng, Y. Chen, L. Zhang, K. Wang, J. Guo, Z. Huang, B. Zhang, W. Huang, K. Jin, T. Dou, M. Hasegawa, L. Wang, Y. Zhang, J. Zhou, L. Tao, Z. Cao, Y. Li, T. Vinar, B. Brejova, D. Brown, M. Li, D. J. Miller, D. Blair, Y. Zhong, Z. Chen, Y. Zhou, H. Zheng, Y. Chen, L. Zhang, K. Wang, J. Guo, Z. Huang, B. Zhang, W. Huang, K. Jin, T. Dou, M. Hasegawa, L. Wang, Y. Zhang, J. Zhou, L. Tao, Z. Cao, Y. Li, T. Vinar, B. Brejova, D. Brown, M. Li, D. J. Miller, D. Blair, Y. Zhong, Z. Chen, F. Liu, W. Hu, Z.-Q. Wang, Q.-H. Zhang, H.-D. Song, S. Chen, X. Xu, B. Xu, C. Ju, Y. Huang, P. J. Brindley, D. P. McManus, Z. Feng, Z.-G. Han, F. Liu, W. Hu, Z.-Q. Wang, Q.-H. Zhang, H.-D. Song, S. Chen, X. Xu, B. Xu, C. Ju, Y. Huang, P. J. Brindley, D. P. McManus,

Z. Feng, Z.-G. Han, G. Lu, S. Ren, Y. Wang, W. Gu, H. Kang, J. Chen, X. Chen, S. Chen, L. Wang, J. Yan, B. Wang, X. Lv, L. Jin, B. Wang, S. Pu, X. Zhang, W. Zhang, Q. Hu, G. Zhu, J. Wang, J. Yu, J. Wang, H. Yang, Z. Ning, M. Beriman, C.-L. Wei, Y. Ruan, G. Zhao, S. Wang, G. Lu, S. Ren, Y. Wang, W. Gu, H. Kang, J. Chen, X. Chen, S. Chen, L. Wang, J. Yan, B. Wang, X. Lv, L. Jin, B. Wang, S. Pu, X. Zhang, W. Zhang, Q. Hu, G. Zhu, J. Wang, J. Yu, J. Wang, H. Yang, Z. Ning, M. Beriman, C.-L. Wei, Y. Ruan, G. Zhao, S. Wang, F. Liu, Y. Zhou, Z.-Q. Wang, G. Lu, H. Zheng, P. J. Brindley, D. P. McManus, D. Blair, Q.-h. Zhang, Y. Zhong, S. Wang, Z.-G. Han, Z. Chen, F. Liu, Y. Zhou, Z.-Q. Wang, G. Lu, H. Zheng, P. J. Brindley, D. P. McManus, D. Blair, Q.-h. Zhang, Y. Zhong, S. Wang, Z.-G. Han, Z. Chen, S. Wang, Z.-G. Han, Z. Chen, S. Wang, Z.-G. Han, and Z. Chen. The schistosoma japonicum genome reveals features of host–parasite interplay. *Nature*, 460(7253):345–351, July 2009.

[188] A. V. Zimin, D. R. Smith, G. Sutton, and J. A. Yorke. Assembly reconciliation. *Bioinformatics*, 24(1):42–45, Jan. 2008.