ABSTRACT

Title of Document:	EVOLUTION OF SEX-BIASED EXPRESSION IN CAENORHABDITIS
	Cristel G. Thomas, Doctor of Philosophy, 2011
Directed By:	Associate Professor Eric S. Haag, Department of Biology

Mating systems have a profound impact on genome structure evolution, both indirectly through their effects on population genetics and directly due to the genetic control of reproductive traits. Most extant Caenorhabditis species are gonochoristic (males and females), while the most studied species, *C. elegans* and *C. briggsae*, are androdioecious (self-fertile hermaphrodites and males). The latter two species display an overall reduced ability to mate, suggesting that the selective pressure on maintaining efficient mating was weakened as selfing arose. The genes underlying these traits were likely to have been expressed in a sex-biased fashion in the gonochoristic ancestor, and we hypothesized that as selfing emerged their regulation was modified or they were lost altogether. This hypothesis is especially interesting given that selfing species have consistently smaller genome sizes than their gonochoristic relatives. I sought to address whether a disproportionate loss of genes with sex-biased expression accompanies the loss of mating-related traits in Caenorhabditis hermaphrodites. I first examine sex-biased expression in a gonochoristic species, *C. remanei*, and identify genes with highly sex-biased expression. I find that these genes are more likely to be missing in selfing species than expected by chance. I then select some of these genes based on their phylogenetic conservation patterns in the genus, and characterize them more thoroughly to shed some light on their functions. Through this study I identify a novel male-associated candidate cis-regulatory element. Lastly, I broaden the scope of the study by determining transcriptome wide sex-biased expression patterns in four Caenorhabditis species. I confirm that *C.elegans* displays a decrease in the proportion of strong female-biased expression, as well as a modification of the expression of genes with male-biased expression both in males and in hermaphrodites, when compared to gonochoristic Caenorhabditis. Taken together, this study illustrates the transcriptomic consequences of a modification of the mating system, and begins to address its effect on genome structure.

EVOLUTION OF SEX-BIASED EXPRESSION IN CAENORHABDITIS

By

Cristel G. Thomas

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2011

Advisory Committee:
Associate Professor Eric S. Haag, Chair
Associate Professor Iqbal Hamza
Professor Thomas Kocher
Associate Professor Stephen Mount
Professor Leslie Pick, Dean's Representative

Acknowledgements

There are many many people I want to thank for helping me go forward while keeping my head in the right direction and making all this a worthwhile and fruitful experience.

None of this would have ever happened without the trust Leslie Pick put in me. I am very grateful that you believed in me, gave me the opportunity to come study at Maryland, and were always available for support and advice.

I would like to thank my advisor Eric Haag for his undying enthusiasm and optimism and his constant support. You opened many doors that I thought were closed or that I didn't even know existed, inspired me to be a better thinker, pushed me to learn many new things, both in the lab and out, let me draw worm cartoons, and most importantly, you taught me that not stressing out also works. I am honored that I could be a Haag lab Ph.D. student.

I also want to thank the other past and present members of my committee, Jonathan Dinman, Iqbal Hamza, Tom Kocher and Steve Mount for their valuable advice, help, ideas, comments and support.

I could have never done most of this work had Ian Korf and Keith Bradnam not spent time teaching me the basics of Unix and PERL. Thank you for welcoming me kindly in Davis and helping me think about worm sex.

An important portion of my dissertation research was made possible by the implication of both Brian Oliver and Harold Smith. Thank you both for trusting me with this projet, and supporting me through its realization. I would also like to thank the members of the Oliver lab for help and valuable input, more specifically Renhua

Li – thank you for being patient and teaching me R ! Warm thanks as well to Carlo Artieri, Lichun Jiang, John Malone, Nicolas Mattiuzzo and Dave Sturgill for valuable scientific discussions.

I want to express my gratitude to the people in the Biology department for making the BPS building a nice place to work, especially Professor Ades, for allowing me to keep the best TA assignment ever, Lois Reid and KeCia Harper for making the american paperwork easy, and Wan Chan for always fixing our computers with a smile.

I would also like to thank the evolution crowd on campus and the worm people in the DC area for valuable feedback, especially Matt Conte, Kelly and Claire O'Quin, Reade Roberts and Masato Yoshizawa for random chats and beers, as well as fellow MOCB grad students, especially Kawther Abdilleh and Alison Heffer for always finding time to get lunch or coffee.

I would like to acknowledge past and present Haag lab members, Alana Doty, Onye Eke, Robin Cook Hill, Joe Ross, Shanni Silberberg, Sarah Stadig, Gavin Woodruff, for their constant good spirit, optimism, support and valuable scientific and less scientific discussions. Special thanks to my wonderful help, Tong Lu. Qinwen Liu, thank you for always showing me the bright side of things, you made my days better on very many occasions. Dorothy Johnson, thank you so much for all the books, new literary perspectives, and random laughs!

On a more personal note, moving to Maryland was a big step to climb, and I couldn't have done it without my family and friends to support me. I want to thank them, the ones who are back in France and without whom my sanity would have been lost somewhere along the way, as well as the new ones I met here and made my time in Maryland all the more pleasant. So Marie-Alice, Benco, Delphyne, Frere, Matt, Olivier, Caroline, thank you all for keeping up with important non-important emails, calls and visits. Papa, Maman, thanks for believing in me all the way. Papy, si tu ne m'avais pas appris à démonter proprement les choses et à dessiner des culottes aux éléphants, je ne serais pas là où je suis maintenant. Merci pour tout !

Lucile, thank you very much for the smiles, coffees, and discussions about science, programming and other random things. Florent, not only were you a great help on a scientific level, both with thinking and willingness to share reagents, you also were there to remind me that there is a light at the end of the tunnel. Bertrand and Robin, thank you so much for your constant good spirit and support throughout the numerous processes of debugging and interminable series of stupid questions. Danielle, thank you so much for all the times you took time off of your busy schedule, the advice, the discussions, and the old ladies evenings.

Last but not least, Keenan, a thank you – or many – would not be enough. Thank you for indulging me with the weekends, odd hours, evenings and nights in the lab, the French time dinner, and the craziness that follows inevitably the graduate lifestyle. And everything else.

Table of Contents

List of Tables

Introduction

Chapter 1

Chapter 2

Chapter 3

List of Figures

Introduction

Chapter 1

Chapter 2

Chapter 3

Introduction

I. Androdioecy and genomic consequences of selfing

Androdioecy, a rare mating system in which hermaphrodites and males coexist, was thought for a long time to be so unstable that it would never be observed (Charlesworth 1984). Indeed, this mating system is believed to be a transition state between hermaphroditism and gonochorism (or dioecy, males and females) along with more frequent gynodioecy (hermaphrodites and females) and even rarer trioecy (hermaphrodites, males and females) (Charlesworth and Charlesworth 1978). Consistent with this, less than a hundred such species have been described so far, most of which are plants and crustaceans (reviewed in Pannell 2002; Weeks, Benvenuto et al. 2006). The first functional androdioecious species described led to the idea that this mating system evolved only from selection for self-fertile hermaphrodites in an initially dioecious system (Pannell 2002). However recent phylogenetic analyses of Limnadia crustaceans (Weeks, Chapman et al. 2009) and Caenorhabditis nematodes (Kiontke, Gavin et al. 2004; Kiontke and Fitch 2005; Kiontke, Félix et al. in revision) indicate that androdioecy evolved from hermaphroditic and gonochoristic ancestors respectively in these genera. Accordingly, androdioecy is a result of different molecular processes in these different species. While ZZ individuals in the clam shrimp *Eulimnadia texana* are males, and ZW and WW are self-fertile hermaphrodites (Sassaman and Weeks 1993), males in *Caenorhabditis elegans* result from X chromosome non-disjunction at meiosis (Nigon 1951). The picture is entirely different in plants, where most outcrossing is a consequence of loss of self-compatibility (reviewed in Charlesworth 2006; Mable 2008).

Regardless of how varied androdioecious systems can be, the theoretical genomic consequences are identical. As alluded to above, the evolution of self-fertility not only provides a way to invade a new ecological niche or survive drastic times, it also has major consequences on genome content and structure. Selfing leads to progressive homozygosis of subsequent generations, which in turn very often results in decline in fitness. The genetic bases for this phenomenon, called inbreeding depression, have been debated for decades (see review by Carr and Dudash 2003), but all stem from the consequences of unmasking deleterious recessive mutations through loss of heterozygosity. In order to survive, selfing species suffering from inbreeding depression have to get rid of residual deleterious mutation through a process called purging. Outbreeding depression is the opposite phenomenom whereby progeny from outcrossing have lower fitness than progeny derived from selfing. Interestingly, Eulimnidia shrimps display ongoing inbreeding depression (Weeks, Marcus et al. 1999) and no evidence of purging selection (Weeks 2004), whereas selfing Caenorhabditis nematodes do not (Johnson and Wood 1982; Johnson and Hutchinson 1993; Chasnov and Chow 2002). They instead suffer from outbreeding depression (Dolgin, Charlesworth et al. 2007), standing in marked contrast to their gonochoristic congeners, which display inbreeding depression (Dolgin, Charlesworth et al. 2007; Baer, Joyner-Matos et al.).

The succession of inbreeding depression and purging of deleterious mutations predict that self-fertilizing species should retain less genetic diversity than dioecious counterparts. Indeed, higher selfing rates lead to a decrease of the effective population size for neutral alleles of autosomal genes (Pollak 1987). Because equilibrium levels of neutral variation depend on the effective population size (Kimura and Ota 1971), the neutral theory of evolution predicts that outcrossing species will display twice as much genetic diversity as selfing species. In addition, other factors are suspected to further reduce genetic diversity in selfing species (Charlesworth and Wright 2001; Artieri, Haerty et al. 2008; Cutter, Wasmuth et al. 2008). For instance, severe population bottlenecks are facilitated by the fact that any single individual can found a new population. Moreover, progressive homozygosis of the genome leads to reduced effectiveness of recombination, and as heterozygote individuals are eliminated over the generations, balanced polymorphisms are lost. As a result the levels of linkage disequilibrium across the genome of selfing species increase. The impact of natural selection is then limited since both deleterious and advantageous loci are effectively linked to one another.

The above dynamics might lead to reduced selection against weakly deleterious mutations, and thereby accelerate rates of evolution in sequences located in non-recombining parts of the genome (Charlesworth and Wright 2001). On the other hand, selection for a new beneficial mutation would be expected to sweep a large swath of the genome along with it, lowering population-level variation. Accordingly, less nucleotide diversity is observed in the genomes of selfing Caenorhabditis than in gonochoristic relatives (Graustein, Gaspar et al. 2002; Jovelin, Ajie et al. 2003; Cutter, Baird et al. 2006; Jovelin, Dunham et al. 2009), and is lowest in the low-recombination central domain of the *C. elegans* chromosomes (Rockman and Kruglyak 2009). Similar findings have been described in plants (e.g. Savolainen, Langley et al. 2000; Baudry, Kerdelhue et al. 2001).

In addition to lower genetic diversity, effective population sizes and genome size are generally inversely correlated, thought to be the result of weaker selection against various forms of selfish elements in small populations (Lynch and Conery 2003). This is consistent with observations of abundance of transposable elements (TE) in regions of the *Drosophila melanogaster* genome where low recombination is observed (Charlesworth, Lapid et al. 1992), and higher copy number of TEs from the Ac-III and Tc1 family respectively in *Arabidopsis thaliana* (Wright, Le et al. 2001) and *C. elegans* (Dolgin, Charlesworth et al. 2008) compared to their equivalent in gonochoristic congeners. However, more recent studies suggest that these findings are not general (see below).

The Caenorhabditis and Arabidopsis genera are the most amenable to thorough examination of differences between closely related selfing and outcrossing species. Interestingly, both the current assessment of patterns of molecular evolution and genome sizes contradict some of the predictions mentioned above. Comparison between sequences from *A. thaliana* and *A. lyrata* to measure rates of protein evolution and codon-bias - thought to be the result of weak selection - found no significant differences between species (Wright, Lauga et al. 2002). Similarly lineage-specific substitution rate and codon usage bias assessment in six Caenorhabditis species did not reveal conspicuous evidence of the impact of inbreeding on the rate of evolution suggesting that the origin of selfing in *C. elegans* and *C. briggsae* is a recent event (less than 4 MYA, Cutter, Wasmuth et al. 2008).

Both studies concluded that more data would be required in order to understand better population histories of selfing species and ultimately rates of genomic evolution.

In both Arabidopsis and Caenorhabditis the genome sizes of selfing species are consistently smaller than that of outcrossing congener species (Johnston, Pepper et al. 2005; Oyama, Clauss et al. 2008; Lysak, Koch et al. 2009; Thomas, Li et al. In prep.). Furthermore, direct comparison of TE content through genome sequencing in *A. thaliana* and *A. lyrata* showed that in addition to being more numerous, TEs were also more active and younger in *A. lyrata* (Hu, Pattyn et al. 2011). While no similar study has been performed to date in Caenorhabditis, the genome sequences of five species have been scrutinized for repeat content, and those of both selfing species *C. elegans* and *C. briggsae* harbour fewer repeats than those of gonochoristic species (Table 1, Caenorhabditis Repeat Libraries) in contrast to predictions by Dolgin et al (2008). Theoretical models developed by Wright and Schoen (1999) explored the extent of TE accumulation in genomes of selfing species depending on the consequences of their insertion on fitness. Models for deleterious recessive insertions, in which homozygous insertions only have an effect on fitness, co-dominance (all insertions) and ectopic exchange (heterozygous insertions only) were tested and only in the latter were TEs found to increase in frequency as selfing rates increased, suggesting that, at least in the case of *A. thaliana* and Caenorhabditis nematodes, ectopic exchange is less likely to be involved, and illustrating the importance of oucrossing for the maintenance of TE in populations. Interestingly, segregation analysis in *A. thaliana* indicated that the process of DNA loss is ongoing (Hu, Pattyn et al. 2011) although the molecular mechanism by which this occurs has not been identified. Similarly, genome shrinkage in Caenorhabditis is likely to be ongoing and driven by a recently described mechanism whereby deletions are preferentially segregating along with the X chromosome in male meiosis (Wang, Chen et al. 2010).

Table 1. Repeat contents of the Caenorhabditis genome sequences. The libraries were obtained from Caenorhabditis Repeat Libraries

	Number of repeats	Total length of repeats (bp)
C. japonica	3,426	1,111,936
C. elegans	1,666	477,915
C. brenneri	5,791	1,627,202
C. remanei	5,662	1,450,462
C. briggsae	2,238	607,711

II. Caenorhabditis genus as a model system

Biological systems allowing direct comparison of androdioecy and gonochorism are extremely limited by the rareness of androdioecy. The Caenorhabditis genus can provide formidable insight into mating system transition, not only because *C. elegans* is a model organism, but also because many species have been described, their phylogenetic relationship are well established, and genomic data are available for a growing number of them.

1. Androdioecy in the Caenorhabditis genus

In Caenorhabditis (Fig. 1), androdioecy arose independently at least three times from gonochoristic ancestors (Cho, Jin et al. 2004; Kiontke, Gavin et al. 2004; Kiontke and Fitch 2005; Kiontke, Félix et al. in revision). The inference of hermaphroditism as a derived character state is further supported by the absence of multi-taxon clades comprising only selfing species (Kiontke and Fitch 2005). Molecular evidence additionally confirm convergent evolution of selfing in *C. elegans* and *C. briggsae* as germline sex determination proceeds through different actors and alternative mechanisms in the two species (Nayak, Goree et al. 2005; Hill, de Carvalho et al. 2006; Guo, Lang et al. 2009; Beadell, Liu et al. In revision). Indeed, while the core sex determination pathway is conserved, regulation of the mechanism allowing hermaphrodites to produce sperm involve different factors in *C. elegans* and *C. briggsae*, some of which species-specific, such as *fog-2* in *C. elegans* (Nayak, Goree et al. 2005) or *she-1* in *C. briggsae* (Guo, Lang et al. 2009). In addition, a few conserved sex determination genes have distinct sex determination phenotypes in those species (Hill, de Carvalho et al. 2006), indicating that the switch between spermatogenesis and oogenesis is controlled in different ways in the two selfing species *C. elegans* and *C. briggsae*.

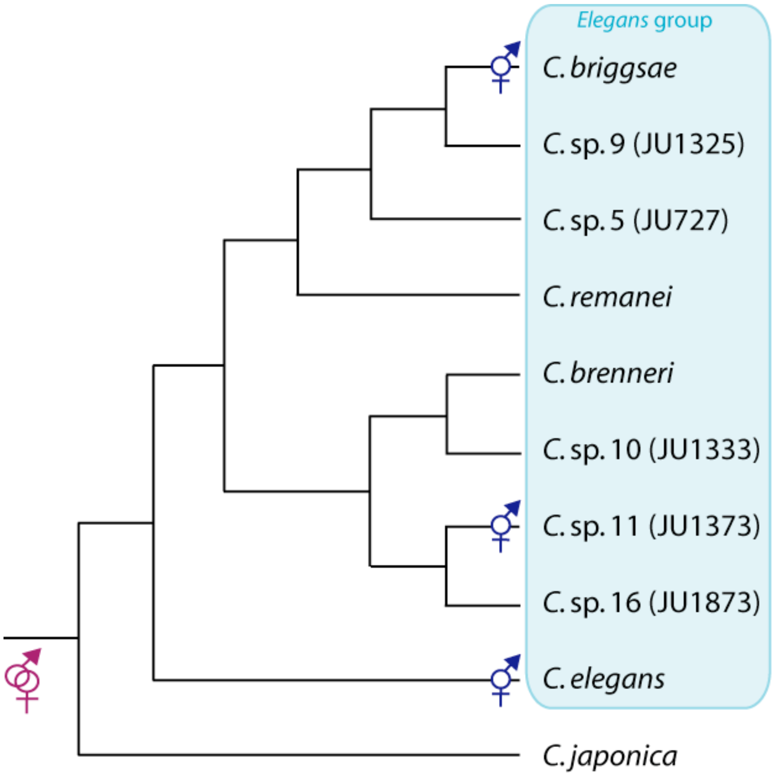


Figure 1. Cladogram of the current phylogenetic relationships between *Caenorhabditis* of the *Elegans* group. Cladogram (modified from Kiontke, Félix et al. in revision) from a maximum likelihood bootstrap analysis. Androdioecy evolved convergently in three lineages from a common gonochoristic ancestor. *C. japonica* belongs to the *Japonica* group and is indicated as an outgroup.

The estimated divergence time between *C. elegans* and its closest relative is about 18 MYA, and the origin of selfing in *C. elegans* and *C. briggsae* is a recent event, estimated at less than 4 million years ago based on a relaxed selection model of codon usage bias (Cutter, Wasmuth et al. 2008). Furthermore, selfing was inferred to have arisen more recently in *C. briggsae* consistent with the Caenorhabditis phylogeny, and further confirmed by a second study based on a wider range of orthologues gene sequences (Artieri, Haerty et al. 2008).

Ten species among the genus have or will soon have their genomes completely sequenced, annotated and available (E. Schwartz, pers. comm., The *C. elegans* consortium 1998; Stein, Bao et al. 2003; Hillier, Coulson et al. 2005; Gerstein, Lu et al. 2010; Ross, Koboldt et al. 2011). Surprisingly, flow cytometry analyses indicate that the genome sizes of gonochoristic species are 30 to 100% larger than that of both androdioecious species (Thomas, Li et al. In prep.). Preliminary assemblies of the *C. japonica, C. remanei* and *C. brenneri* genomes also suggest larger gonochoristic genome sizes. Some of this is explained by the substantial amounts of heterozygosity retained in the sequenced strains despite 20 generations of inbreeding (Barriere, Yang et al. 2009). However, even when heterozygosity is considered, the gene content of the *C. remanei* and *C. japonica* genomes is still higher than that of *C. elegans* or *C. briggsae* (CGT, pers. obs.). While genome sizes and emergence of selfing could be unrelated, it would be more parsimonious to consider that the genome sizes of both *C. elegans* and *C. briggsae* shrunk after they adopted selfing (3 independent events of genome shrinkage), rather than the opposite hypothesis of a parallel genome expansion phenomenon in all gonochoristic species (6 independent events of genome expansion at least). This hypothesis is supported by the polarity of character changes between selfing species and their gonochoristic congeners. For instance, plugging is lost in some strain of *C. elegans* due to insertion of a retro-transposon in the gene coding for the mucin protein that composes most of the plug (see below). Whether this drastic diminution in genome size is due to deletions of repetitive sequences or TEs as observed in *A. thaliana* compared to its obligate outcrossing congener species *A. lyrata* (Hu, Pattyn et al. 2011), or to loss of *bona fide* nematode genes remains unclear (but is addressed by this study). If it is the latter, however, an interesting hypothesis is that many genes that are lost are involved in sexual behaviour and cross-fertile reproduction. Indeed, as self-fertile individuals, neither *C. elegans* nor *C. briggsae* hermaphrodites need to rely on outcrossing with males to maintain their population, leading to a relaxation of selection on traits related to mating and mating efficiency (see below).

Sex is determined by the ratio of X chromosomes to autosomes in Caenorhabditis, hermaphrodites and females being XX and males XO (Nigon 1951). Males result either from meiotic X chromosome non-disjunction, leading to gametes carrying no X chromosome - and very few males in the subsequent generation, or through fertilization of XX individuals by males, increasing the fraction of males in the resultant progeny to up to 50% (Ward and Carrel 1979). In selfing species the male frequency is rather low, close to that of X chromosome non-disjunction rate in a standard lab strain, N2 (male frequency ≤ 0.002: Chasnov and Chow 2002; Teotonio, Manoel et al. 2006; non-disjunction rate 0.001-0.004: Hodgkin, Horvitz et al. 1979; Rose and Baillie 1979; Cutter and Payseur 2003; Teotonio, Manoel et al. 2006). This indicates that although male functions are maintained, maintenance of males in those population is weak in selfing species. Outcrossing does still occur at low levels in

C. elegans and *C. briggsae* (Denver, Morris et al. 2003; Barriere and Felix 2005; Sivasundar and Hey 2005; Cutter, Wasmuth et al. 2006; Barriere and Felix 2007). Although males in nature are rare (Barriere and Felix 2005; Barriere and Felix 2007) in some strains or natural isolates of *C. elegans* males persists at a higher rates than in the standard N2 strain, and display various abilities to mate (Hodgkin and Doniach 1997; Teotonio, Manoel et al. 2006). A natural population of *C. elegans* carrying a mutation *in mab-23* which renders the males unable to mate has however been reported (Hodgkin and Doniach 1997; Lints and Emmons 2002), illustrating the fact that males are not needed, at least for short-term survival of the population.

The reason why males are still retained in androdioecious species is a subject of open debate (Chasnov and Chow 2002; Stewart and Phillips 2002; Wegewitz, Schulenburg et al. 2008; Anderson, Morran et al. 2010). Strong selection against dioecy in *C. elegans* has been demonstrated by using mutations in *fog-2* or *him-5* (Chasnov and Chow 2002; Stewart and Phillips 2002), which respectively transform hermaphrodites into functional females (Schedl and Kimble 1988) or increase the frequency of spontaneous males (Hodgkin, Horvitz et al. 1979). In these assays, conducted in lab environments, self-fertile hermaphrodites consistently invade populations generation after generation, even when starting with a population composed of nearly 50% of males (Chasnov and Chow 2002; Stewart and Phillips 2002; Teotonio, Manoel et al. 2006). While in benign conditions males are maintained in the population at the X chromosome non-disjunction rate, and cannot increase in frequency because of their inefficiency at mating, another theory is emerging. When environmental conditions are not optimal (i.e. scarce food resources, overpopulation and temperature), Caenorhabditis nematodes have the possibility during development to go into a phase called dauer (Cassada and Russell 1975). Males survive dauer arrest better than hermaphrodites, and outcrossing rates increase after starvation (Morran, Cappy et al. 2009). In addition, in populations subjected to environmental stresses or increased mutation load, progeny resulting from crosses between males and hermaphrodites have a higher fitness than that from selfing (Morran, Parmenter et al. 2009). In these outcrossing populations, male frequency increases as a result of direct selection for cross-fertilization, as opposed to increased rates of X chromosome non-disjunction (Morran, Cappy et al. 2009). Thus, outcrossing allows both avoidance of inbreeding depression and more rapid adaptation to new environments for *C. elegans* - as is the case for most species.

2. Mating behavior evolution

Hodgkin (1983) noted that mating in *C. elegans* relies on physical ability to mate (i.e. body shape) as well as on chemosensory and neuronal signals. Mating behaviour in *C. elegans* has been extensively studied (Barr and Garcia 2006), and males display the most obvious mating behaviour. Briefly, upon contact between the rays and ventral sensilla of the male tail with the hermaphrodite cuticle, the male presses his tail against the hermaphrodite's body and starts moving backwards. This motion allows him to scan the hermaphrodite to find the vulva. If the male reaches the head or the tail of the hermaphrodite without having located the vulva, he coils his body ventrally and starts scanning the other side of the hermaphrodite, without breaking the contact. When the male tail eventually makes contact with the hermaphrodite vulva, the male stops moving and inserts his spicules in the vulval slit.

The male then ejaculates and leaves on the vulva what Hodgkin and Doniach (1997) described as a "large blob of material", the copulatory plug. Larger, male-derived sperm is stored in the hermaphrodite spermatheca, along with the hermaphrodite's own sperm, upon which it has a competitive advantage because of its larger size (LaMunyon and Ward 1998; LaMunyon and Ward 1999). In the absence of a mate, males wander away from food source presumably in search of new mates (Simon and Sternberg 2002).

However well described, mating is not an efficient process in androdioecious nematodes. Since mating is not required anymore for reproduction, mating-related traits should be under much weaker selection. Such traits should display increased phenotypic and molecular variation, both within androdioecious species and between related gonochoristic and androdioecious species. Accordingly, in C. *elegans* and *C. briggsae*, both genders have lost their ability to behave as reliable mating partners in comparison to gonochoristic Caenorhabditis species (Chasnov and Chow 2002; Stewart and Phillips 2002; Chasnov, So et al. 2007; Cutter 2008, see below). Most laboratory crosses between males and hermaphrodites in *C. elegans* or *C. briggsae* use immobilized or strongly uncoordinated hermaphrodites or excess of males to make up for the low frequency of successful matings. Furthermore, males from androdioecious species do not discriminate between genders and tend to initiate mating behaviours upon contact with any cuticle, including their own (Garcia, LeBoeuf et al. 2007). When the object of the attentions of a male is another male or himself, the contact between the excretory pore and the male sensory neurons in the tail triggers the rest of the mating behaviour sequence and the male deposit a plug on the excretory pore.

Several steps in the mating process are degraded in androdioecious species. Females from dioecious species attract conspecific males as well as *C. briggsae* and *C. elegans* males, while hermaphrodites do not elicit such a reaction from males from any species (Chasnov and Chow 2002; Chasnov, So et al. 2007; Garcia, LeBoeuf et al. 2007). Females from *C. remanei* and *C. brenneri* secrete a sex pheromone to attract males that seems to either not be produced by hermaphrodites, or to be less efficient (Chasnov, So et al. 2007). *C. elegans* hermaphrodites do produce a mixture of ascarosides that attract both conspecific males as well as males from *C. brenneri* and *C. remanei*, and those from *C. briggsae* and *C. japonica* to a lesser extent (Srinivasan, Kaplan et al. 2008). However no study has been undertaken yet to assess whether the attractants secreted by the females from dioecious species are chemically similar, and comparably potent to attract males to that secreted by *C. elegans* hermaphrodites.

Once males from dioecious species have found a female partner, they are able to immobilize it to facilitate spicule insertion and consequent copulation (Garcia, LeBoeuf et al. 2007). Garcia, LeBoeuf et al. (2007) suggested that males produce a 'soporific factor' that allows them to immobilize the female and facilitates spicule insertion by widening the vulval opening. However, hermaphrodites seem to not be responsive to this signal, and most of the matings fail because the hermaphrodite moves away from its partner before copulation occurs. Interestingly, *C. briggsae* males induce the same inactive behaviour in females from both gonochoristic species tested, whereas *C. elegans* males do not (Garcia, LeBoeuf et al. 2007). Similarly, a strain of *C. briggsae* otherwise able to mate seems to lack sex-drive and displays neither mating nor self plugging behaviours (Garcia, LeBoeuf et al. 2007). In addition, hermaphrodites tend to eject male-sperm if their reproductive tracts already contain their own, particularly if they are young (Barker 1994; Kleeman and Basolo 2007). Self-sperm-depleted hermaphrodites, while more inclined to mate, still may eject male sperm after outcrossing (Kleeman and Basolo 2007) which seems completely counter-productive in terms of reproduction assurance.

Lastly, only males wander away in search for mates in androdioecious species. In contrast, both males and females in dioecious species wander in the absence of potential mate, indicating that hermaphrodites do not search for mates (Lipton, Kleemann et al. 2004).

All these studies confirm that androdioecious species have lost (or are in the process of losing) some characteristics that would otherwise make them efficient cross-fertilizers. The genes underlying these traits and/or their regulation may have evolved in response to relaxation of selection (Cutter 2008) and some are still responsive to selection (LaMunyon and Ward 2002). A few labs have taken an interest in the sex-specific transcription of *C. elegans* (Reinke, Smith et al. 2000; Jiang, Ryu et al. 2001; Reinke, Gil et al. 2004; Thoemke, Yi et al. 2005) and of parasitic nematodes (Nisbet, Cottee et al. 2008). These workers have collectively identified somatic- and germline-enriched mRNAs, both in hermaphrodites and males, as well as sperm- and oocyte-enriched. However, given that androdioecious species have clear mating defects, and that sex-specific mutants are rare in *C. elegans*, genes associated with mating success may not be easily identifiable in *C. elegans*. In contrast, identification and study of the genes underlying these traits will be more efficient in a gonochoristic species such as *C. remanei*. Nisbet et al. (2008) further underlined the importance of identifying sex-specific genes in Caenorhabditis to be able to better compare and understand the reproduction and sex determination processes in the related parasitic nematodes, in order to better design new anthelmintics and control strategies.

III. RNA-seq as a tool to detect differential expression

Recent breakthroughs in DNA sequencing technology have led to the advent of massive parallel sequencing. Total DNA or RNA from any sample can now be sequenced at great depth in single runs generating millions of short sequences (36-200 bp depending on the technology). These new methods, termed DNA-seq or RNA-seq, are especially attractive because of their complete independence from any reference sequence. They are being used exponentially for genome sequencing projects, tumor characterizations, transcriptomic analyses, assessment of differential expression, identification of splice variants, and many other applications (e.g. Lister, O'Malley et al. 2008; Mortazavi, Williams et al. 2008; Sultan, Schulz et al. 2008; Li, Fan et al. 2010; Trapnell, Williams et al. 2010; Iorizzo, Senalik et al. 2011). The possibilities and challenges allowed by RNA-seq have been reviewed recently (Wang, Gerstein et al. 2009; Oshlack, Robinson et al. 2010; Salzberg 2010), and the goal here is only to give a brief overview of the methodology.

Briefly, total or poly-A enriched RNA are extracted and purified, and then used to generate cDNA libraries that may or may not be normalized. A fragmentation step is often added to make long sequences more amenable to RNA-seq. These libraries are then deep-sequenced using any of three technologies, SOLiD (Applied Biosystems), 454 (Roche LifeScience) or Solexa / Illumina (Illumina). The method marketed by Illumina/Solexa uses a solid-phase bridge amplification in which adapters are used to link cDNA fragments to a flow cell. The amplification takes place locally (i.e. where the DNA strand is bound to the slide) and 'colonies' of DNA fragments are generated at random on the flow cell, all of them unique because generated from a unique DNA molecule. Both sequencing technologies developed by 454 Life Sciences and Roche Applied Science use a step of emulsion PCR to amplify the cDNA library. One of the adapters ligated to the cDNA fragment allows its binding to beads – ideally one molecule per bead. The DNA-bound beads are then emulsified with the PCR reagents, such that, ideally, one drop of aqueous phase contains only one bead, and hence a unique template for the PCR reaction. Both these amplification techniques allow for a very large number of amplification reactions to occur at the same time. The sequencing step in the protocol proposed by Illumina/Solexa consists in monitoring the incorporation of labeled terminators, cycle after cycle, using the flow cell on which the amplification of the cDNA library was performed as template. The pyrosequencing method from 454 Life Sciences is similar in the sense that the sequencing is monitored as the reaction of polymerization occurs. However, the deoxynucleotides used for the extension are not terminators, and the number of incorporated bases per cycle is measured by the intensity of the luminescent signal. Lastly, the SOLiD technology of Roche Applied Science is based on the specificity of hybridization of a known octamer to the single stranded DNA template being sequenced.

Any of the above platforms will produce a vast quantity of short sequences corresponding to one sample. RNA-seq in itself thus does not take many steps, but each harbors its own biases. For instance, the fragmentation step can be done before or after cDNA synthesis, by sonication, nebulization, hydrolysis or DNase I treatment, and each of those have their own caveats (Wang, Gerstein et al. 2009). Some of the kinks have been identified and incorporated in further analytical steps, but the technology is new enough that the field is still unsure about what analytical pipeline is the best, be it for the RNA-seq itself or the subsequent analysis of the data.

Numerous open source softwares are available to analyse data obtained by RNA-seq, allowing mapping of short sequences, or reads, to a reference sequence. Mapping the reads enables one to attribute them to genes and subsequently estimate gene expression levels. In the absence of gene annotations or reference sequences other methods can be used. The massive amount of sequence information generated by RNA-seq can be utilized to generate gene models for the genome used as a reference, based on mapping information. Alternatively, *de novo* reference sequences can be assembled (reviewed in Martin and Wang 2011).

Different softwares handle RNA-seq data in various manners, but have to face the same challenges. Genome sequences represent a snapshot of the organism's status at one point in time, and samples sequenced might contain polymorphisms not captured at the time of sequencing. In addition, short reads might map equally well to several locations in the genome because of repeats, recent gene duplicates or conserved domains. Further, reads from RNA-seq experiments represent a spliced transcriptome and isoforms have to be taken into account. Lastly there is of course the possibility of sequencing errors. While many ways have been developed to take these parameters into consideration for assessing gene expression, no clear consensus has been reached yet as to what the best method is (Oshlack, Robinson et al. 2010), especially since different experiments or experimental designs might call for specific ways to analyze the data produced by RNA-seq.

Finally, like any other high-throughput gene expression detection method, RNA-seq is not necessarily only limited by the measuring aspect itself, but also by the experimental design and the power of statistics (Gibson and Weir 2005). The number of technical or biological replicates is critical in assessing differential expression properly, as the statistical method and the definition of significance are.

IV. Present study

Caenorhabditis provides a unique opportunity to study the consequences of mating system switch at the genomic level. The present study is driven by several hypotheses, related to one another:

I propose that as selfing arose in this genus, lower selective pressure on mating efficiency led to lower selective pressure on the maintenance of the regulation of expression of genes involved in mating processes. This is illustrated by overall poor performances of both males and hermaphrodites of androdioecious species in all processes related to mating, compared to males and females of gonochoristic species (see section II.)

In addition, I suspect that in selfing species, a subset of these genes is likely to have been lost, partially contributing to the decrease in genome size observed in *C. elegans* and *C. briggsae*. Although it is very possible that new genes arose in selfing species, loss of genes – or decreased expression level of genes – is more consistent with loss of traits observed in selfing species.

I expect that traits related to sex and mating will be conferred by genes with sex-biased expression, since these traits are specific to one or the other sex. I further hypothesize that loss of sex-related traits in selfing species was accompanied by loss of genes with sex-biased transcript abundances.

Lastly, I conjecture that in selfing lineages, the rate of loss of genes inferred to have had sex-biased transcript abundances should be greater than expected by chance, since the main difference between selfing species and gonochoristic species is the mating system they employ.

In the first chapter, I use *C. remanei* as a proxy for the gonochoristic common ancestor of the *Caenorhabditis* genus in a preliminary survey to study these genes and their evolution. I first identify genes with sex-biased expression by comparing *C. remanei* male- and female-specific transcriptome. I then evaluate the extent to which these genes with sex-biased expression are conserved in the *Caenorhabditis* genus and assess whether the frequency of loss of sex-biased genes in selfing species is significant when compared to their overall rate of gene loss.

Genes identified as having sex-biased expression and without homologues in androdioecious species are characterized in the second chapter. Their phylogenetic patterns are examined and their expression patterns defined. This comprehensive curation sheds some light on the function of genes specifically lost in selfing species and on the regulation of genes with male-biased expression in the Caenorhabditis genus.

Lastly, in the third chapter I explore sex-biased gene expression in the Caenorhabditis genus in a more thorough manner and identify genus-wide patterns of expression. Further, this more complete study allows placement of sex-biased expression of an androdioecious species in the light of gonochoristic congeners. I then identify interesting trends that allow a better understanding of genomic consequences of selfing emergence.

Chapter 1: Preliminary survey of sex-biased expression in *C. remanei*

I. Summary

Both *Caenorhabditis elegans* and its congener *C. briggsae* are androdioecious while their last common ancestor and most extant Caenorhabditis are gonochoristic (Cho, Jin et al. 2004; Kiontke, Gavin et al. 2004; Kiontke, Félix et al. in revision). The capability of gonochoristic Caenorhabditis to evolve self-fertilization may stem from their latent potential to produce and activate self-sperm in a timely developmental fashion from female gonads (Baldi, Cho et al. 2009). Thus, a minimum of two evolutionary steps are enough for selfing to arise in Caenorhabditis. However, a multitude of other evolutionary steps would be required in order for selfing to be maintained and become the main reproductive mode, just as substantial genomic and genetic consequences would be expected to accompany selfing emergence (reviewed in Charlesworth and Wright 2001). A complete understanding of the evolution of selfing would require determining how the regulation of sex determination has been modified historically to allow for germline bisexuality, as well as identifying other traits that have been modified through the rise and maintenance of self-fertilization as a breeding system.

In addition to having smaller genome sizes than the gonochoristic extant species in the genus (Thomas, Li et al. In prep.), *C. elegans* and *C. briggsae* both display a weakening of their overall mating efficiency and ability (Hodgkin and Doniach 1997; LaMunyon and Ward 1999; Chasnov and Chow 2002; Lipton,

Kleemann et al. 2004; Geldziler, Chatterjee et al. 2006; Chasnov, So et al. 2007; Garcia, LeBoeuf et al. 2007; Kleeman and Basolo 2007; Palopoli, Rockman et al. 2008), which has been seen as the hallmark of a selfing syndrome (Cutter 2008). Understanding the evolutionary loss of ancestral traits in the wake of an adaptation is an important complement to the identification of the developmental and genetic novelties that directly produce the adaptation. Some of the genes underlying mating-related traits are most likely not identifiable in either of the androdioecious species, because those traits are degraded in those species. However, using an extant gonochoristic species like *C. remanei* as a proxy for the gonochoristic ancestors of *C. elegans* and *C. briggsae* may allow their identification. In this study, we hypothesized that genes involved in traits related to mating processes, and lost in hermaphrodites, are likely to be strongly sex-biased in their expression in gonochoristic species, since such mating-related traits are sex-specific. Moreover, at least a subset of them will be lost in *C. elegans* and/or *C. briggsae,* assuming they confer traits lost or degraded in those species. In addition, we also hypothesized that genes with a sex-biased expression in gonochoristic species were lost at a higher rate in selfing species than expected by chance. The latter is based on the assumption that if indeed such genes should be involved in mating processes, they should be under relaxed selection more so than essential genes.

I used transcriptome deep-sequencing technology to identify genes with high sex-biased expression in *C. remanei*, and defined candidates for loss in androdioecious species based on their phylogenetic distribution within the Caenorhabditis genus.

In this work, I generated a low-depth RNA-seq dataset for *C. remanei* males and females, and developed an analytical pipeline that detected the expression of 6,268 genes, about a third of which were identified as significantly differentially expressed between the sexes. Among the genes that displayed a strong sex-bias in their expression, most had a homologous gene in *C. elegans* whose expression was similarly sex-biased as defined by previous micro-array studies (Reinke, Smith et al. 2000; Jiang, Ryu et al. 2001; Reinke, Gil et al. 2004), consistent with the idea that transcriptional variation is mostly conserved across species (Gibson and Weir 2005). In addition, *C. remanei* genes with highly sex-biased expression were found to be more likely to be missing homologues in one or both of the selfing species, suggesting that they might be preferentially lost through selfing emergence and evolution.

II. Material and Methods

1. Nematode culture and strains

C. remanei strain PB4641 was maintained on NGM agar plates (3 g/L NaCl, 2.5 g/L peptone, 22 g/L agar, 5 mg/L cholesterol, 2 mg/L uracil, 0.1 M $CaCl_2$, 0.1 M $MgSO_4$ and 25 mM KH_2PO_4) according to standard *C. elegans* methods (Wood 1988), using 2.2% agar to discourage burrowing. Unless otherwise noted, the NGM plates were seeded with OP-50 bacteria and worm stocks kept at 20˚C. To maintain outbreeding as much as possible and avoid inbreeding depression, large worm populations were maintained by transferring large chunks from two independent plates per passage.

2. RNA extraction

L4 to adult male or female worms were individually picked in M9 buffer (86 mM NaCl, 42 mM Na_2HPO_4, 22 mM KH_2PO_4, 1mM $MgSO_4$). Worms were washed twice in M9 and once in RNase free water and resuspended in 30 μL of RNase free water. 250μL of TRI-Reagent (Molecular Research Center) were added and the samples were frozen at -80˚C. The samples were then thawed, and pools of worms picked independently combined in order to increase the yield of the RNA extraction. 3,000 females and 5,000 males were then lysed using a plastic pestle. RNA was purified using phenol/chloroform extraction and subsequent isopropanol precipitation. The RNA was resuspended in RNase free water.

3. cDNA library and deep-sequencing

Non-normalized, sex-specific cDNA libraries were generated at the Washington University Genome Sequencing Center (WUGSC). Briefly, the total RNA was treated with DNaseI and reverse transcribed using SMART 5' and 3'-oligo-dT primers (Clontech), which allow for preferential amplification of full-length mRNA. Biotin-coupled primers complementary to the SMART oligos and containing a *Mme*I restriction site were then used in a second PCR round. The cDNA library was digested with *Mme*I to cut out the primers as well as most of the poly-A tails, and purified with magnetic beads. Lastly, the cDNA libraries were nebulized to generate random short fragment to sequence. Each cDNA library was sequenced on an Illumina Genome Analyzer, using the Solexa sequencing technology, twice. Each run produced 5 to 6 million 35bp- and 50bp-long reads, respectively for each of the two sequencing rounds.

4. Data analysis

4.1. Analysis – overview

There is currently a growing number of softwares and programs available to analyze RNA-seq data (reviewed in Oshlack, Robinson et al. 2010). However, at the time this analysis was performed, none were suited to fit the characteristics of our study. I followed an analysis outline similar to that of the published RNA-seq studies at the time (Fig. 1) (Cloonan, Forrest et al. 2008; Lister, O'Malley et al. 2008; Marioni, Mason et al. 2008; Morin, Bainbridge et al. 2008; Mortazavi, Williams et al. 2008; Nagalakshmi, Wang et al. 2008; Wilhelm, Marguerat et al. 2008), using PERL scripts developed by Ian Korf and myself (Appendix 1).

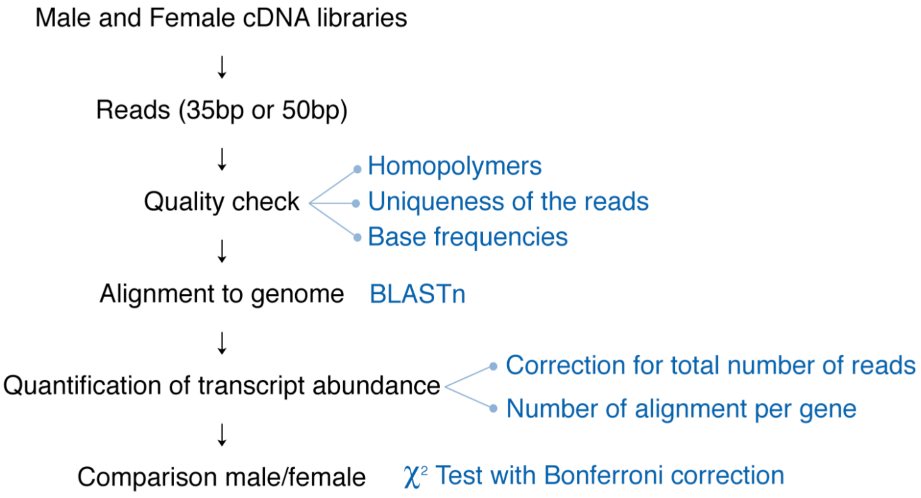

Male and Female cDNA libraries
Reads (35bp or 50bp)
Quality check
Homopolymers
Uniqueness of the reads
Base frequencies
Alignment to genome
BLASTn
Quantification of transcript abundance
Correction for total number of reads
Number of alignment per gene
Comparison male/female
χ^2 Test with Bonferroni correction

Figure 1. Flow chart of the RNA-seq analysis pipeline.

Solexa sequencing yields a large batch of short reads, which, in the case of mRNA deep-sequencing, have to be assigned to genes in order to retain any informational value. After checking the quality of the datasets, the reads were aligned to the reference sequence. I used the number of reads aligned to a gene as a measure of its expression level, and compared for each gene the expression levels in the male and female datasets. I then used a simple statistical test to assess whether the genes were significantly sex-biased, and among the genes diplaying a higher than 10-fold sex-bias, I chose those with interesting phylogenetic patterns to study further.

4.2. Analysis – Data

We chose to sequence the PB4641 *C. remanei* strain, an inbred derivative of EM464 also used for the *C. remanei* genome project. Despite 20 generations of inbreeding, approximately 10% of the PB4641 genome sequence actually represents alternative alleles (Barriere, Yang et al. 2009). The fastest and most efficient way to obtain pure pools of *C. remanei* males or females is to pick them individually by hand under a stereomicroscope. Alternative bulk methods, such as filtering (Reinke, Gil et al. 2004) or particle sorting (E.S. Haag, Y. Lin, unpublished obs.) are faster but introduce unacceptable cross-contamination.

A limit of manual sorting however is that the earliest stage at which the sex of wild-type worms can be determined is the larval L3-L4 stage. As a result, expression of the genes involved in early development, in late embryos to the L3-L4 larval stages will not be detected. However, the genes conferring the traits this study aims to identify are more likely to be expressed at later stages of development. The cDNA libraries prepared from L4 to adults males and females should then represent genes expressed both in the soma and in the germline, as well as in sperm and oocyte. In order to be able to measure expression levels, the cDNA libraries were not normalized. Each library was prepared twice independently and each of those sequenced twice. The first sequencing produced 35bp reads and the second 50bp reads.

4.3. Analysis – Quality control

Our quality control is based on a few assumptions. First, the homopolymer frequencies in the datasets should be low, consistent with that of the coding regions of the genomes as well as with the preparation of the cDNA library, protocol during which the poly-A tails of the mRNAs are cleaved off. Secondly, we expect very few of the reads to exist in more than one copy as each 35bp or 50bp tag represents a random start position of sequencing in the transcriptome. Lastly, we assume that the reads represent an unbiased sample of the whole transcriptome, so that the percentage of each nucleotide at each of the 35 or 50 positions of the sequence of the reads should be consistent with the overall frequency of each base in the transcriptome.

Table 1. Homopolymer Frequencies. The length of homopolymers is indicated by *k*. Datasets 1 (35bp) or 2 (50bp) are indicated in parentheses.

Datasets	*k*-mer size (bp)	Poly-A	Poly-C	Poly-G	Poly-T
Male set (1)	8	1.2 %	0.08 %	0.12 %	23.78 %
Female set (1)		3.26 %	0.15 %	0.13 %	16.9 %
Male set (2)		0.29 %	0.01 %	0.06 %	13.75 %
Female set (2)		0.31 %	0.02 %	0.07 %	12.83 %
Male set (1)	10	1.07 %	0.07 %	0.02 %	22.7 %
Female set (1)		3.05 %	0.13 %	0.02 %	16.13 %
Male set (2)		0.1 %	0.004 %	0.02 %	12.78 %
Female set (2)		0.07 %	0.01 %	0.02 %	11.56 %
Male set (1)	12	1.04 %	0.06 %	0.01 %	1.52 %
Female set (1)		2.89 %	0.12 %	0.01 %	1.19 %
Male set (2)		0.04 %	0.003 %	0.01 %	12.2 %
Female set (2)		0.03 %	0.01%	0.01 %	10.86 %

The homopolymer analysis revealed a surprisingly high fraction of 10-mers of thymidine in both 35bp and 50 bp datasets (Table 1). The frequencies of each base at each position on the reads (Fig. 2) also showed a departure from expectations. For the first 10bp, the most abundant nucleotides fluctuated dramatically, and taken in sequence correspond to the primer used to prepare the cDNA library. This primer shouldn't have been sequenced after digestion by *Mme*I. The reads whose sequence starts with the 10 first bp of the adapter sequence represent 25.8% of the first dataset overall (Table 2). The second round of cDNA libraries preparation and re-sequencing was supposed to be of higher and better quality but retained similar levels of the same artifact : 18.1% of the reads of the second dataset start with the 10 first bp of the adapter sequence. This sequence does not exist in the *C. remanei* genome and should not impair the downstream analysis.

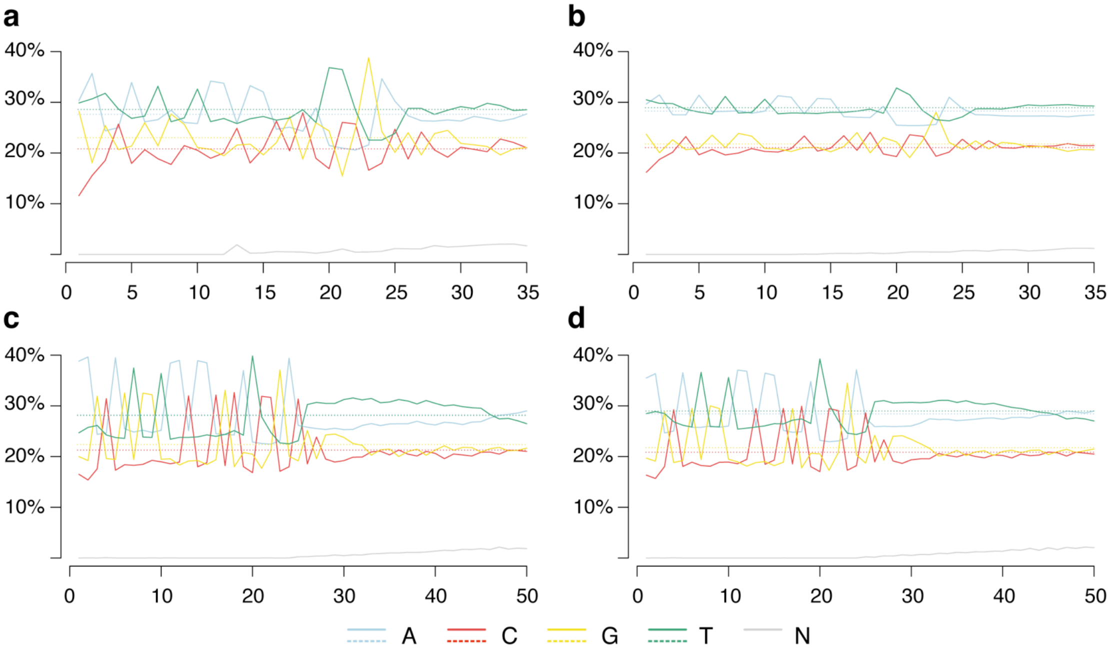


Figure 2. Nucleotides frequencies in datasets. Nucleotide frequency per position on the reads per dataset. Plain lines indicate nucleotide composition at each position and dotted lines frequency of each nucleotide in the dataset. The most abundant base for the first 25 nucleotides corresponds to the sequence of an adapter inavertently retained after cDNA library formation. **a**. Female cDNA library, dataset 1. **b**. Male cDNA library, dataset 1. **c**. Female cDNA library, dataset 2. **d**. Male cDNA library, dataset 2.

To relieve the load on CPU usage during later phases of the analysis, and to enrich for biologically relevant reads, we filtered the datasets to eliminate the non-unique reads. The number of unique reads represent 17.2% and 37.2% of the female and male 35bp datasets respectively, and 33.6% and 32.8% of the female and male

50bp datasets respectively (Table 2). This measure, though drastic, is compatible with the spirit in which this analysis was undertaken, a preliminary analysis of sex-specific transcriptomes to discover highly expressed, sex-biased genes. Given that the *C. elegans* transcriptome size has been estimated to be 27Mb, for a genome size of 100Mb (The *C. elegans* consortium 1998), by extrapolation from its genome size the transcriptome size of *C. remanei* should be 35.37Mb. The amount of unique reads generated by each sequencing set thus allowed a 6X coverage (35bp dataset) and a 21X coverage (50bp dataset) of the *C. remanei* transcriptome.

Table 2. RNA-Seq Statistics.

	Dataset 1		**Dataset 2**	
	Female set	*Male set*	*Female set*	*Male set*
Total number of reads	11,692,352	22,790,853	21,959,536	22,753,811
Total number of unique reads (fraction of total)	2,015,673 (17.2%)	6,141,457 (37.2%)	7,376,435 (33.6 %)	7,470,895 (32.8%)
Estimation of adapter contamination (fraction of total number of reads)	29.9%	21.6%	18.8%	17.3%
Estimation of transcriptome coverage	2X	4X	10.5X	10.5X
Fraction of unique reads aligning to a gene	55.2%	70.9%	64.9%	74.1%
Fraction of alignments to a gene that are unique	78 %	77.4%	86.4%	81.9%
Correcting factor	0.272	0.728	0.493	0.507
Number of genes detected	6,350		6,308	
Male-biased genes (over 10-fold)	849 (406)		1,644 (708)	
Female-biased genes (over 10-fold)	2,138 (327)		1,718 (244)	

4.4. Analysis – Alignment to the *C. remanei* genome

We aligned the unique reads to the publicly available *C. remanei* supercontig sequences (WS204 release) using BLASTn. We kept all alignments, including multiple alignments of a same read, if they were 95% identical and encompassed either the beginning or the end of the sequence tag. Additionally, we used a scoring table of +1 for matches and -1 for mismatches. Initially we required a minimum alignment score of 16 or higher. All together these criteria should enable detection of reads aligning completely to exons or straddling splice junctions, and allow pertinent alignment despite sequencing errors. The alignments were assigned to a given gene when they were located within a region encompassing 500bp upstream of the predicted start site to 500bp downstream of the stop codon.

4.4.a. Tuning of the alignment parameters

A preliminary alignment analysis was performed only on the first dataset (35bp) to determine parameters for optimal alignment of the reads to the genome. Using the aforementioned alignment parameters 22.3% of the reads matched a landmark defined as a gene region in the genome. Possible sources of unaligned reads include contamination from *E. coli* fed to the worms, adapter contamination and other artifacts of the sequencing methodology. Out of the 31,620 genes reported in the WS204 version of the annotation of the *C. remanei* genome, the same 6,424 were matched at least once, both in the male and in the female set, and 75.1% to 86.5% of the unique reads aligned at least once to a gene landmark (Table 3). Lastly, 68% of the alignments were located in regions of the genome where no gene was predicted to exist (Table 3).

Table 3. Comparison of alignment score threshold of 16 and 30.

Alignment score	16 and up		30 and up	
	Female set	*Male set*	*Female set*	*Male set*
Total number of alignments	20,402,686		5,750,425	
Alignments to genes	32 %		55.9 %	
-- Representing fraction of total reads	*22.3 %*		*17.7 %*	
Fraction of alignments to genes that are unique	31.4 %	29.3 %	78 %	77.4 %
Maximum number of alignments at the same location	2,966	2,147	984	603
Number of genes detected	6,424	6,424	5,959	6,232
Fraction of unique reads matching at least once to the genome	75.1 %	86.5 %	552 %	70.8 %

Surprisingly, the matches were not randomly distributed. Rather, an exact location of the genome may be hit up to about 3,000 times. Furthermore, the distribution of number of alignments per alignment score was U-shaped (Fig. 3a) indicating an unexpected large proportion (37.6%) of alignments with a score of 16, along with a more expected high proportion of alignments with a score of 33 (corresponding to 1 mismatch) and 35 (no mismatch). The scoring matrix defined for alignments explains the lower fractions of odd-numbered scores: a 34bp-long perfect alignment will have a score of 34, but a 35-bp long alignment with one mismatch has 34 matches (+ 34) and 1 mismatch (-1).

Reads aligning to splice junctions would have scores lower than 35, and their scores should be statistically evenly distributed, between 16 and 34, consistent with random sampling of the transcriptome. The massive over-representation of alignment scores of 16 therefore cannot be accounted for only by splice junction detection. It is very likely that these alignments are not biologically relevant, both because of the

U-shaped distribution of the alignment lengths, and because smaller sequences are more likely to match an inexact location of the genome, even with a high identity threshold. Given the distribution of the alignments per alignment scores, I set a new score threshold at 30 to carry out the analysis, as these alignments are more likely to represent biologically relevant hits to the genome (Fig. 3b). Only 28.2 % of alignments previously defined matched the new criterium (Table 3). Accordingly, the number of unique reads matching at least once to the genome decreased. However, the number of alignments to regions where no genes are predicted to exist, as well as the number of multiple hits at one location, representing irrelevant alignments, decreased considerably (Table 3). Furthermore, the proportion of unique alignments located within a gene landmark increased dramatically. Lastly, the number of genes hit in both male and female sets were different, yielding an overall number of genes detected of 6,350. All together, these numbers suggested that keeping alignments with scores beween 16 and 29 introduced a high background of biologically irrelevant matches.

4.4.b. Alignment to the C. remanei genome

Given the different performances of low (16) vs. high (30) alignment scores, I subsequently used a threshold of 30, along with the same BLASTn parameters : +1 for a match, -1 for a mismatch, and 95% identity, keeping only the alignments for which the first or last bases were a perfect match to the genome.

I performed the alignment of the reads from the second 50 bp-long reads dataset and the distribution of number of alignments per alignment score confirms that the majority of the alignments have a score of 46 (2 mismatches), 48

(1 mismatch) and 50 (perfect alignment) (Fig. 3b). The proportion of alignments with scores between 30 and 45 was low, confirming the validity of a score threshold of 30.

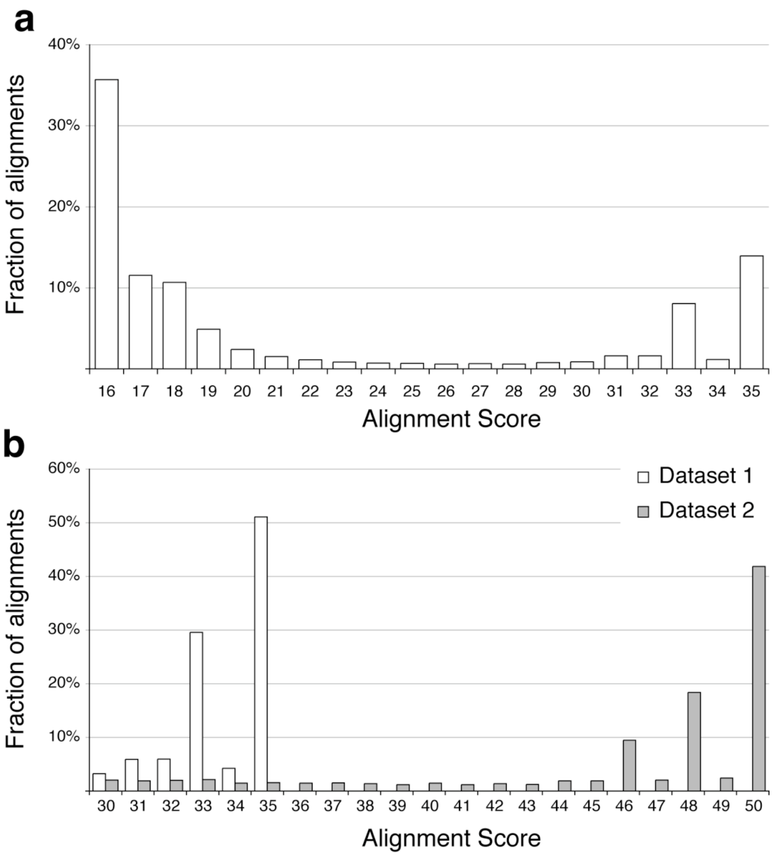


Figure 3. Distribution of alignment scores in both datasets. a. The fraction of read alignments to the *C. remanei* genome are shown for scores ranging from 16 to 35 (dataset 1, white). **b.** The fraction of read alignments to the *C. remanei* genome are shown after application of the threshold at 30 for scores ranging up to 35 (dataset 1, white) and 50 (dataset 2, grey).

4.5. Analysis – Quantification of gene expression

Quantification of transcripts by gene tag sequencing is based on tag counts per gene (reviewed in de Hoon and Hayashizaki 2008). In our case, the reads are expected to be randomly distributed across the coding regions as opposed to be representing only the 5' or 3' of each transcript. The expression level of each gene was defined by the number of uniquely mapping reads aligning within a region encompassing 500bp upstream of the predicted start site to 500bp downstream of the stop codon. The read counts per gene were adjusted by a correcting factor to compare meaningfully the female transcriptome to the male transcriptome. The correcting factor for each male and female set within a dataset was defined as the total number of independent loci hit within the gene regions in the set divided by the total number of loci hit in both the male and the female set (Table 1). The corrected counts for males and females datasets were compared gene by gene and significant sex-bias was assessed by a Bonferroni corrected χ^2 test ($p < 0.001$).

5. Reads alignment to genome and gene expression quantification, alternative method

When they became available, we also used TopHat v. 1.2.0 (Trapnell, Pachter et al. 2009) and Cufflinks v.1.0.3 (Trapnell, Williams et al. 2010) to align the reads to the genome and assess gene expression levels respectively. This was achieved using the *C. remanei* genome reference sequence (WormBase release WS224) and annotation file (CR2, ENSEMBL, www.ensembl.org). By default, TopHat allows reads to map to several location in the reference sequence and will report up to 20 such alignments. To correct for multiple mapping effects, Cufflinks attributes a value proportional to the number of alignments of a read per read position. For example, if a read maps to five different locations, it will contribute 20% of a read at each position in terms of expression value. Both TopHat and Cufflinks were run with the default parameters except for the minimum intron size which was set at 40bp to accommodate for the average reported intron length of *C. elegans* (Cutter, Dey et al. 2009).

6. Assessment of phylogenetic conservation patterns

To assess patterns of conservation, I gathered data both from WormBase.org and Treefam.org, when available. Treefam gathers curated phylogenetic trees and has information on some *C. remanei* genes. WormBase provides among other things homology information for every gene, from several resources (WormBase-Compara, TreeFam, Inparanoid_7 and OMA). The categorization of genes in different conservation patterns was assessed for several different releases of WormBase for genes with strong expression differentials between sexes.

The total number of genes found to have no homologues in one or both of the selfing species varied from release to release - 116 in WS210, 110 in WS213, 105 in WS215, 114 in WS224, reflecting the variation in *C. remanei* genome annotations over time. While the number of genes varied, the 105 genes of the smallest set were found in all four lists. In order to facilitate further studies, I kept the list of 116 genes from release WS210 because the corresponding release was "frozen" (see Chapter 2). Briefly, because information pertaining to *C. elegans* and other species available on WormBase.org are constantly amended or corrected, the databases corresponding to each new realease of WormBase are available for a limited period of time with a dynamic interface, until the next release becomes available. At regular intervals, a selected release is saved as a datafreeze ("frozen") and made accessible as a dynamic website at all times, allowing access to all features expected from WormBase (as opposed to static text files).

III. Results

1. RNA-seq analysis

In order to identify genes whose expression levels display a significant difference between sexes in *C. remanei*, we generated two datasets from independent cDNA libraries, both composed of a male set and a female set. Both datasets retained substantial numbers of reads corresponding to the sequence of the adapter used to generate the cDNA libraries, confirmed by the low fraction of unique reads in each set (Table 2). 55% to 74% of the unique reads aligned to a region defined as a gene, based on the WS204 annotations of the *C. remanei* genome. Because of the broad definition we used of what a gene region is, we verified later for genes of interest the accuracy of the detection of expression levels by generating mapping profiles. 51.1% and 41.8 % of the alignments to the *C. remanei* genome from the first and second dataset respectively were perfect matches, 29.6% had 1 mismatch (Fig. 3). Only the reads matching once to the genome were considered for further analysis. The WS204 version of the annotations predicts 31,620 genes to exist in the *C. remanei* genome. 6,350 were detected overall in the first dataset, and 6,308 in the second.

2. Sex-biased expression

The expression level of 47% and 53 % of the genes in the first and second dataset, respectively, were found to be significantly sex-biased. Known male- and female-specific genes with high expression such as those encoding MSP (Major Sperm Protein) in males or yolk proteins (vitellogenins) in females, were recovered among the most extremely dimorphic genes. We arbitrarily set a fold-difference between male and female transcript abundance to 10, as we were mostly interested in genes displaying a strong sex-bias in their expression. 24.5% in the first dataset and 28.3% in the second were highly sex-biased (over 10-fold). Overall, more genes showed a female expression bias, but the genes male-biased in their expression tended to display a more ample bias (Table 2, also see Fig. 4).

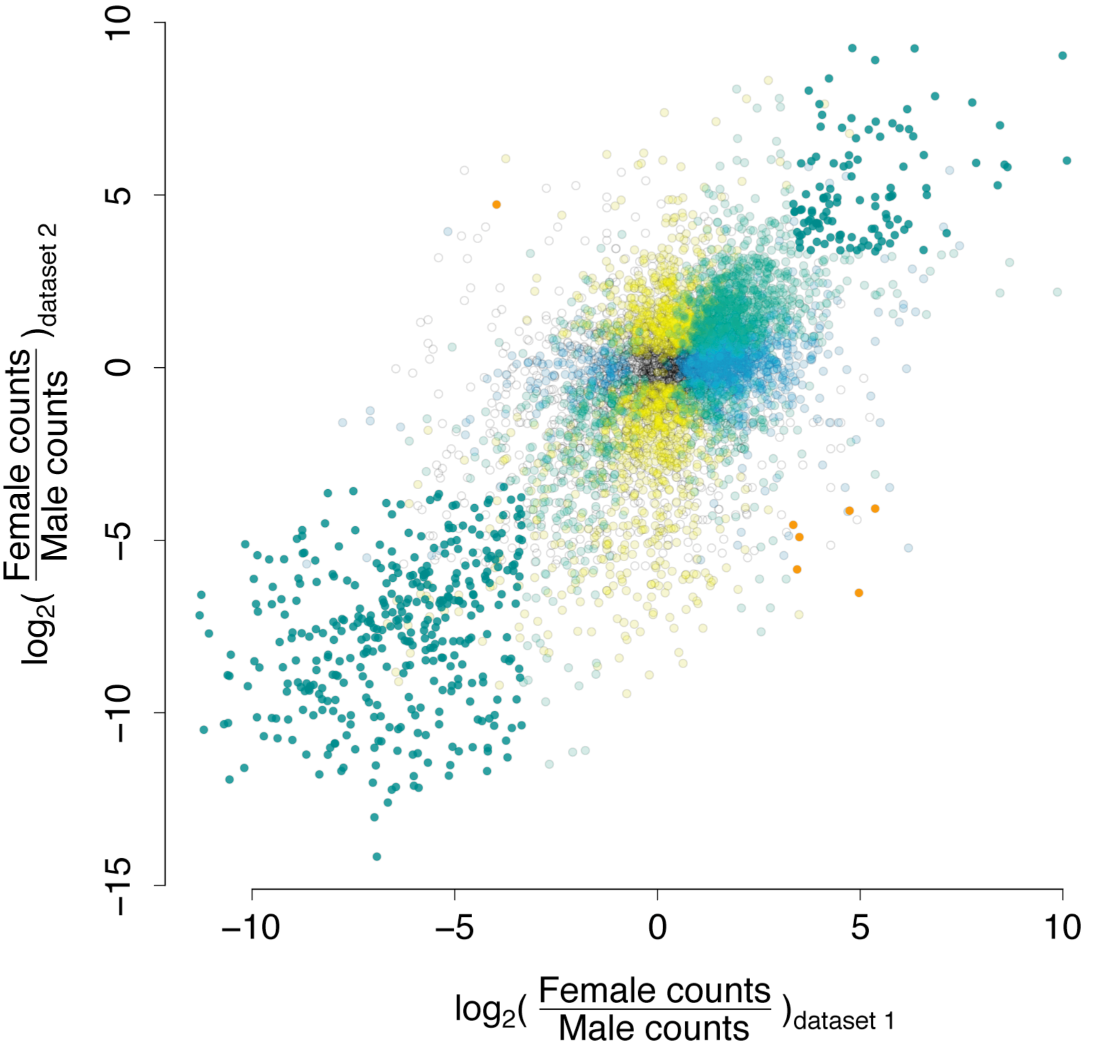


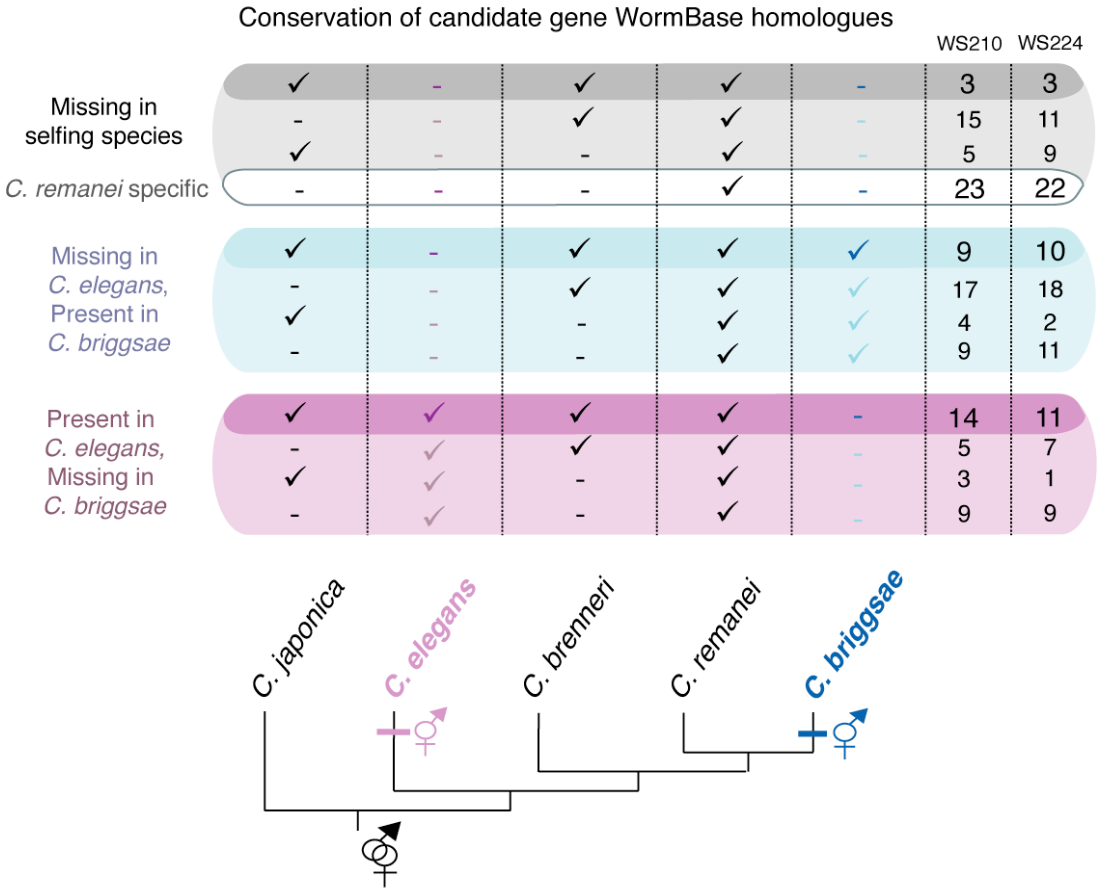
Figure 4. Datasets correlation using BLASTn for read mapping. The value of gene expression differentials between female and male cDNA libraries in each dataset are plotted against each other. Only genes detected in both datasets are represented. Genes whose expression is significantly sex-biased only in dataset 1 are represented by blue dots, only in dataset 2 by yellow dots, and in both by light green dots. Genes whose expression differential is above 10 in both datasets are represented by dark green dots when their bias was the same, and by orange dots if they displayed opposite bias between datasets.

3. Overlap datasets

Because the genes whose expression we were interested in detecting should have a consistent high sex-bias, I pooled the two datasets for further analysis. 2,001 genes were significantly sex-biased in both datasets, representing about a third of the detected *C. remanei* genes. The amplitude of the sex-bias between datasets for a given gene did vary, but 1,728 genes, representing 86.3 % of the genes whose expression was significantly sex-biased in both datasets, displayed the same bias and all but 7 of the 483 genes with a bias over 10 fold were concordant in both datasets (Fig. 4). The genes presenting an opposite high sex-bias in dataset 1 and 2 were not kept for subsequent analyses. As observed for each individual dataset, there were overall more genes whose expression was female-biased, while the male-biased genes tended to have a higher fold difference. In contrast, 112 and 364 genes respectively displayed highly (over 10 fold) female-biased and highly male-biased expression in both datasets (Fig. 4).

4. Conservation of genes with sex-biased expression in Caenorhabditis

Presence or absence of a WormBase homologue in *C. japonica, C. brenneri, C. briggsae* and *C. elegans* was verified for all of the 476 genes with strong sex-biased expression. Theoretically, there are 16 possible patterns of presence/absence of homologues. In practice, I was primarily interested in those for which homologues of candidate genes are missing in one or both of the selfing species (Fig. 5). Because the genomes of *C. brenneri* and *C. japonica* are far from being completely sequenced and annotated, I considered absence of a homologue in both these species as non-informative. 75.6% of the genes with high sex-biased expression had a WormBase homologue in all 5 species, and 116 genes had a WormBase homologue in neither *C. elegans* nor *C. briggsae* (Fig. 5).



	C. japonica	C. elegans	C. brenneri	C. remanei	C. briggsae	WS210	WS224
Missing in selfing species	✓	-	✓	✓	-	3	3
	-	-	✓	✓	-	15	11
	✓	-	-	✓	-	5	9
C. remanei specific	-	-	-	✓	-	23	22
Missing in *C. elegans*, Present in *C. briggsae*	✓	-	✓	✓	✓	9	10
	-	-	✓	✓	✓	17	18
	✓	-	-	✓	✓	4	2
	-	-	-	✓	✓	9	11
Present in *C. elegans*, Missing in *C. briggsae*	✓	✓	✓	✓	-	14	11
	-	✓	✓	✓	-	5	7
	✓	✓	-	✓	-	3	1
	-	✓	-	✓	-	9	9

Figure 5. Select patterns of conservation of candidate genes orthologues within Caenorhabditis. The numbers of genes falling in each category as of the latest datafreeze of WormBase available (WS210) as well as for the WS224 release are indicated.

Among those 116 genes, there was no significant over-representation of female- or male-biased expression when compared to the distribution of sex-bias in the whole datasets. In addition, while genes with sex-biased expression diplayed no significant difference when compared to the set of genes whose expression was detected, genes with highly sex-biased expression were more likely to be missing in one or both selfing species (Fig. 6, Fisher's exact test, p < 0.001). This difference was contributed to mostly by genes whose expression was highly male-biased among genes missing in one selfing species, and highly-female biased among genes missing in both (Fig. 6). Lastly, both genes with strong sex-biased expression, and among them those whose expression was highly female-biased differed significantly from detected genes among genes found to be *C. remanei* specific (Fig. 6, Fisher's exact test, $p < 0.05$ and $p < 0.001$ respectively).

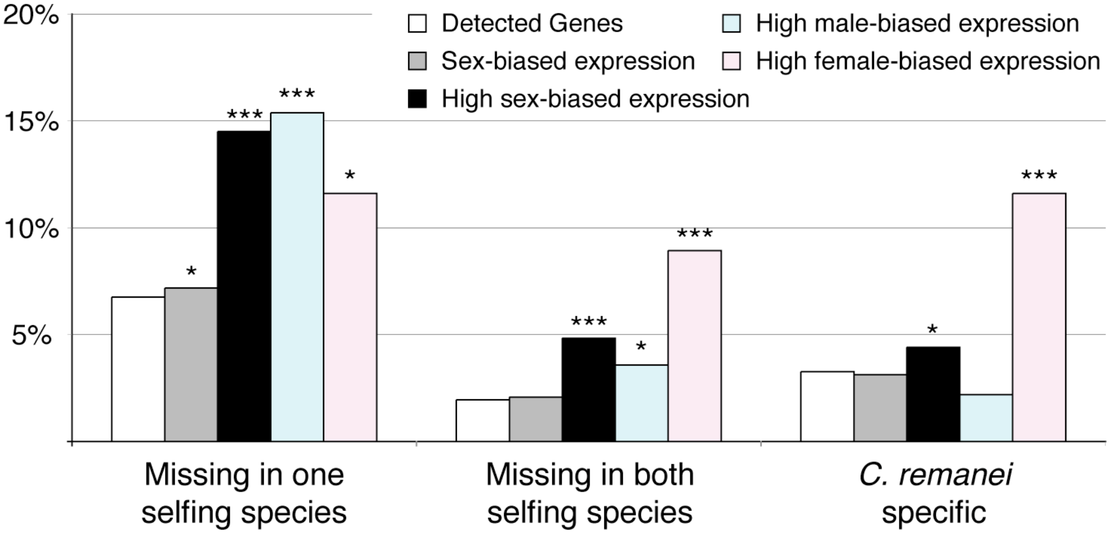


Figure 6. *C. remanei* genes with highly sex-biased expression are more likely to be missing in selfing species. Comparison of the patterns of conservation of *C. remanei* genes whose expression was detected (white), sex-biased (grey), highly sex-biased genes (black), higly male-biased (blue) or highly female-biased (pink) as of the WS224 Wormbase release. The fraction of genes which have a homologue in one or none of the selfing species are represented, as well as the fraction of genes found to be specific to *C. remanei*. Significance of difference to detected number of genes was assessed by Fisher's exact test. (* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$).

I listed GO terms (Ashburner, Ball et al. 2000) associated with the genes lacking homologues in both selfing species and whose expression was found to be strongly sex-biased in order to shed some light on their potential function. Genes with no homologues in *C. elegans* or *C. briggsae*, as well as genes whose expression levels were sex-biased or highly sex-biased, were less likely to be attributed a GO term when compared to genes whose expression were detected (Fig. 7, Fisher's exact test, p < 0.001 and p <0.01 respectively). The only GO terms that were significantly over-represented in genes without homologues in selfing species were pertaining to retro-transposon functions (Fig. 7, Fisher's exact test, p < 0.05). As a comparison GO terms indicating genes encoding membrane proteins were under-represented among genes with no homologues in selfing species, and in genes with sex-biased expression and highly sex-biased expression (Fig. 7, Fisher's exact test, p < 0.05, p < 0.001 and p < 0.05 respectively).

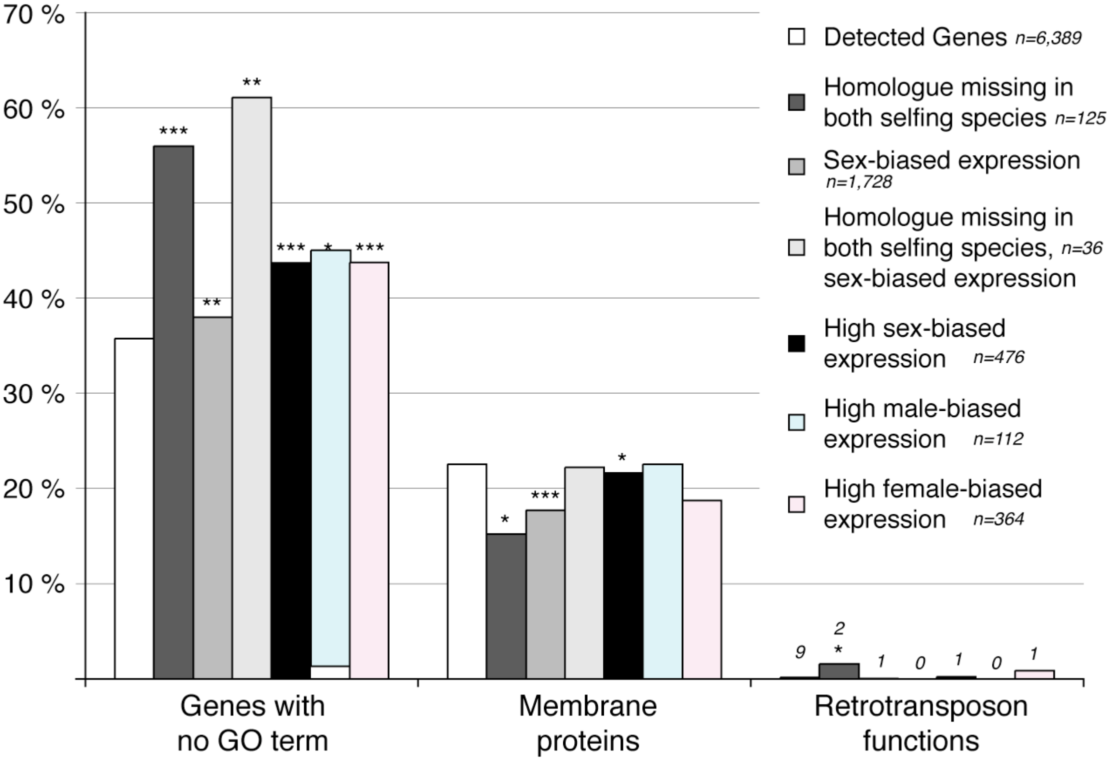


Figure 7. GO term analysis. The fraction of genes not associated to GO terms, predicted to be integral to membrane and whose GO terms pertained to retrotransposon functions are indicated for all *C. remanei* genes whose expression was detected (white), sex-biased (grey), highly sex-biased (black), highly female-biased (pink) and highly male-biased (blue), as well as for genes whose homologues were missing in both selfing species (dark grey) and sex-biased (light grey). Significance of difference to detected number of genes was assessed by Fisher's exact test (* p < 0.05; ** p < 0.01; *** p < 0.001). Number of genes are indicated in italic.

5. Conservation of sex-bias between *C. elegans* and *C. remanei*

Thorough microarray studies were previously performed in *C. elegans* to identify sex-biased genes (Reinke, Smith et al. 2000; Jiang, Ryu et al. 2001; Reinke, Gil et al. 2004). I compared the sex-bias of the *C. remanei* genes over 10-fold sex-biased and with a *C. elegans* homologue, to the sex-bias of their *C. elegans* homologues. Most of the genes whose expression bias was assessed in *C. elegans* displayed a similar sex-bias to their *C. remanei* homologues (Fig. 8) Out of 357 genes with a strong sex-biased expression in *C. remanei*, 20 genes (5.6%) showed a reversal of sex-bias and 25 (7%) were not identified as significantly differentially expressed in *C. elegans*. Among those with opposite sex-biased, 7 had GO terms pertaining to embryonic development, and 4 were predicted to encode membrane proteins (data not shown). However, the number of genes concerned was too small to run a statistical analysis.

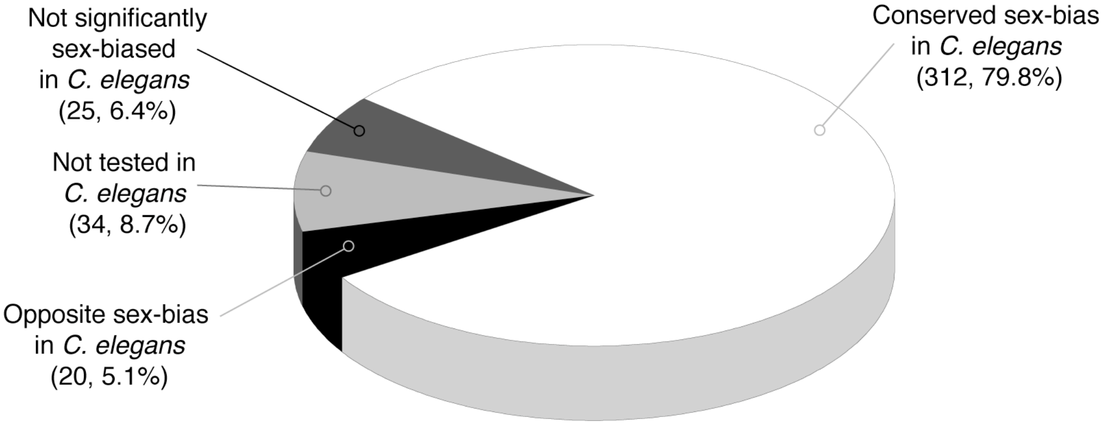


Figure 8. Conservation of sex-bias between *C. remanei* and *C. elegans*. Pie chart indicating the numbers of *C. remanei* genes with highly (over 10-fold) sex-biased expression whose *C. elegans* homologous genes either display the same (white) or the opposite (black) sex-bias, are not significantly sex-biased in *C. elegans* (dark grey), or have not been tested (grey). The expression bias of the *C. elegans* genes was determined by previous microarray studies (Reinke, Smith et al. 2000; Jiang, Ryu et al. 2001; Reinke, Gil et al. 2004).

6. Alternative analytical pipeline

In more recent years, numerous programs were created and published to study deep-sequencing data in a more automated way (Velvet (Zerbino and Birney 2008), Oases (Garg, Patel et al. 2011), AbySS (Simpson, Wong et al. 2009), Trinity (Grabherr, Haas et al. 2011), SHRiMP (Rumble, Lacroute et al. 2009), SOAP (Li, Li et al. 2008), Bowtie (Langmead, Trapnell et al. 2009), Cufflinks (Trapnell, Williams et al. 2010), TopHat (Trapnell, Pachter et al. 2009)). I chose to use TopHat and Cufflinks (see Methods) because of their speed and low CPU requirements to re-analyze our datasets and evaluate in retrospect the quality of our preliminary survey. I mapped all the reads of the two datasets independently to the *C. remanei* genome assembly using TopHat and quantified expression levels of *C. remanei* annotated genes using Cufflinks. Gene expression was estimated in FPKM (Fragments Per Kilobase of transcript per Million mapped reads, Trapnell, Williams et al. 2010). For consistency, I used the same statistical test to assess differential expression between male and female sets in both datasets.

Table 4. Comparison of analytical pipelines. The number of genes with highly sex-biased expression indicated reflect genes with consistent biases in both datasets.

	First analysis	TopHat / Cufflinks	In common (fraction of preliminary survey)
Number of genes detected	6,268	21,711	6,158 (98.2 %)
With: - sex-biased expression	2,001	3,847	1,019 (50.9 %)
- highly sex-biased expression	483	1,207	296 (61.3%)
- highly male-biased expression	364	852	217 (59.6 %)
- highly female-biased expression	112	355	79 (70.5 %)

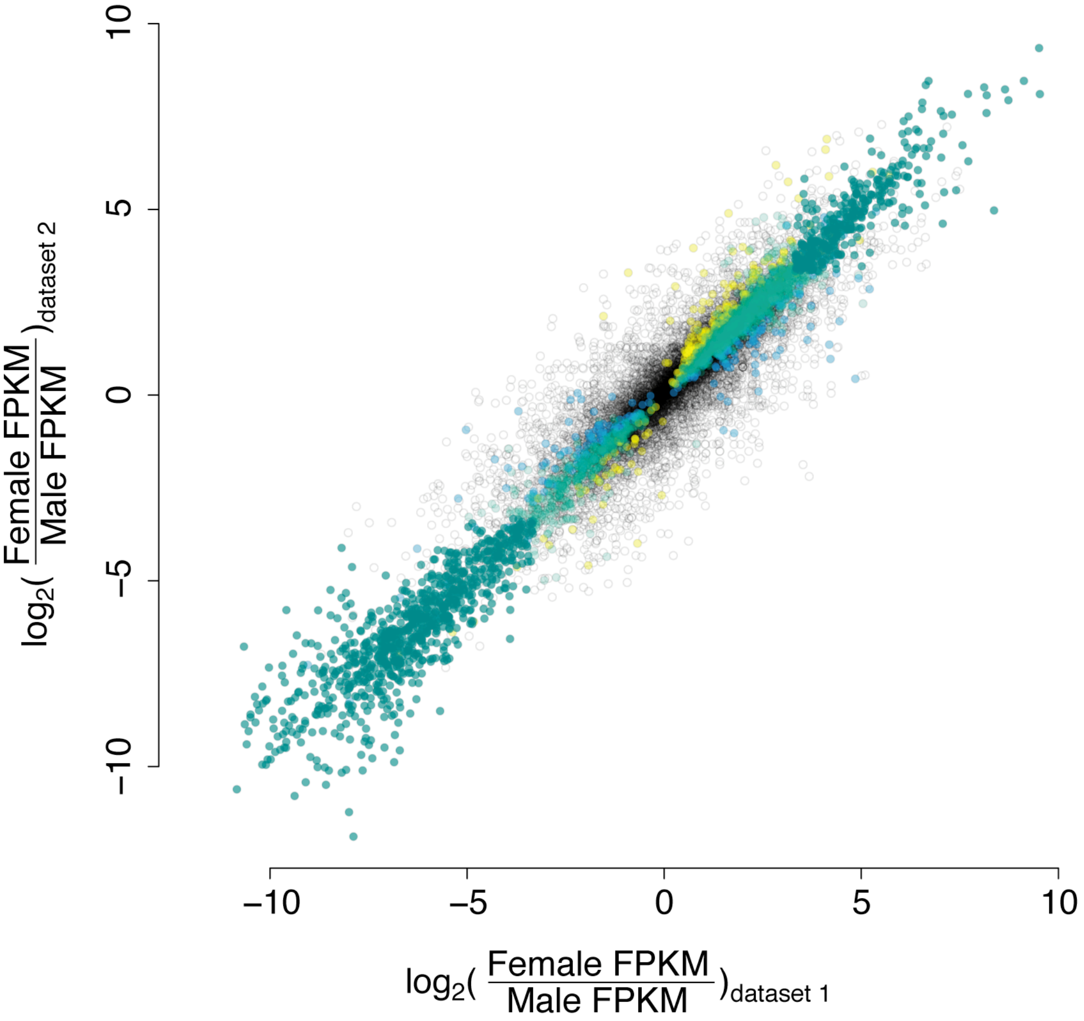


Figure 9. Datasets correlation using TopHat and Cufflinks for gene mapping. The value of gene expression differentials as assessed by TopHat analysis between female and male cDNA libraries in each dataset are plotted against each other. Genes whose expression is significantly sex-biased only in dataset 1 are represented by blue dots, only in dataset 2 by yellow dots, and in both by light green dots. Genes whose expression differential was above 10 in both datasets are represented by dark green dots when their bias was similar. No gene whose expression differential was above 10 fold in both datasets had a divergent sex-bias between dataset.

I found that the expression values attributed to genes in both 35bp- and 50bp-reads datasets were very well correlated as measured with the TopHat / Cufflinks pipeline (Fig. 9). In addition, the expression of a total of 21,711 genes was detected in common to both datasets, as opposed to that of 6,268 in our preliminary pipeline. The expression of most of the genes detected in the first analysis was also detected by the

TopHat / Cufflinks analysis (Table 4). 3,847 of the 21,711 genes displayed a significant expression differential between males and females, 1,892 of which over 10-fold (Table 4). 1,019 genes had a significant sex-biased expression in both analysis, 296 of these over 10-fold, representing respectively about 51% and 61% of the number of genes in the corresponding categories in the preliminary survey (Table 4). Among the genes whose expression differential was over 10-fold in both analyses, the direction of the sex-bias was the same in both analyses.

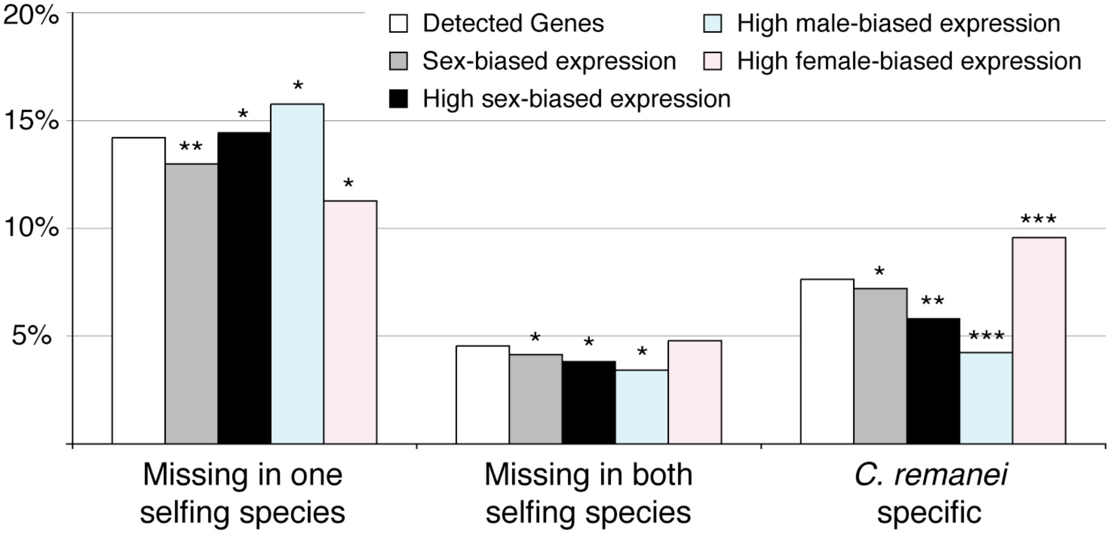


Figure 10. Conservation pattern of genes whose expression was detected through Tophat / Cufflinks analysis. Comparison of the patterns of conservation of *C. remanei* genes whose expression was detected (white), sex-biased (grey), highly sex-biased genes (black), higly male-biased (blue) or highly female-biased (pink) as of the WS224 Wormbase release. The fraction of genes which have a homologue in one or none of the selfing species are represented, as well as the fraction of genes found to be specific to *C. remanei*. Significance of difference to detected number of genes was assessed by Fisher's exact test (* p < 0.05; ** p < 0.01; *** p < 0.001).

To be consistent and be able to compare further both analytical pipelines, I also determined the conservation patterns of the genes in the list yielded by the TopHat / Cufflinks analysis. Among those genes, there was a slightly significant over-representation of highly sex-biased expression in genes missing in one selfing species (Fig. 10, Fisher's exact test, $p < 0.05$). In addition, genes with sex-biased expression were found to be less likely to be missing in one or both of the selfing species when compared to the set of genes whose expression was detected (Fig. 10, Fisher's exact test, $p < 0.01$ and $p < 0.05$ respectively). Lastly, genes whose expression was highly sex-biased differed significantly from detected genes among genes found to be *C. remanei* specific, and this under-representation was contributed to mostly by genes with high male-biased expression (Fig. 10, Fisher's exact test, $p < 0.05$ and $p < 0.001$ respectively).

I also listed GO terms associated to genes whose expression were detected with the TopHat / Cufflinks analysis and found similar trends to what was observed previously. Genes with no homologues in *C. elegans* or *C. briggsae* were less likely to be attributed a GO term when compared to genes whose expression were detected (Fig. 11, Fisher's exact test, $p < 0.01$). GO terms pertaining to retro-transposon functions were over-represented in genes without homologues in either selfing species, as well as in genes with sex-biased expression (Fig. 11, Fisher's exact test, $p < 0.001$), and under-represented in those with high sex-biased expression (Fig. 11, Fisher's exact test, $p < 0.01$) and high female-biased expression (Fig. 11, Fisher's exact test, $p < 0.01$). Genes with sex-biased expression were also less likely to be attributed GO terms indicating genes encoding membrane proteins (Fig. 11, Fisher's exact test, $p < 0.001$), as well as genes with highly sex-biased expression (Fig. 11, Fisher's exact test, $p < 0.001$), contributed to by genes with highly female-biased expression (Fig. 11, Fisher's exact test, $p < 0.001$).

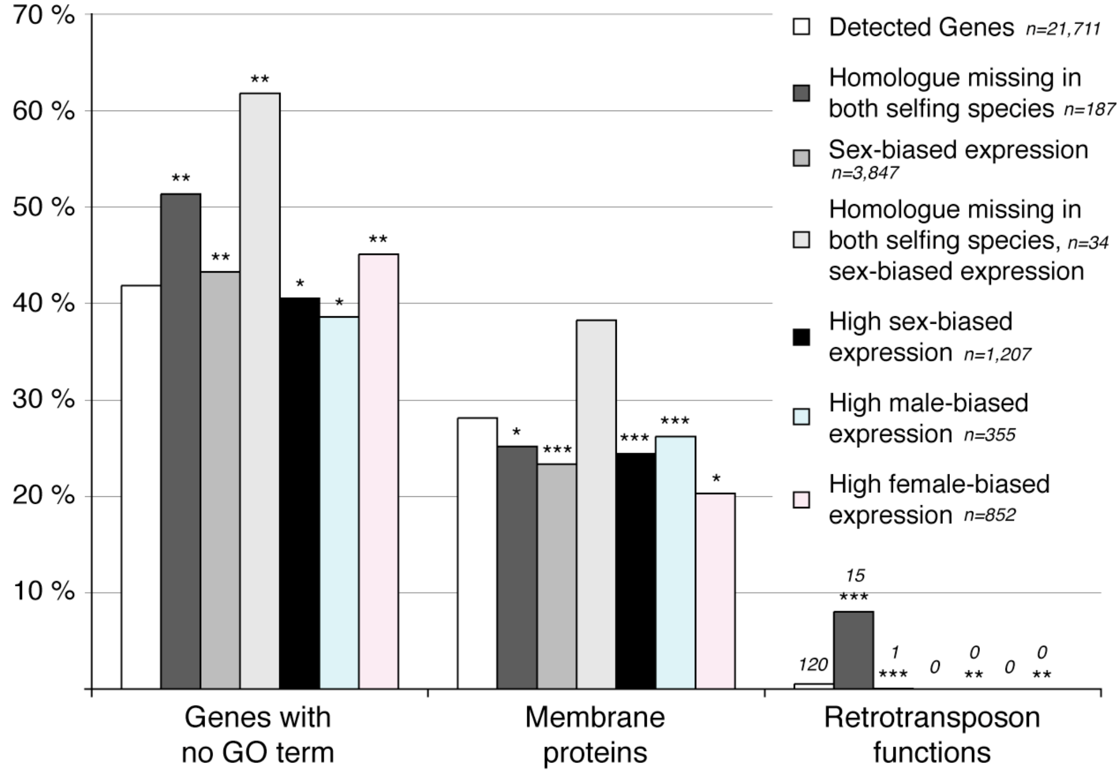


Figure 11. GO term analysis of genes whose expression was detected through Tophat / Cufflinks analysis. The fraction of genes not associated to GO terms, predicted to be integral to membrane and whose GO terms pertained to retrotransposon functions are indicated for all *C. remanei* genes whose expression was detected (white), sex-biased (grey), highly sex-biased (black), highly female-biased (pink) and highly male-biased (blue), as well as for genes whose homologues were missing in both selfing species (dark grey) and sex-biased (light grey). Significance of difference to detected number of genes was assessed by Fisher's exact test (* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$). Numbers of genes are indicated in italic.

IV. Discussion

1. Sex-biased expression in *C. remanei*

The reported proportions of the genome displaying sex-biased expression vary from species to species and tissues to tissues, ranging from as little as 4% in mouse gonads (Rinn, Rozowsky et al. 2004) to close to 90% in *D. melanogaster* whole flies (Ayroles, Carbone et al. 2009). These values depend highly on the method used to measure gene expression, typically micro-array in the case of sex-biased expression, as well as on the statistical threshold chosen to define sex-bias. In this study, I used RNA-seq to measure gene expression level, and defined sex-biased expression as a deviation from the expected male:female expression ratio, as opposed to chosing an arbitrary 2-fold difference threshold. I observed that 17.7 % to 27.8% of the genes whose expression was detected were significantly differentially expressed between sexes, depending on the way the data was analyzed. While these numbers are within the very wide expected range, they differ enough to call for a deeper survey of sex-biased expression in order to assess more accurately the extent of sex-biased expression in *C. remanei*. Such a survey is presented in Chapter 3.

Despite their methodological differences, the patterns of *C. remanei* sex-biased expression observed with the two analytical approaches (first analysis and TopHat / Cufflinks) were broadly consistent. In addition, sex-biased expression in other species exhibit similar trends. More genes tended to have a female-biased expression than male-biased expression, as in *C. elegans* (Reinke, Gil et al. 2004), several species of Drosophila (Parisi, Nuttall et al. 2003; Ranz, Castillo-Davis et al. 2003; Zhang, Sturgill et al. 2007; Ayroles, Carbone et al. 2009; Jiang and Machado 2009), mosquitoes (Hahn and Lanzaro 2005), sticklebacks (Leder, Cano et al. 2010), and mice (Yang, Schadt et al. 2006). Genes with male-biased expression on the other hand showed a wider magnitude of expression levels (Fig. 4 and Fig. 9) as is also seen in several species of Drosophila (Parisi, Nuttall et al. 2003; Zhang, Sturgill et al. 2007; Ayroles, Carbone et al. 2009) or in pigs (Perez-Enciso, Ferraz et al. 2009). Interestingly, this pattern is exactly the opposite in birds, where there are more genes with male-biased expression (Mank, Hultin-Rosenberg et al. 2007; Naurin, Hansson et al. 2011), and in Xenopus where fewer genes with female-biased expression display a greater spread in their expression values (Malone, Hawkins et al. 2006). Both these cases represent the ZZ/ZW sex determination system. Dosage compensation has been proven to not be effective in several species with such a system (Ellegren, Hultin-Rosenberg et al. 2007; Itoh, Melamed et al. 2007; Mank and Ellegren 2009; Zha, Xia et al. 2009; Itoh, Replogle et al. 2010; Wolf and Bryk 2011), suggesting a fundamental difference between XX/XY and ZZ/ZW systems in the way sexual dimorphism is attained, and subsequently in the expected corresponding sex-biased expression patterns.

2. Sex-biased expression in the Caenorhabditis genus

The only other species in the Caenorhabditis genus for which sex-biased expression has been measured is *C. elegans,* in which microarray studies identified genes expressed in the germline and soma of hermaphrodites and males (Reinke, Smith et al. 2000; Jiang, Ryu et al. 2001; Reinke, Gil et al. 2004). Because those studies were conducted independently, and used a different experimental protocol (microarrays) the expression values of *C. elegans* orthologues of *C. remanei* genes whose expression were detected in our analysis were not directly comparable. However, the direction of sex-bias – that is, male or female / hermaphrodite – can provide an indication for the overall trend of conservation of sex-bias. Sex-biased expression seems to be mostly conserved in other species, although reversal of sex-biased expression between species has been shown to occur in Drosophila (Ranz, Castillo-Davis et al. 2003; Zhang, Sturgill et al. 2007; Jiang and Machado 2009) at varying rates depending on the threshold chosen to define sex-bias. For most orthologous *C.elegans-C.remanei* pairs the direction of sex-bias was conserved, but a small fraction of genes with a strong sex-biased expression in *C. remanei* showed either a reversal of sex-bias or were not identified as significantly differentially expressed in *C. elegans*. A more directly comparable, deeper assessment of expression level in *C. elegans* and *C. remanei* would be required in order to define accurately the extent to which expression differential and transcriptional variance is conserved between these species.

3. Evolutionary dynamics of genes with highly-sex-biased expression

In *C. elegans*, genes with high male-biased expression evolve faster than other genes (Cutter and Ward 2005), as do they in Drosophila (Ranz, Castillo-Davis et al. 2003; Clark, Eisen et al. 2007) and mammals (Schug and Overton 1997; Torgerson, Kulathinal et al. 2002; Good and Nachman 2005; Khaitovich, Hellmann et al. 2005). As a result, *C. elegans* genes with male-biased expression are less likely to have *C. briggsae* orthologues than any other category of genes (Cutter and Ward 2005). These genes are also more likely to be species-specific paralogues (Cutter and Ward 2005), indicating a greater rate of loss, duplicaton and divergence than other genes.

In line with these observations, *C. remanei* genes with strongly sex-biased expression were more likely to be missing in one or both of the selfing species (*C. elegans* and *C. briggsae),* and among those missing in one selfing species, genes with a high male-biased expression were over-represented (Fig. 6). Interestingly, among *C. remanei* specific genes, only genes with a strong female-biased expression were over-represented. These trends, however, did not hold when analyzing the data with TopHat and Cufflinks (Fig. 10). While the lack of conservation of genes with highly-sex biased expression as a whole was not expected, genes with a strong male-biased expression differential were expected to behave differently than other genes. An artefact due to both the low-depth of our survey and the inadequacy of applying the same statistical test to define expression differential could explain the discrepancy between both analytical pipelines. A broader analysis would be required to decipher which of these patterns is true.

C. remanei genes deemed as missing in one or both selfing species were conserved in at least one other Caenorhabditis species. Therefore, they were lost in at least one of the selfing species. An important point to note here is that "losses" cover several different evolutionary events: the orthologous gene might still exist but be so diverged that it is not recognized as such. Other options include pseudogene formation, disruption of the coding sequence by insertion of a transposable element, complete or partial excision from the genome, or disruption of the coding sequence through chromosomal rearrangements, any of those resulting from relaxed selection pressure on these genes. Incomplete genome annotations may also cause apparent loss. For example, in the *C. elegans* genome annotations, and to a lesser extent in that of *C. briggsae*, gene predictions corresponding to transposable elements are systematically removed. Such an annotation effort has not yet been extended to the other Caenorhabditis gene predictions and as a result there are more gene predictions corresponding to transposable elements in those species compared to *C. elegans,* irrespective of the actual lower transposable element content of *C. elegans* (CGT, pers. observation). The over-representation of GO terms pertaining to retro-transposon functions within *C. remanei* genes missing in both selfing species could then be in part explained by the discrepancy in annotations. Interestingly, the over-representation of these GO terms does not extend to *C. remanei* genes missing in both selfing species whose expression was highly sex-biased. More comprehensive studies of transposable element genome content would shed some light on these observations but were beyond the scope of this study and not followed up on.

4. Accuracy of detection of gene expression method

4.1. Limits of the study system

The different assumptions and limits under which this study was undertaken were in part due to the current status of the *C. remanei* genome, and in part to the low depth and low quality of both RNA-seq datasets received form the WUGSC.

The reference sequence for the *C. remanei* genome is a set of 3,670 supercontigs, whose lengths vary from 2 kb to 4.5 Mb, and which contains residual heterozygosity (Barriere, Yang et al. 2009). In *C. elegans* the average gene length is 1.9 kb (Lynch 2007), suggesting that a fraction of *C. remanei* real genes will not be detectable as such by gene prediction softwares because of the reference sequence fragmentation. Subsequently the current number of *C. remanei* gene predictions is likely to be misestimated. As a result the expression of some genes should not be detectable through analytical pipelines requiring the use of a genome reference sequence.

The cDNA libraries sequenced in this study represented genes expressed during the L4 and adult stages, in the soma or in the germline of *C. remanei* nematodes as well as in embryos, eggs and sperm since the collected animals were at the adult stage. As a result, we were not expecting to be able to define an expression level measure for every single gene predicted in the *C. remanei* genome. The goal of this work was to compare male to female expression using the same reference sequence and identify a few select candidate genes. These limit constitute merely an obstacle to the number of genes whose expression can be detected rather than one to the accuracy of our undertaking.

4.2. Limits of the analytical pipeline for detection of gene expression

RNA-seq is a new technique that has proven to have greater gene expression detection power than now traditional micro-arrays (see reviews by Wang, Gerstein et al. 2009; Oshlack, Robinson et al. 2010; Salzberg 2010). However the time at which this project was started coincided with the very beginning of RNA-seq and the profuse amount of very convenient tools and softwares that are nowadays publicly available was rather limited then. The analytical pipeline used in this study to assign reads to genes in order to be able to compare male to female expression level was designed with known limitations and allowed the actual use of only a fraction of the generated datasets (Table 2) and alignments (Table 3). Namely, splice junctions were kept in mind, but the alignments that would have covered them in the first dataset (35bp reads) were filtered out when setting an alignment score to 30; the gain in biological relevance outweighed the amount of background from low score alignments. In addition, “gene windows” were defined so as to be able to take into account the reads matching to either the 5’ or the 3’ UTRs, as well as to compensate for the imprecision of the genome and include potential alternative isoforms. Any alignment to the genome reference sequence located outside of such gene windows were not taken into account for further analysis. Lastly, while only unique reads were mapped to the reference sequence, multiple hits of a same read to several location in the reference sequence were kept and all counted equally so long as the alignments were within the set alignment parameters.

The short length of the reads is a limit to RNA-seq, as it is impossible to attribute accurately a short sequence to either perfect match it might have in the reference sequence, and different softwares deal with this issue in different ways (Oshlack, Robinson et al. 2010). While this might introduce a bias in defining exactly the expression level of genes sharing similar sequences, the increase in irrelevant alignments is limited to the shared stretch of sequence. In this work, the expression level of a gene is defined by all the alignments to the corresponding gene window. Since alignments to a sequence shared by two genes count equally towards their expression level, only alignments to gene-specific sequence allow discrimination among them. Multiple alignments might potentially lead to a misestimation of sex-biased expression, either by amplifying or by flooding the expression signal observed in both sexes.

4.3. Performance in the light of a newer better faster method.

The number of genes whose expression was detected was much higher using TopHat and Cufflinks (Table 4), and the expression level of genes as detected in the two datasets (35 vs. 50bp) were found to be better correlated between datasets than using the first method (compare Fig. 4 to Fig. 9). The massive difference in the number of genes was detected and how well their expression levels between datasets was correlated can be attributed to fundamental differences in the way reads are attributed to genes by TopHat and how expression levels are determined by Cufflinks compared to our method. TopHat was designed to identify splice junctions using short reads, by mapping reads to the reference sequence a first time to identify potential exons, and a second time to confirm splice junctions. Cufflinks defines expression levels based on the assumption that the distribution of the reads reflects the distribution of the fragments of cDNA that have been sequenced, and corrects for this distribution by assuming an approximate gaussian distribution (Trapnell, Williams et al. 2010). In addition, Cufflinks corrects for potential sequence-specific bias introduced at the cDNA library stage (Roberts, Trapnell et al. 2011). Lastly, the way Cufflinks handles multi-mapping reads in attributing expression values to genes (see methods) decreases the bias in measuring expression level of genes sharing sequences. In short, using TopHat and Cufflinks led to a more accurate mapping of the reads from the low-depth RNA-seq survey, which was however still limited by the status of the *C. remanei* genome sequence.

Most of the genes whose expression were detected by the first method were also detected using TopHat and Cufflinks (Table 4). The way the expression level was quantified by both methods is different and comparing expression value for expression value would be potentially misleading, as could be the expression differentials between males and females. The first method allowed the unambiguous identification of 1,728 genes with a sex-biased expression, representing 27.6% of the genes whose expression was detected. In comparison, the same statistical analysis applied to the TopHat / Cufflinks analysis yielded 17.7% of sex-biased expression. In addition, only half of the genes whose expression was found to be sex-biased in the first analysis were also identified as having sex-biased expression in the second analysis. Along the same lines, 62% of the genes found to have a high sex-biased expression in the first analysis also had a high sex-biased expression in the second one, with no particular over-representation of high female- or male-bias. Sex-biased expression was defined by a χ^2 test and Bonferroni correction, thought to be rather stringent, based on the number of genes whose expression was detected. It is possible that such a statistical analysis is too stringent when applied to expression level values already corrected for the different biases inherent to RNA-seq data as they are with Cufflinks. This is supported by the fact that later RNA-seq experiment at higher sequencing depth found 49% of the *C. remanei* transcriptome to be significantly sex-biased (see Chapter 3, Thomas, Li et al. In prep.).

In conclusion, re-analyzing the low-depth RNA-seq data with TopHat and Cufflinks showed in retrospect that the two technical replicates (35 vs 50bp) were well correlated. Further, it confirmed the need for curating carefully the set of candidate genes identified to confirm sex-biased expression.

Conclusions

All together, this study revealed trends of sex-biased expression pattern in *C. remanei* and shed some light on the evolutionary dynamics of genes with a high expression differential between sexes. The data obtained in this preliminary survey confirmed that, as observed in other species, the expression of more genes is biased towards the monogametic sex while the expression pattern in the heterogametic sex is more spread. It also indicated that genes with a strong expression differential between sexes in *C. remanei* tended to be missing orthologues in one or both of the selfing species. This is consistent with the presence of a selfing syndrome in *C. elegans* and *C. briggsae*. However, the functions of the genes found to be missing have yet to be determined, nor have they been directly implicated with the decline in mating efficiency in selfing species. The low-depth of this survey however, together with the inconsistencies observed between the analytical pipelines call for a more thorough investigation of sex-biased expression in the genus to be able to paint an accurate landscape of the transcriptomic consequences of the emergence of a new mating system.

Chapter 2: Characterization of evolutionary labile male genes in Caenorhabditis

I. Summary

Within nematodes of the Caenorhabditis genus, two mating systems co-exist. Some species are androdioecious, their populations being mostly composed of hermaphrodites and rare males, while most other extant species are gonochoristic (i.e. producing true females that must mate to reproduce). Selfing arose independently from distinct gonochoristic ancestors in two species, *C. elegans* and *C. briggsae* (Cho, Jin et al. 2004; Kiontke, Gavin et al. 2004; Nayak, Goree et al. 2005; Hill, de Carvalho et al. 2006; Guo, Lang et al. 2009; Kiontke, Félix et al. in revision). Several lines of evidence suggest that *C. elegans* and *C. briggsae* suffer from gradually losing traits related to mating (Hodgkin and Doniach 1997; LaMunyon and Ward 1999; Chasnov and Chow 2002; Lipton, Kleemann et al. 2004; Geldziler, Chatterjee et al. 2006; Chasnov, So et al. 2007; Garcia, LeBoeuf et al. 2007; Kleeman and Basolo 2007; Palopoli, Rockman et al. 2008), hallmark of a selfing syndrome (Cutter 2008). In addition, estimates of genome sizes of Caenorhabditis species by flow cytometry indicate that androdioecious species have markedly smaller genomes than gonochoristic species (Thomas, Li et al. In prep.). Because there is a strong correlation between genome shrinkage and loss of ancestral traits related to mating, we hypothesized that genes coding for proteins involved in mating processes would be preferentially lost as self-fertilization emerged and stabilized in *C. elegans* and *C. briggsae*.

To address whether such genes existed, I first used RNA-seq to discover genes with highly sex-biased expression in a gonochoristic species, *C. remanei.* Indeed, specific mating processes are limited to one sex or the other, and the genes underlying such traits are expected to be strongly sex-biased in their expression (see Chapter 1). In addition, such genes, candidates for conferring traits related to mating, are more likely to be missing in one or both of *C. elegans* and *C. briggsae* than expected. In this chapter, I further explore the consequences of self-fertility by asking whether genes with highly sex-biased expression in *C. remanei* and lost in selfing species are indeed related to mating processes.

In this study, I curated 116 genes whose expression was found previously to be highly sex-biased in *C. remanei*, and which were defined through a preliminary screen as missing homologues in *C. elegans*, *C. briggsae* or both. I examined their conservation patterns and selected those 15 that best fit our working hypothesis for further characterization. All of these genes have strongly male-biased expression, consistent with faster rates of evolution of male-biased genes in the Caenorhabditis genus (Cutter and Ward 2005). Their predicted gene structures and expression bias were confirmed. In addition, their temporal expression patterns were determined. Interestingly I discovered through this work a novel gene family characterized by coding for a novel conserved repeated sequences. I also was able to identify a novel putative cis-regulatory element strongly associated with genes male-biased in their expression. Through this work, I show that some genes with a strong male-biased expression in *C. remanei*, while expected to have diverged significantly in the Caenorhabditis genus (Cutter and Ward 2005), are conserved in other gonochoristic species and lost specifically in selfing species, and that their role is related to mating processes.

II. Material and Methods

1. Nematode culture and strains

Caenorhabditis strains (Table 1) were maintained on NGM agar plates (3 g/L NaCl, 2.5 g/L peptone, 22 g/L agar, 5 mg/L cholesterol, 2 mg/L uracil, 0.1 M $CaCl_2$, 0.1 M $MgSO_4$ and 25 mM KH_2PO_4) according to standard *C. elegans* methods (Wood 1988), using 2.2% agar to discourage burrowing. Unless otherwise noted, the NGM plates were seeded with OP-50 bacteria and worm stocks kept at 20˚C. To maintain outbreeding as much as possible and avoid inbreeding depression, large worm populations were maintained by transferring large chunks from two independent plates per passage.

Table 1. Strains of Caenorhabditis nematodes. The *C. elegans unc-119*(ed3) strain was kindly shared by the Hamza lab (University of Maryland), JU1422 was kindly shared by Marie-Anne Felix (Institut Jacques Monod, Universite Paris 7), the others were obtained from the Caenorhabditis Genetics Center (University of Minnesota).

Species	Strain	Allele / Comment
C. japonica	DF5081	Inbred, derived from DF5080, used for genome sequence
C. elegans	N2	WT
C. elegans	CB4108	*fog-2*(q71)
C. elegans	HT1593	*unc-119*(ed3)
C. brenneri	PB2801	Inbred, derived from LKC28, used for genome sequence
C. remanei	EM464	WT
C. remanei	PB4641	Inbred, derived from EM464, used for genome sequence
C. briggsae	AF16	WT
C. sp. 9	JU1422	Inbred, derived from JU1325

2. DNA extraction

Worms were washed off of plates in M9 buffer (86 mM NaCl, 42 mM Na_2HPO_4, 22 mM KH_2PO_4, 1mM $MgSO_4$), washed 2 to 3 times in M9 buffer to get rid of residual bacteria, then once in disruption buffer (200mM NaCl, 50mM EDTA, 100mM Tris pH 8.5). The worms were then resuspended in disruption buffer supplemented with SDS (200mM NaCl, 50mM EDTA, 100mM Tris pH 8.5, 0.5% SDS), and frozen at -80˚C. The samples were then thawed, proteinase K added at a final 150µg/ml concentration, and incubated at 65˚C until the worms were dissolved. The DNA was then purified using phenol/chloroform extraction followed by ethanol precipitation, and resuspended in TE 10:1.

3. Long-range PCR

Primers were designed using gene predictions and data available on WormBase. Manufacturer's instructions from the DyNAzyme EXT PCR kit (Finnzymes) were followed to amplify long DNA fragments from genomic DNA.

4. DNA Sequencing

Primers were designed using gene predictions and data available on WormBase. BigDye v3.1 Cycle Sequencing chemistry (Applied Biosystems) was used for DNA sequencing according to manufacturer's protocol. Sequencing was performed on ABI 3750 machine according to standard protocol, and trace files were examined using Geneious (Drummond AJ 2011).

5. RNA extraction

Males and females at the L4 stage, 24 hours (Young Adults) and 72 hours (Adults) later as well as synchronized mixed sex populations of eggs, larval stages L1, L2, L3, L4, and Young Adults were collected independently five to seven times each. Worms were synchronized by NaOH / bleach treatment and washed off the plates in M9 buffer at various developmental stages, or individually picked in M9 buffer for males and females preparations. Eggs were isolated according to standard protocol (Wood 1988). Worms or eggs were washed three times in RNase free water and resuspended in 50 µL of RNase free water. 250µL of TRI-Reagent (Molecular Research Center) were added and the samples were frozen at -80˚C. The samples were thawed, pelleted and lyzed using a plastic pestle. RNA was purified using phenol/chloroform extraction and subsequent isopropanol precipitation. All of the samples were treated with DNaseI (New England Biolabs). Phenol/chloroform extraction followed by isopropanol precipitation was performed a second time and the RNA was resuspended in RNase free water.

6. Real-Time Quantitative RT-PCR

For each RNA sample but two first strand cDNA was synthesized using 1 µg of total RNA using Superscript III (Invitrogen) in 50 µl according to manufacturer's instructions, in a 96-well plate. The N_0 values of the two samples for which less than 1 µg of total RNA was available were later corrected for input quantity. 2µl cDNA were used as template with the Light Cycler 480 SYBR Green I kit (Roche) according to manufacturer's instructions. Primers spanning at least one exon-exon junction were designed so that the amplicons sizes were between 161 and 220 bp (Appendix 2).

Negative control reactions with no template for each primer pair were performed, as well as positive control reactions with four conserved genes whose biases were high in *C. remanei* and conserved in *C. elegans,* and three unbiased constitutively expressed conserved genes. One 96-well plate containing all the cDNAs was run by gene. Data were collected from a Roche Light Cycler 480 machine using manufacturer's software, and analyzed using the program LinRegPCR (Ramakers, Ruijter et al. 2003; Ruijter, Ramakers et al. 2009).

7. RNA Interference

cDNAs specific of genes of interest were amplified using primers with T7-polymerase priming sites (Appendix 2) from total RNA, and used as template. Double-stranded RNA was synthetized using the Megascript T7 kit (Ambion) according to manufacturer's protocol. dsRNA was purified according to the phenol:chloroform / isopropanol clean-up protocol included in the kit and resuspended in RNAse-free TE (100mM Tris, pH7.5, 10mM EDTA). dsRNA was injected at varying concentrations in the gut of young adult females or hermaphrodites. Injected mothers were moved to fresh plates 6 hours post-injection, and every 12 hours subsequently, along with males in the case of females so as to not limit sperm availability.

8. Multi-site Gateway Cloning

Transcriptional reporters were built using the GateWay technology (Invitrogen). Genomic DNA directly upstream of the start site of genes of interests as well as directly downstream of their stop codons were PCR amplified using primers carrying attB2F, attB3R, attB1R or attB4F sites (Fig. 1a, also see Appendix 2). The location of the primers were chosen so that the smallest of either about 1 kb or the intergenic region between the start or stop codon of the gene of interest and that of the closest upstream or downstream gene, including 5' and 3' UTRs of genes of interest, were amplified. These PCR fragments corresponding to 5' or 3' intergenic regions were cloned respectively into pDONR P4-P1R or pDONR P2R-P3 in a BP reaction according to manufacturer's instructions to generate 5' and 3' entry clones. pCM1.35, a middle entry clone containing *Ce-his-58,* the sequence of histone H2B, fused in frame to that of green fluorescent protein was obtained from the Seydoux lab (Johns Hopkins Medical School, Merrit and Seydoux 2010).

The destination vector pBCN30 was kindly shared by Jennifer Semple (Dupuy lab) and was built from pDEST R3-R4 by adding the sequences of a "worm-friendly" mCherry under the control of the *Ce-myo-2* promoter as well as those conferring resistance to puromycin and G418 under the control of *Ce-rpl-28* promoter (Fig. 1b). This vector was used successfully to drive mCherry expression in the pharynx of *C. elegans, C. briggsae* and *C. remanei* (Semple et al 2010). The entry vectors and pBCN30 were used in a LR reaction to generate pCGT vectors according to manufacturer's protocol (Fig. 1b). pCGT vectors were transformed into DH5α cells (New England Biolabs). Successful LR reactions were identified by colony PCR from 12-64 independent colonies depending on the pCGT transformation. Positive colonies were grown in 5mL overnight cultures, their plasmid DNA extracted and purified using the QIAGEN mini-prep kit according to manufacturer's standard protocol. The sequences of pCGT vectors were verified by restriction digest and sequencing before use.

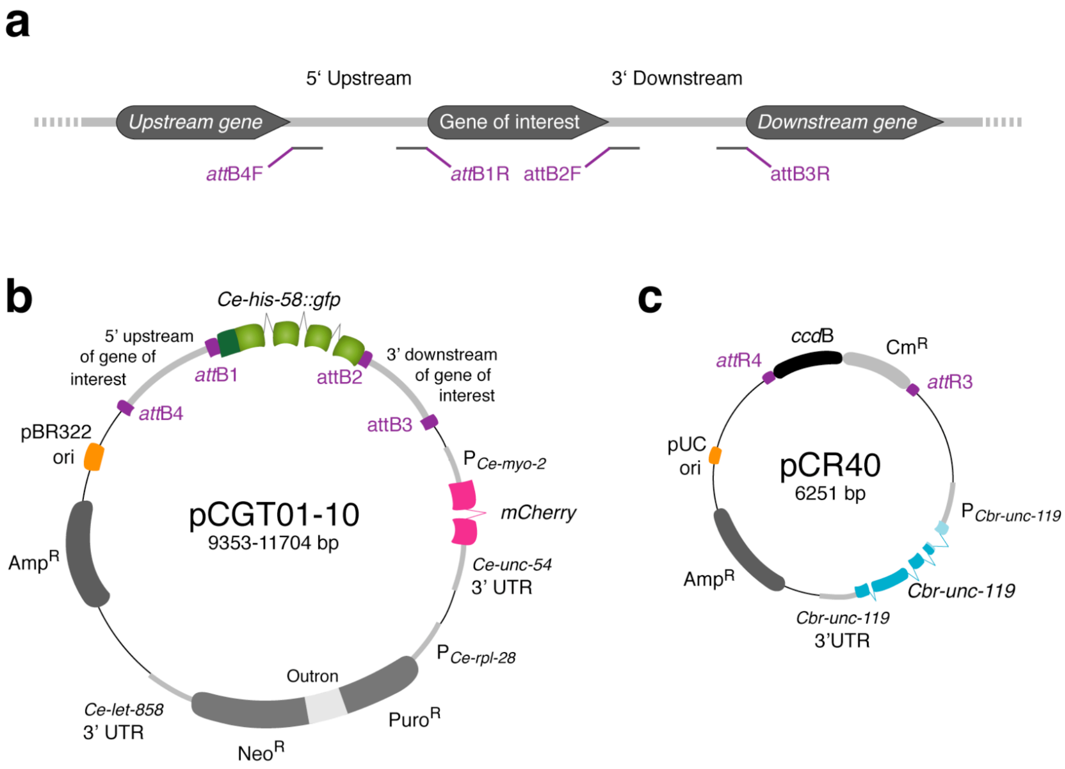


Figure 1. Transcriptional reporter strategy. **a**. Schematics of the gateway cloning strategy. For each gene of interest, primers carrying appropriate attB sites were designed, as shown. See text for details. **b**. Schematics transcriptional reporter vectors used to generate animals carrying transcriptional reporters for genes of interests. pCGT vectors 1-10 carry approximately one kilobase of the sequences directly upstream and downstream of the genes of interests, around a *Ce-his-58::gfp* fusion, as well as a mCherry reporter. **c**. pCR40 carries the *Cbr-unc-119* gene used for rescuing the *unc-119* phenotype, and was a gift from Christopher Richie (NIDDK, NIH).

9. Biolistic transformation of *C. elegans*

Generation of transgenic animals was carried out following a bombardment protocol from the Seydoux lab (Johns Hopkins Medical School, Merritt and Seydoux 2010), using a bombardment apparatus kindly supplied by Iqbal Hamza (University of Maryland). Briefly, *Ce-unc-119*(ed3) worms were grown on large NGM agar plates covered with a thick layer of OP-50 bacteria at 25°C. Large populations were then transferred to liquid S medium and shaken at 140rpm, 25°C. 0.5mL of young gravid adults were harvested for each bombardment and co-bombarded with 5μg of one of the pCGT vectors and 5μg of pCR40. pCR40 was a gift from Christopher Richie (Golden lab, NIDDK, NIH) and carried a *Cbr-unc-119* rescuing fragment (Fig. 1c). Worms were plated on large peptone-enriched NGM agar plates (3 g/L NaCl, 3.75 g/L peptone, 22 g/L agar, 5 mg/L cholesterol, 2 mg/L uracil, 0.1 M $CaCl_2$, 0.1 M $MgSO_4$ and 25 mM KH_2PO_4) covered with a thick layer of OP-50 bacteria., and incubated at 25°C for 10-12 days. Plates were then screened for wild-type animals whose pharynx expressed mCherry. Such individuals were transferred to fresh plates and their progeny observed to identify stable transformants, whose progeny was mostly composed of worms expressing mCherry in their pharynx.

10. Mapping profile

Using the first base of the start site as position 0, the position of the first base of each alignment attributed to a gene was plotted against its corresponding position along the gene sequence, generating the equivalent of a "heat maps" for candidate genes. The mapping profiles were compared to gene model predictions from WormBase in order to assess where the alignments were located in the gene (UTRs, exons or introns). See Figure 2b for an example.

11. Promoter analysis

Several programs to identify over-represented k-mers in a set of given sequences are available online. We used Weeder (Pavesi, Mereghetti et al. 2006; Pavesi, Zambelli et al. 2007), Gibbs (Thompson, Rouchka et al. 2003; Thompson, Palumbo et al. 2004; Newberg, Thompson et al. 2007; Thompson, Newberg et al. 2007), BioProspector (Liu, Brutlag et al. 2001) and MEME (Bailey and Elkan 1994) to analyze various sets of sequences to identify sites significantly over-represented. We also used TESS (Schug and Overton 1997) and JASPAR (Sandelin, Alkema et al. 2004; Bryne, Valen et al. 2008) to look for binding sites of known male-promoting factors in various sets of sequences, as well as Perl scripts developped by CGT after kind suggestion from Steve Mount (Appendix 1).

12. Note about Wormbase

WormBase (www.wormbase.org) is a dynamic website based on data that keeps on being improved. Updates are released every 2-3 months. Datafreezes, which are saved, stable copies of databases at a given point in time are available. To avoid constant tedious catch-up, we settled on the latest datafreeze at the time (WS210) to conduct conservation of synteny assessment. As shown in Chapter 1, the list of genes of interest was mostly the same release after release. In addition, in so far as the following experiment aimed at identifying accurate orthologues for candidate genes, the actual release used was less important in terms of homology assignment for genes of interest.

III. Results

1. Conservation of synteny as a discovery tool

Wormbase homology assignment is based on protein sequence similarity. To more accurately assign orthologues for genes of interest, I used WormBase information to establish synteny maps for 116 highly sex-biased genes missing in one or both of the selfing species (Fig. 2a). In some cases the syntenic landscapes allowed further understanding of the evolutionary dynamics of these genes and enabled either attribution of orthologues missed by WormBase, or confirmation of the lack of orthologues in one or both of the selfing species. To restrict further analyses to a smaller set of high-confidence genes, I also examined the correlation of mapped reads with gene structure predictions (Fig. 2b), as well as with results from *in silico* resources such as Pfam (Finn, Mistry et al. 2010), Prosite (de Castro, Sigrist et al. 2006; Sigrist, Cerutti et al. 2010), SignalP (Bendtsen, Nielsen et al. 2004), Phobius (Kall, Krogh et al. 2004) and Treefam (Li, Coghlan et al. 2006; Ruan, Li et al. 2008). Pfam and Prosite allow for identification of conserved protein domains, and the latter also identifies putative phosphorylation sites. SignalP and Phobius predict position of signal peptide, as well as membrane segment in the case of Phobius, based on protein sequence composition. Treefam gathers curated phylogenetic trees and has information on some *C. remanei* genes (Fig. 2c).

Out of the 116 genes defined as missing a homologue in one or both of the selfing species, nine actually had homologues, not indicated as such on WormBase as of the WS210 datafreeze. One gene was discarded because its mapping profile showed that the detected gene was in fact its neighbour, and six more because there was not enough information to determine orthology with certainty. 54 genes displayed patterns of conservation that did not fit the working hypothesis and were not studied further. 21 belonged to large gene families which tend to have varying number of members in different Caenorhabditis species (The *C. elegans* consortium 1998; Stein, Bao et al. 2003; also see review by Thomas 2008). Those were not studied further because orthology could not be assigned with certainty. Interestingly however, 13 genes belonged to the same gene family, defined by the presence of a repeated sequence. This gene family will be referred to henceforth as the novel-repeat containing (VRM) gene family. 4 of those genes, *Cre-vrm-1* to *4*, were studied further, along with 13 other genes whose patterns of conservation and/or potential functions were of interest (Table 2).

2. Confirmation of gene structure prediction

To check the accuracy of gene structure predictions available on WormBase, I designed primers spanning exon/intron junctions and amplified and sequenced cDNA fragments of all 17 candidate genes from total RNA from mixed-stage populations (Appendix 2, see Methods). Gene predictions were found to be accurate, with the exceptions of *Cre-vrm-1*, *Cre-vrm-2* and CRE04599 for which one exon/intron boundary was wrong – and hence corrected by this work. Attempts to amplify the first 4 exons of CRE23534 were not successful using a forward primer corresponding to the predicted ATG. Another forward primer designed to anneal to the first base of a downstream ATG of the predicted start site, in frame with the rest of the cDNA sequence, did allow amplification of a cDNA fragment whose sequence corresponded to that of CRE23534, suggesting either a mistake in the start site of the WormBase prediction, or an unannotated splice variant. Lastly, no amplification products were recovered from multiple combinations of primers for CRE28795 and CRE1138, confirming the lack of correlation between mapping profiles of these genes to their gene predictions. As a result, neither CRE28975 nor CRE1138 were studied further, reducing the list of genes of interest to 15.

Table 2. Candidate genes characteristics. Orthologues were determined as described in text, *in silico* predictions were compiled from Pfam, Prosite, SignalP3.0 and Phobius. In the case of CRE28795 and CRE11138 (grey), bias were not confirmed by qRT-PCR because of wormbase sequence misprediction (see text). *Cre-mss-2* was added to the list based on its genomic location (see text).

Gene ID	**RNA-seq bias**	**qRT-PCR bias**	**Note**	**Orthologues**	***in silico* predictions**
Cre-mss-1	Male	Male, L4-Adults	Tandem duplicates	*C. brenneri*, one copy	Signal peptide
Cre-mss-2	*Not in dataset*	Male, Adults			Signal peptide
Cre-mss-3	Male	Male, Adults			Signal peptide
Cre-mss-4	Male	Male, Adults			Signal peptide
Cre-vrm-1	Male	Male, Adults	Paralogues	*C. japonica, C. elegans, C. brenneri, C. briggsae* – different number of repeats	Novel-repeat containing genes
Cre-vrm-2	Male	Male, Adults			
Cre-vrm-3	Male	Male, Adults			
Cre-vrm-4	Male	Male, Adults			
CRE28795	Male	*N/A*		*C. japonica* and *C. briggsae*	Membrane protein
CRE28796	Male	Male, L4-Adults		*C. japonica* and *C. brenneri*	Membrane protein
CRE11138	Male	*N/A*		*C. brenneri*	Membrane, hydrolase
CRE01361	Male	Male, L4-Adults		*C. japonica*, *C. brenneri* and *C. briggsae*, pseudo-gene in *C. elegans*	S/T phosphatase
CRE05483	Male	Male, L4-Adults		*C. japonica, C. brenneri* and *C. briggsae*	Y phosphatase
CRE12267	Male	Male, L4-Adults	Tandem duplicates	1 in *C. elegans,* 2 in *C. brenneri*	
CRE12278	Male	Male, L4-Adults			
CRE04599	Male	Male, L4-Adults		*C. japonica*, *C. elegans, C. brenneri*	S/T kinase
CRE23534	Male	Male, L4-Adults		*C. japonica*, *C. elegans, C. brenneri*	Lipase

Figure 2. Characterizing a gene of interest - CRE01361 as an example. **a**. Syntenic landscape were built for each gene of interest (see text). Groups of orthologous genes are circled in purple boxes, the orthologues of the gene of interest in yellow. In this example, CRE01361 has a orthologue in *C. briggsae*. The genomic DNA sequence of the pseudo-gene in *C. elegans* presents 60.4% identity to that of the CRE01361 locus. **b**. Heat maps were generated for each gene by plotting the number of alignments at each position of the genomic location of a gene of interest from both male and female cDNA library, and correlated to in silico predicted gene structure and potential existing ESTs data. Attention was paid to predicted protein structure and function, here a metallophosphatase domain. **c**.

When available, Treefam trees were looked at to decipher phylogenetic relationship of the gene of interest and its homologues in *C. elegans* (green) and *C. briggsae*. Tree modified from www.treefam.org.

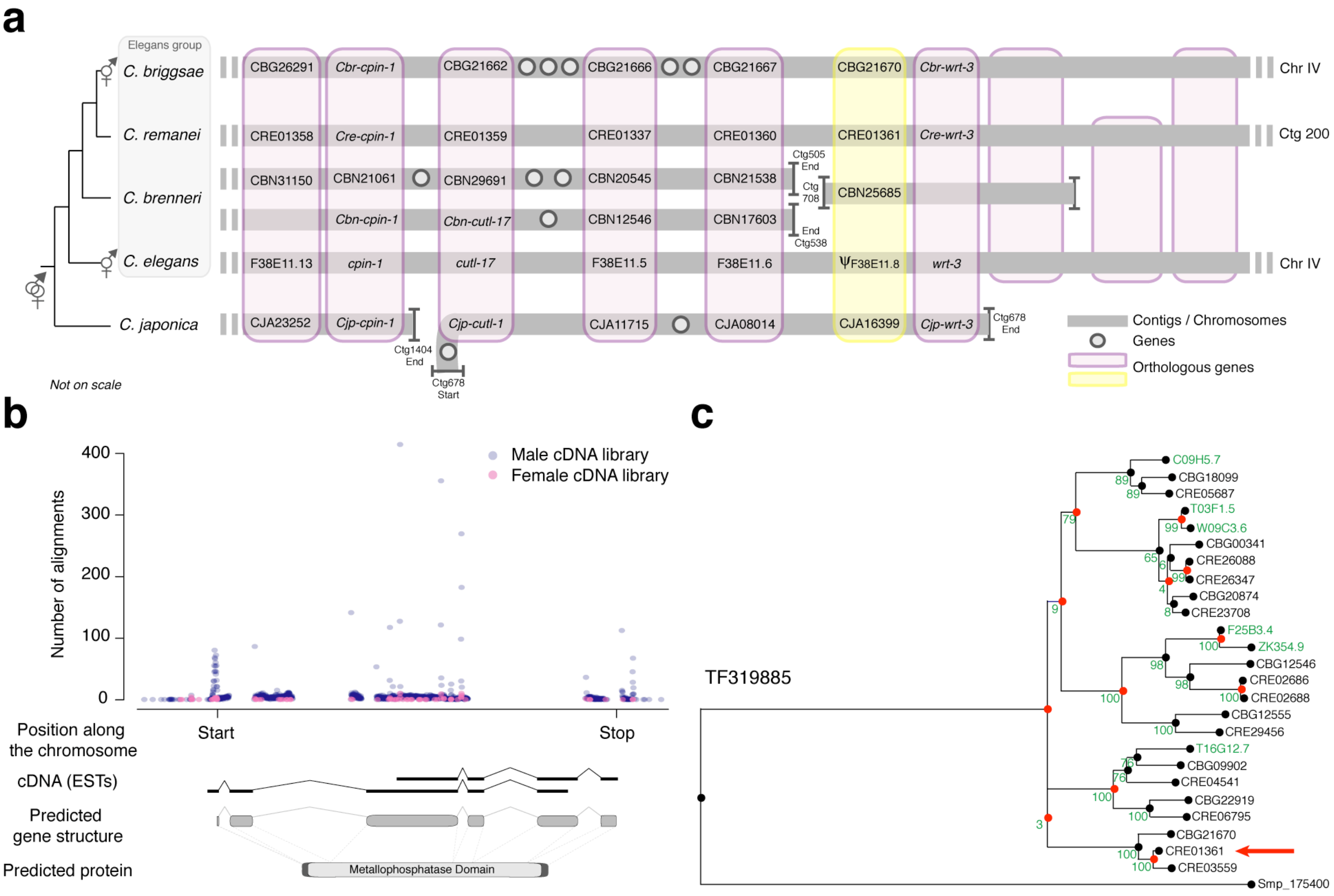


Figure 2.

As expected, *Cre-unc-94*, *Cre-nol-1* and *Cre-vha-8* were expressed throughout larval development as well as by adults, by both males and females (Fig. 3a). CRE24268 and CRE08297 by females from the L4 stage and on (Fig.3b), and CRE05585 and CRE01558 were expressed increasingly by males from the L4 stage and on (Fig. 3c). All of the candidate genes that we selected were expected to diplay a male-biased expression. We confirmed that only males expressed them (Fig. 4a, also see Appendix 2), either from the L4 stage or shortly after (young adults) and on into adulthood (Fig. 4b, also see Appendix 2).

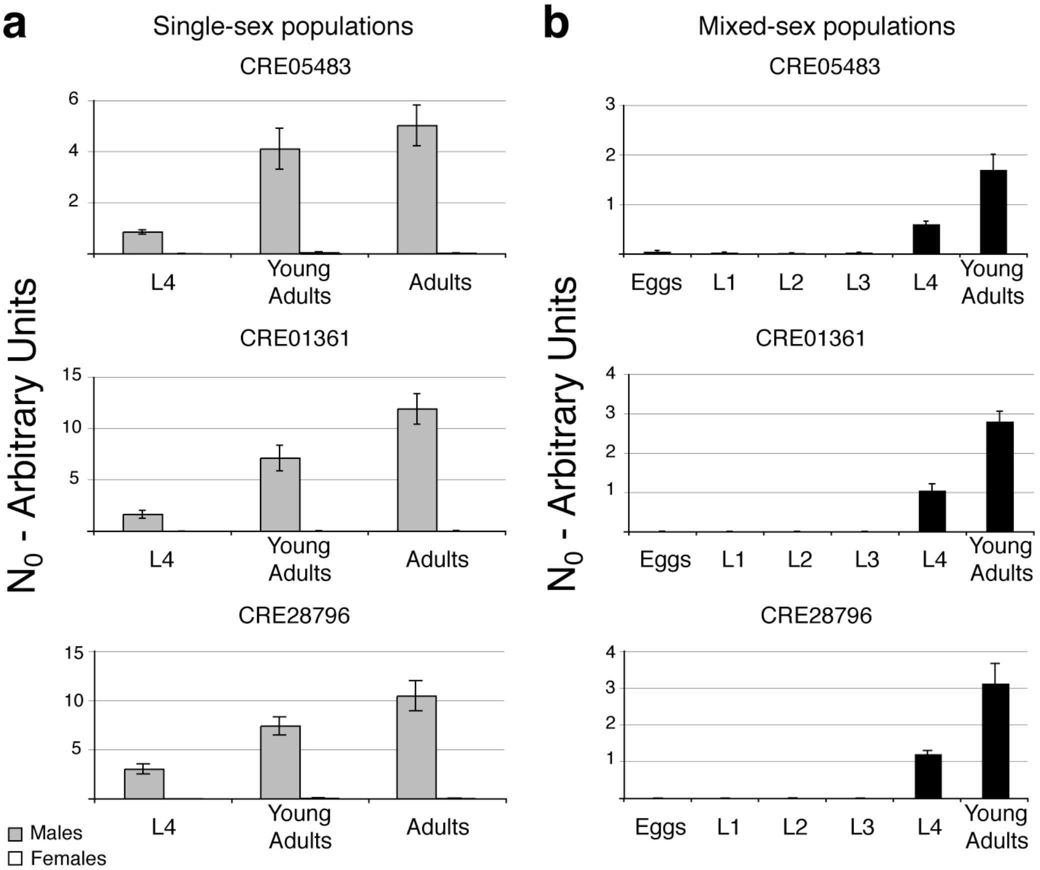


Figure 4. Developmental gene expression patterns for selected candidate genes missing in selfing species. The expression level of genes of interest was measured by quantitative RT-PCR in males (grey) and females (white) at different stages (**a**) as well as in a developmental series of mixed-sex populations (**b**). The standard error representing 4 to 7 replicates is indicated (see Methods). Note: female expression levels in (a) are not zero, but extremely low.

4. Spatial expression pattern

The temporal expression patterns obtained by quantitative RT-PCR are consistent with expression of the genes of interest in, or in association with male-specific structures that develop during the last larval stage, such as male-specific neurons, male tail and mating structures, and testis (Emmons 2005). However Caenorhabditis nematodes are too small and fragile to easily dissect tissues in quantities adequate for most expression assays. In order to identify precisely what cells or tissues are expressing candidate genes, we turned to other methods. Repeated attempts to characterize spatial expression patterns via mRNA in situ hybridization were not successful. Therefore transgenic transcriptional reporters were emphasized.

Though closely related to the model organism *C. elegans, C. remanei* is currently not amenable to the standard *C. elegans* transformation protocols, gonad micro-injection and biolistic rescue of *unc-119* mutants (Praitis 2006). However, biolistic transformation, using antibiotic resistance as a selection marker (Semple, Garcia-Verdugo et al. 2010) may be more applicable, as it does not require propagation of an uncoordinated mutant unable to mate. I developed transcriptional reporters for 8 candidate genes and 2 control genes using the multisite GateWay technology from Invitrogen (Fig. 1b, Table 3). The destination vector into which the putative promoters and 3' UTRs of genes of interest were cloned was generated by Jennifer Semple (Dupuy lab). This allows selection of transformants by antibiotic resistance (Semple, Garcia-Verdugo et al. 2010). However, because of time constraints and because an experiment by another student in the lab, Gavin Woodruff, necessitated the use of similar expression domain controls, we chose to transform

C. elegans unc-119 mutants by co-bombardment with pCR40 (vector carrying *unc-119* rescue fragment), using appropriate controls.

Table 3. Correspondance between genes and vector names.

Gene ID	Vector name
Cre-vrm-4	pCGT01
Cre-vrm-1	pCGT02
Cre-vrm-3	pCGT03
Cre-pie-1	pCGT04
CRE28796	pCGT05
Cre-mss-1	pCGT06
Cre-mss-2	pCGT07
Cre-mss-3	pCGT08
Cre-mss-4	pCGT09
Cre-spe-11	pCGT10

In *C. elegans*, the *pie-1* promoter drives expression in all germ cells from the late L1 stage to the adult stage (Merritt, Rasoloson et al. 2008) while its 3'UTR leads to expression in oocytes and embryos (Merritt and Seydoux 2010). The *Ce-spe-11* promoter drives expression in developing sperm in both hermaphrodites and males (Merritt, Rasoloson et al. 2008). Both *Ce-spe-11* and *Ce-pie-1* are conserved in *C. remanei*, and were used as controls for expression in the male and female germline respectively. Because transcriptional regulation seems to be well-conserved across the genus (Ruvinsky and Ruvkun 2003), expression domains of *Ce-pie-1* and *Ce-spe-11* should be recapitulated in *C. elegans* using the promoting sequences from *Cre-pie-1* and *Cre-spe-11.*

Transformation of *C. elegans unc-119* was carried out with three vectors, pCGT03, pCGT05 and pCGT09 (Table 3) in addition to those corresponding to *Cre-pie-1* and *Cre-spe-11*. At the time of writing, the lines obtained from biolistic transformation have not been characterized.

5. Functional characterization

To address the function of the 15 genes of interest, I injected dsRNA specific to the genes of interests into adult *C. remanei* females, in order to knock down expression in their progeny (Winston, Sutherlin et al. 2007; Felix 2008). However, none of the injections produced an obvious phenotype.

6. Specific examples

Based on their genomic location, two groups of genes stood out from the other genes of interest. *Cre-mss-1* to *4* (for Male Secreted Short proteins) constitute the first group of interest. *Cre-vrm-1*, *Cre-vrm-2*, *Cre-vrm-3* and *Cre-vrm-4* are part of the second group, which is the VRM family (for Variable Repeat numbers, Male-enriched proteins).

6. 1. MSS family

The four genes in this group are tandem duplicates on Contig 9 of the *C. remanei* genome assembly, between *Cre-twk-37* and CRE29434 (Fig. 5a). Among these, only three were detected in our first analysis (Table 2). The fourth, *Cre-mss-2*, was however detected by the TopHat / Cufflinks analysis (Table 4). Furthermore, because of its genomic location, and because *Cre-mss-2* displayed strong similarity to the other three genes, it was included in the analysis. Its expression pattern is male-specific, like that of the other genes of the family (Fig. 5b).

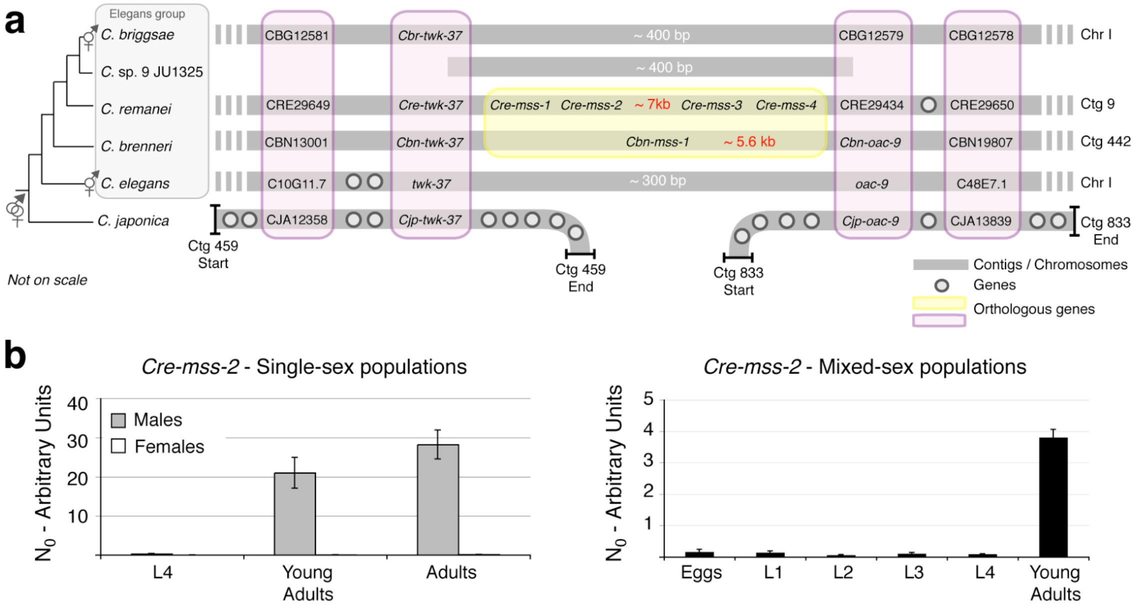


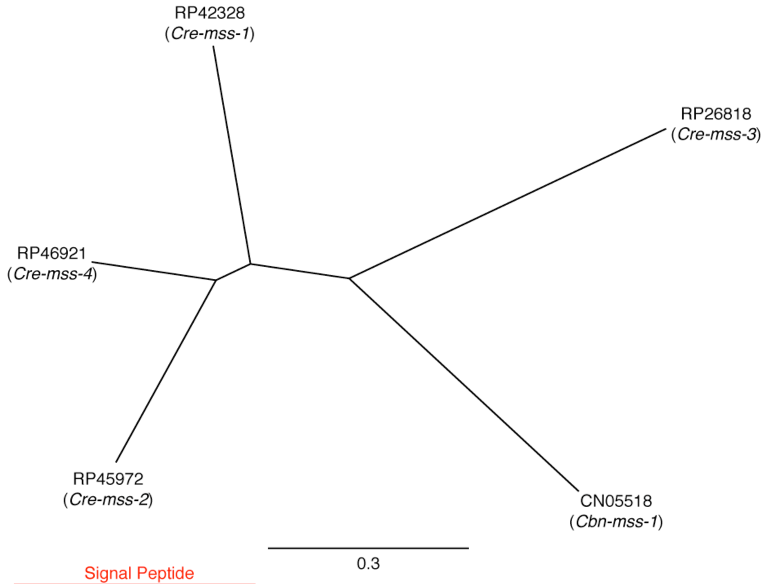
Figure 5. *Cre-mss* genes characteristics. **a**. Syntenic landscape of the MSS genomic locus. Groups of orthologous genes are circled in purple boxes, the orthologues of the genes of interest in yellow. The distance between *twk-37* and *oac-9* orthologues in *C*. sp.9 was inferred by sequencing. **b**. The expression level of all 4 genes of interest was measured by quantitative RT-PCR in males (grey) and females (white) at different stages as well as in a developmental series of mixed-sex populations. The same pattern was observed for all 4 genes. Only results for *Cre-mss-2* are diplayed for the sake of simplicity. The standard error representing 4 to 7 replicates is indicated (see Methods).

The MSS family genes probably diverged recently, based on overall conservation between protein sequences (Fig. 6). Synteny analysis (see above) showed that while the surrounding genes were very well conserved in *C. briggsae, C. brenneri, C. elegans* and *C. japonica*, the genes of the MSS family themselves did not have homologues in the selfing species. In *C. brenneri* one gene, *Cbn-mss-1*, is found between *Cbn-twk-37* and *Cbn-oac-9*, orthologues of respectively *Cre-twk-37* and CRE29434. The protein sequence of *Cbn-mss-1* shared 30% pairwise identity with the protein sequences coded for by the genes of the small cluster (Fig. 6). In addition, the close phylogenetic relationship of *Cbn-mss-1* to the genes of the small cluster was further confirmed by its position in a neighbor-joining tree based on a MUSCLE alignement of the protein sequences encoded by these 5 genes (Fig. 6).

Table 6. Expression level of small cluster genes. Significant expression differential between males and females is indicated by an asterisk.

Gene IDs	Preliminary survey			TopHat / Cufflinks analysis		
	Male counts	Female Counts	Male Counts / Female Counts	Male FPKM	Female FPKM	Male FPKM/ Female FPKM
Cre-twk-37	Expression not detected			2.4	0	N/A
Cre-mss-1	4702.3	51.4	91*	3250.7	45.8	74*
Cre-mss-2	Expression not detected			657.2	6.6	99*
Cre-mss-3	2441.8	7.3	332*	1185.3	2.0	597*
Cre-mss-4	3548.9	20.2	175*	6225.8	62.7	99*
CRE29434	Expression not detected			180.1	0.45	395*

In *C. japonica*, syntenic analysis indicated that the homologues of *Cre-twk-37* and CRE29434 were on different contigs of the *C. japonica* assembly, contig 459 and contig 833. The sequences of these two contigs are not related to each other as they failed to align to each other. In addition, only CJA12538 and *Cjp-twk-37* on contig 459, and *Cjp-oac-9* and CJA13839 on contig 833 had orthologues in *C. briggsae, C. remanei, C. brenneri* and *C. elegans* as indicated on Figure 7a. Furthermore, a BLAST search using both tBLASTn and BLASTp using the sequences of the proteins coded for by the *mss* genes against the *C. japonica* databases returned no hits. Altogether this suggested that the *mss* genes had no orthologues in *C. japonica*. However, given the fact that both *twk-37* and *oac-9* belong to large gene families (see TF313063 for *oac-9* and TF316115 for *twk-37,* www.treefam.org), in addition to the poor status of the *C. japonica* genome assembly and gene prediction, it is likely that the homology attribution of *C. japonica* genes is not always accurate, and might not be in the case of *Cjp-twk-37* or *Cjp-oac-9*. In addition, the incomplete genome sequence of *C. japonica* does not allow to draw an accurate conclusion from the analysis of syntenic conservation as absence of a gene or a protein sequences from the *C. japonica* databases does not imply its non-existence.



```
                    Signal Peptide
RP26818(Cre-mss-3)  MFRKNTFLFMTFALLVIISRAQVDSSDDVESTTEFDYDYDGDNDEQVEVTEEPEKEE----------ETPTEKPEEET
RP42328(Cre-mss-1)  MIRKTTILFLALALLATVARGADGDNVEAGDAQLPSGGAQKQGGENKSTEEKK----------------STEAPKLET
RP45972(Cre-mss-2)  MIQKTTILFMALALLAIFAHGFP--TPPSGGADDANKDEDTPADTPEDTTDDNAAAGVPGTPDLPRLPQLPIPAPEAPK
RP46921(Cre-mss-4)  MIQKTTILFMALTLLAIFARGQN--NVTAGGQNNTAQGPP-----------------------------PAAAPG---
CN05518(Cbn-mss-1)  MIRKTTIFFLFLALLAIAHAADPENNTDTGTSGVGKSGGD------------------------------IAKPTLA-

RP26818(Cre-mss-3)  TAEVPIEETKTTEAESTEEEKKTTPAPAPIAPRVTAAPQVADPPPAEVEDVVKIENTTSSSTFMSYGFGASVVLLAAL
RP42328(Cre-mss-1)  PAENAPKDTQ----------------PDAGSNSDAATP----------EAPAEGGNGTSSSGFMSYGLGASVVLLAAL
RP45972(Cre-mss-2)  APEDPKDGTSNQGGKDEEKSKSTEAPPATAESGSNTTP----------EAPVEGGNGTSSSAFMSYGLGASVVLLAAL
RP46921(Cre-mss-4)  --------------------------PAATEVEGTGEP---------------AGNGTSSSAYLNYGLGASFVLLAAL
CN05518(Cbn-mss-1)  --------------------------PEHHEPEKEEKP---------VSDEKKPTNNTASAAQ--YAFGASLVLLATL
```

Figure 6. MSS proteins alignment. The alignment was performed using MUSCLE with a BLOSUM62 scoring matrix. Amino-acids with 100% similarity are indicated in a black box, 80-100% in dark grey, and 60-80% in light grey. The neighbor-joining tree derived from this MUSCLE alignment is represented (Drummond AJ 2011). A conserved putative signal peptide is present at the N-terminus of all homologues indicated by a red box.

C. sp. 9 is the gonochoristic sister species of *C. briggsae* (Cutter, Yan et al. 2010; Woodruff, Eke et al. 2010). Its phylogenetic position makes it an ideal candidate to test whether the genes of the MSS family are specific to gonochorist species. Because the genomic region seemed to be conserved in the Elegans group, I designed forward and reverse primers to conserved regions of *twk-37* and *oac-9* respectively to amplify the region between these genes in *C.* sp. 9. This region could be amplified using the same primer pair in *C. briggsae, C. elegans, C. remanei* and *C.* sp.9, using both short and long extension time. As expected, long products were observed for *C. remanei*, and short ones for *C. elegans* and *C. briggsae* (Fig. 5a). Amplification in *C.* sp. 9 yielded a short band of about the same size than in *C. briggsae*. Sequencing of the band revealed that it shared 61.1% identity with the *C. briggsae* sequence, and lacked a gene between *twk-37* and *oac-9*. In addition, a deletion was identified in the *C.* sp.9 fragment compared to that of *C. briggsae*, confirming that the amplified sequence was not a *C. briggsae* contamination.

6.2. VRM gene family

Cre-vrm-1, *Cre-vrm-2*, *Cre-vrm-3* and *Cre-vrm-4* are located on contig 23 of the *C. remanei* genome assembly. These genes were defined as belonging to a group of 16 paralogues on WormBase, all but one of which are located on contig 23 within a stretch of about 136 kb. This region contains 28 genes according to WormBase release WS210. Preliminary analyses were limited to the four genes initially identified. Using tBLASTn or BLASTp, all four genes returned the same best hit from the *C. elegans* databases, Y1A2B.5, a "novel protein coding gene". Interestingly, the sequence of Y51A2B.5 seemed to align to several locations in *Cre-vrm-2*, suggesting a repeated pattern in the *Cre-vrm-2* sequence. A dot-plot representing the alignment of the protein sequences of *Cre-vrm-1*, *Cre-vrm-2*, *Cre-vrm-3* and *Cre-vrm-4* against themselves and each other revealed that a sequence of about 110 aa seemed to be repeated three to five times within each protein sequence (Fig. 7). Based on the patterns observed on the dot-plot, sequences corresponding to potential repeats were extracted from the protein sequences of

Cre-vrm-1, *Cre-vrm-2*, *Cre-vrm-3* and *Cre-vrm-4*, as well as from those of the other

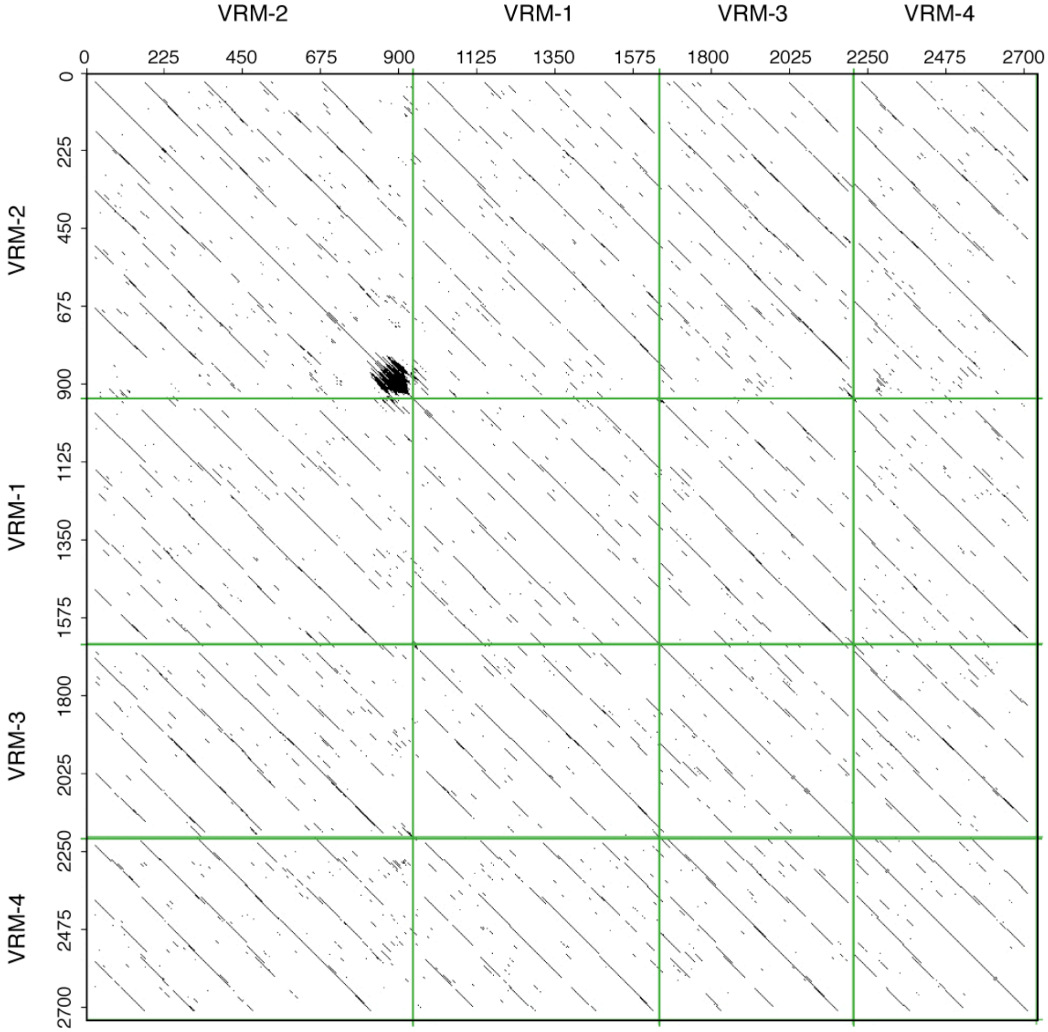


Figure 7. Dot-plot representation of VRM protein sequences. The protein sequences of VRM-1, VRM-2, VRM-3 and VRM-4 were concatenated in a long sequence. A dot-plot representing similarities between sequences was generated using JDotter (Brodie , Roper et al. 2003).

12 paralogues, and aligned using MUSCLE. The protein sequence of Y51A2B.5 was included (Fig. 8). The predicted protein sequences of some of the paralogues on contig 23 contained up to eight repeats. A neighbor-joining tree was generated from the MUSCLE alignment of these 50 sequences (Fig. 9). The repeats tended to mostly cluster independently from the protein sequence in which they were located,

indicating that they duplicated within an ancestral gene before the expansion of this gene family. In some cases - see for instance CRE14471 on Fig. 9 - repeat duplication within the gene seemed to have happen after gene divergence. The alignment did not allow me to decipher further the phylogenetic relationship between the *C. elegans* protein and the *C. remanei* proteins within the contig 23 cluster, or between some of the protein sequences of said cluster.

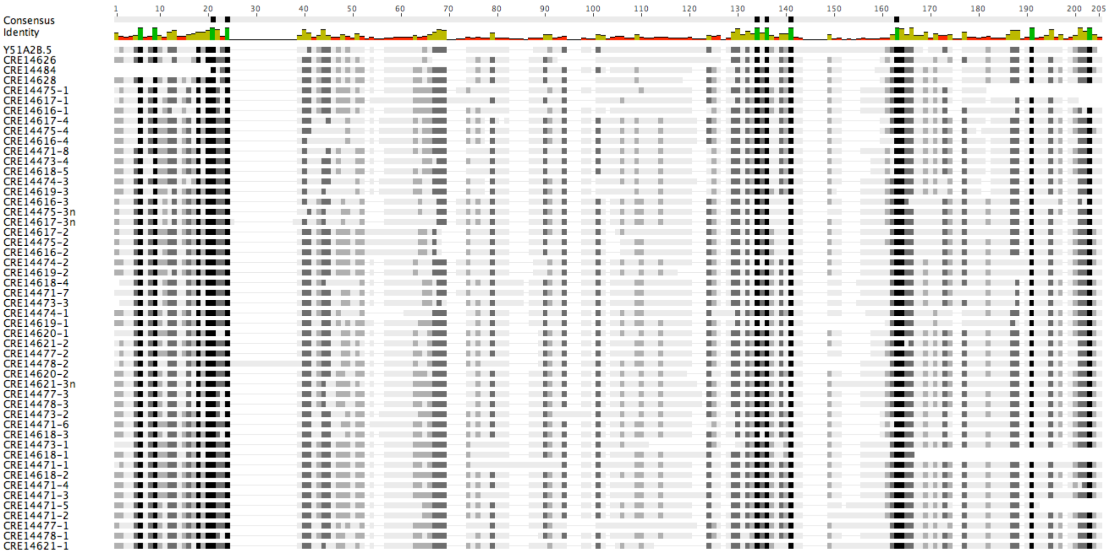


Figure 8. Alignment of individual repeats of the VRM protein sequences. The repeats from the 15 paralogues located on contig 23, as well as the sequence of Y51A2B.5 were aligned using MUSCLE and a BLOSUM62 scoring matrix. Amino-acids with 100% similarity are indicated by a black box, 80-100% in dark grey, and 60-80% in grey. Light grey indicates similarity under 60% and white no amino-acid (Drummond AJ 2011).

Since the sequence of the repeat seemed well conserved, the 50-sequence alignment was also used to create a HMMER profile (http://hmmer.janelia.org) in order to identify other proteins from the same family in the Caenorhabditis genus. Using HMMER, I identified 141 predicted protein sequences in the *C. remanei* database corresponding significantly to the VRM profile generated by HMMER. Similarly, there were respectively 15, 10, 108 and 20 predicted protein sequences in the *C. japonica, C. elegans, C. brenneri* and *C. briggsae* databases (Table 5). The number of repeats varied between species, and interestingly, not only did the gonochoristic species tend to have more novel-repeat-containing protein than selfing species, the maximum number of repeats seemed to be higher as well.

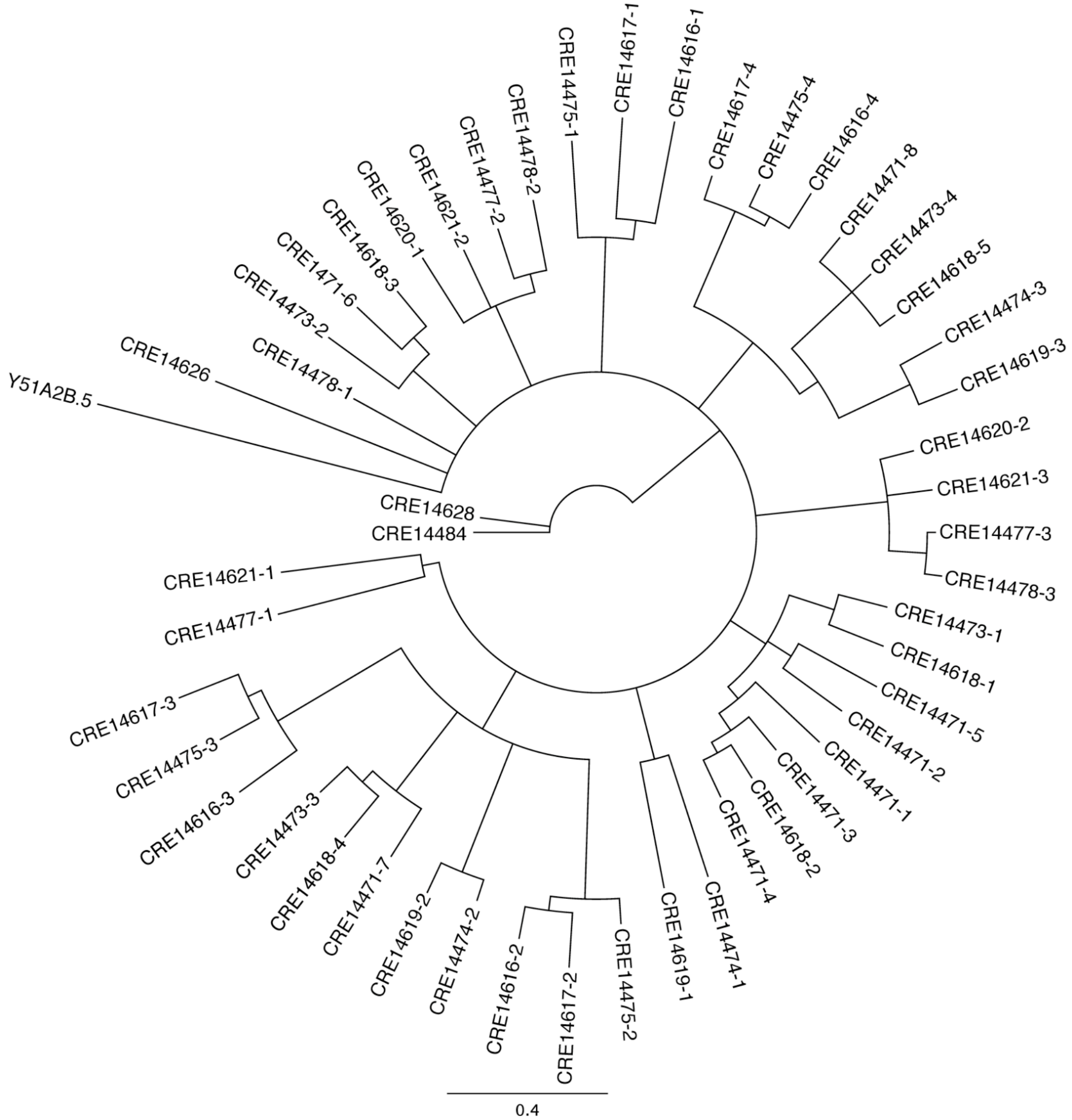


Figure 9. Phylogeny of individual VRM protein Repeats. Neighbor-joining tree generated from the MUSCLE alignment of the 50 sequences (Fig. 8) with Geneious (Drummond AJ 2011). The number indicated after the protein ID name is the number of the repeat within the protein sequence, 1 being the most N_{ter}.

Finally, I checked whether the expression of other genes than the initial four belonging to this family were detected in our preliminary survey. 19 of them were identified, and the expression of 18 of these was found to be male-biased. In addition, the expression of 87 of them was detected when the RNA-seq data was analyzed with TopHat / Cufflinks, and was found to be female-biased in only one case, male-biased in all other cases.

Table 5. HMMER search summary of the VRM repeat HMMER profile.

	C. japonica	*C. elegans*	*C. brenneri*	*C. remanei*	*C. briggsae*
Number of predicted proteins	15	10	108	141	20
Total number of repeats	16	10	125	189	24
Number of repeats within a protein sequence	1-2	1-2	1-4	1-8	1-3

7. Promoter analysis of genes with male-biased expression

The MSS and VRM family genes are all expressed in a strongly male-biased manner, but the two families are otherwise unrelated. The putative promoters of this set of genes might allow the identification of binding sites of potential male-promoting trancription factors. The *mss* genes seemed to be recently duplicated paralogues. However, the average pairwise sequence identity across any two of the intergenic regions corresponding to their putative promoters (57.7% on average) is close to that observed for sets of random intergenic sequences from the *C. remanei* genome (55.7% on average). The regulation of *mss* gene expression might be achieved through the same factors and binding sites, especially more so since neighbor genes in *C. elegans* tend to be co-expressed (Lercher, Blumenthal et al.

2003; Fukuoka, Inaoka et al. 2004). The intergenic sequences between each gene of the locus comprised between *Cre-twk-37* and *Cre-oac-9*, and the putative promoter of *Cre-twk-37* constituted a first set of sequences. The second set included the sequences of the first intron of each gene of the MSS family. A third set of sequences was defined by the intergenic sequences of all 28 genes spanning the region with the largest VRM family cluster of 15 paralogues. A fourth set included only intergenic sequences pertaining to the 15 VRM genes. Lastly, the first introns of the VRM genes formed a fifth set of sequences (Table 6).

Table 6. Sets of sequences used for promoter analysis.

Set 1	Putative promoters of *Cre-twk-37*, CRE29434 and small cluster genes
Set 2	First introns of small cluster genes
Set 3	Putative promoters of the 28 genes on contig 23 around novel-repeat containing genes cluster
Set 4	Putative promoters of the 15 novel-repeat containing genes on contig 23
Set 5	First introns of 15 novel-repeat containing genes on contig 23

In order to identify putative binding sites, I first looked for known binding sites of candidate transcription factors. Several male-promoting transcription factor are known in *C. elegans*, and corresponding binding sites have been identified for a few : ELT-1 (del Castillo-Olivares, Kulkarni et al. 2009), MAB-3 (Yi and Zarkower 1999) and TRA-1 (Zarkower and Hodgkin 1993). I used online resources to identify such binding sites in each set of sequences using TESS (Schug and Overton 1997) and the JASPAR database (Sandelin, Alkema et al. 2004; Bryne, Valen et al. 2008).

No TRA-1 sites were found in any of the sets of sequences. In addition, no known sites were identified in the set of sequences consisting of first introns. Putative MAB-3 binding sites were found in the likely promoters of *Cre-twk-37*, *Cre-mss-1*, *Cre-mss-3*, *Cre-mss-4* and CRE29434, and one ELT-1 site in that of *Cre-mss-1*

(Fig. 10). The expression of both *Ce-twk-37* and *Ce-oac-9* was found to be male-biased in previous microarray sudies (Jiang, Ryu et al. 2001; Reinke, Gil et al. 2004),

consistent with the presence of binding sites of male-promoting factors in their putative promoters. However, the expression of *Cre-twk-37* or CRE29434 was not detected in the first analysis of our datasets, possibly because their expression level was not high enough. Their expression was detected when the data was analyzed with TopHat and Cufflinks, and confirmed that it was much weaker than that of any of the genes of the small cluster, and not significantly sex-biased in the case of *Cre-twk-37* (Table 4). Since expression of the small cluster genes was detected, there might be other factors than MAB-3 or ELT-1 involved in the regulation of their expression. Similarly, while a multitude of MAB-3 putative binding sites were identified in the putative promoters of the 28 genes of contig 23, no ELT-1 sites were found at all, and MAB-3 binding sites were not present in the putative promoters of all of the genes estimated to have a male-biased expression.

I looked for putative binding sites of unknown factors in the same sets of sequences described above, hypothesizing that such a novel site should be over-represented. In a first step, I used online resources (Weeder, Bioprospector and Gibbs) to identify sites of varying length *k* that might be over-represented in different set of sequences. I then used TESS to locate in the same sets of sequences the sites of interest identified in the first step, and finally wrote a Perl script (Appendix 1) to locate and estimate the frequency of the sites in the genome.

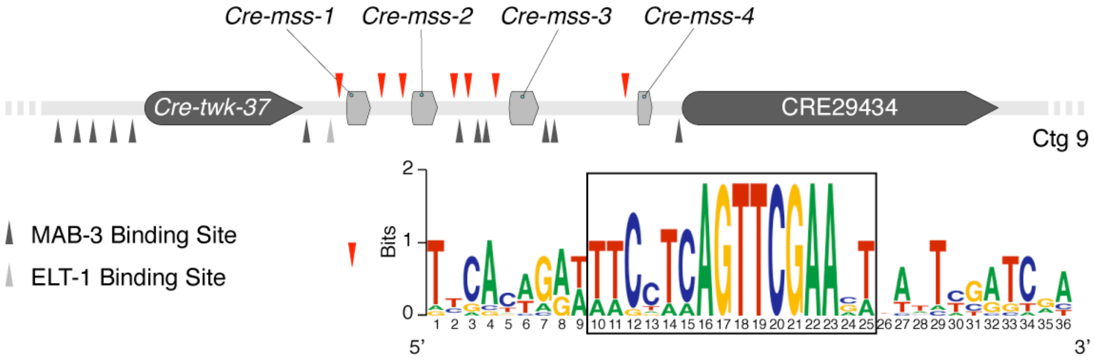


Figure 10. Promoter analysis of genes with male-biased expression missing in selfing species. Schematic of the genomic locus of the *Cre-mss* genes cluster (to scale). Genes are represented by rounded boxed arrow starting at the ATG and ending on the pointing side at the stop codon. ELT-1 and MAB-3 binding sites are indicated by grey and black triangles respectively. The location of the novel putative binding sites are indicated by red triangles. The sequence logo is based on an alignment of the 13 sites found in putative promoters of both the small cluster genes and the novel-repeat containing genes. A box indicates the core sequence of the putative binding site used for further analysis.

All three programs used (Weeder, Bioprospector and Gibbs) identified a similar palindromic site in the set of sequences 1 and 4 (Table 6), corresponding respectively to the putative promoters of *Cre-mss-1* to *4* and those of the *Cre-vrm* genes. I located the sites in the sequences using TESS and found that they were all located in the promoters of genes with male-biased expression (Fig. 10). In the MSS gene locus, the sites are all located within intergenic sequences, and seem to occupy random location with respect to that of the TSS of the genes of the locus. In contrast, of the 8 sites identified in the locus of contig 23 where VRM genes are located, six of them are located exactly 19 bp away from the predicted ATG of the closest gene, one is 16 bp away and the last one 9 bp away (data not shown).

In order to shed some light on the significance of these sites, the sequence logo generated from the alignment of the sites found so far was used to derive a 16 bp consensus site (indicated on Fig. 10 by a black box). The incidence of the consensus site was estimated in the genome reference sequences of *C. japonica*, *C. elegans*, *C. brenneri*, *C. remanei* and *C. briggsae*, and its frequency in each genome compared to the expected frequency of such a site based on overall nucleotide composition of each genome (Table 7). The incidence of the consensus site was found to be significantly higher than expected in every genome (χ^2 test, $p < 0.001$, Table 7).

Table 7. Frequency of site in the Caenorhabditis genomes. The site defined as [ATC](A|T)(C|G)[ACGT](A|T)(A|C)AGTTCGAA[ACG](T|A) was counted in each genome reference sequence with a Perl script. The expected number of sites are based on nucleotide frequencies in each genome, and the p-values correspond to a χ^2 test.

	Total number of sites	Expected number of sites	p-value
C. japonica	141	102	0.0001
C. elegans	212	80	1.13×10^{-48}
C. brenneri	363	136	3.51×10^{-84}
C. remanei	234	111	1.69×10^{-31}
C. briggsae	311	85	1.1×10^{-133}

If the novel motif functions as a transcription factor binding site, it would be expected to be enriched in intergenic regions, within putative promoters of genes. The sites corresponding to the consensus were categorized for each species as belonging to exons, introns, or putative promoters, the latter defined as sites located between genes and upstream of a transcription start site (Table 8). In every species, sites were found both within genes and intergenic regions. The proportion of sites located within genes ranged from 41 to 56% depending on the species. In *C. elegans*, the amount of exonic DNA was estimated to represent roughly 27% of the genome (The *C. elegans* consortium 1998). Assuming similar proportions for the other Caenorhabditis nematode species, the consensus site seemed to be significantly enriched in non-exonic DNA in every species (χ^2 test, $p < 0.05$ for *C. japonica*, $p < 0.01$ for *C. elegans, C. brenneri* and *C. remanei*, $p < 0.001$ for *C. briggsae*, Table 8).

Table 8. Location of the sites within the Caenorhabditis. The expected number of sites in exons is based on estimated transcriptome size in Caenorhabditis (27%, REF). The p-values correspond to a χ^2 test for number of sites in exonic DNA.

	C. japonica	*C. elegans*	*C. brenneri*	*C. remanei*	*C. briggsae*
In genes	44	96	95	77	129
...In exons	*14*	*25*	*39*	*32*	*29*
In putative promoters	53	76	125	112	125
...Within 1kb of TSS	*20*	*21*	*36*	*48*	*35*
Expected number of site in exons	26	46	59	51	69
p-value	0.0172	0.0017	0.0081	0.0077	less than 0.0001

Should the above site be involved in the regulation of expression of genes with a male-bias, it should be preferentially located within putative promoters of genes diplaying male-biased expression. Sex-biased expression information is available both in the case of *C. elegans* and *C. remanei* (Reinke, Smith et al. 2000; Jiang, Ryu et al. 2001; Reinke, Gil et al. 2004, this study). I compared the sex-bias distribution of the genes whose expression was detected to that of the genes whose expression was detected and whose putative promoter contained a site. In the case of *C. remanei*, I used both the list of genes from our initial analysis and from the TopHat / Cufflinks analysis. In *C. elegans* (Jiang, Ryu et al. 2001) as in *C. remanei* (this study), genes whose putative promoter contains a site were significantly more likely to be male-biased (χ^2 test, $p < 0.001$, Table 9). In addition, the trend held for sites located within 1 kb of the predicted transcription start sites in both species (χ^2 test, $p < 0.001$, Table 9).

Table 9. Correlation between site locations and male-biased expression. p-values correspond to a χ^2 test.

	C. remanei, initial analysis		*C. remanei*, TopHat / Cufflinks analysis		*C. elegans* (Jiang et al 2001)	
	Male	Female	Male	Female	Male	Female
Number of transcript detected	668	1,060	2,284	1,563	1,651	520
Total	6,268		21,711		12,486	
Number of genes with a site within putative promoter	13	1	19	1	26	11
Total	28		77		42	
p-value	3.24 e^{-6}		2.15 e^{-7}		7.61 e^{-29}	
Number of genes with a site within 1kb of TSS	9	1	15	1	9	3
Total	17		39		13	
p-value	4.93 e^{-8}		4.1 e^{-8}		9.10 e^{-11}	

IV. Discussion

1. Evolutionary histories of genes with male-biased expression and no homologues in selfing species

1.1. On the importance of curation.

Although chromosomal rearrangements distinguishing *C. elegans* and *C. briggsae* are common (Coghlan and Wolfe 2002; Stein, Bao et al. 2003), the global content of chromosomes remained mostly the same and micro-synteny is mostly conserved (Hillier, Miller et al. 2007). Given the current Caenorhabditis phylogeny (Cho, Jin et al. 2004; Kiontke, Gavin et al. 2004; Kiontke, Félix et al. in revision), *C. elegans* and *C. briggsae* are more distantly related than *C. elegans* and *C. remanei* or *C. brenneri*. Therefore, micro-synteny is expected be mostly conserved between Caenorhabditis species in the Elegans group. As a result, assessing synteny around genes of interest is a powerful tool to determine more precisely their phylogenetic conservation pattern.

The strains of *C. remanei*, *C. japonica* and *C. brenneri* used for genome sequencing retained heterozygosity (Barriere, Yang et al. 2009). While heterozygous regions are not likely to span only one gene in the genome, the strategy used for assembly led to short contigs representing different alleles of the same genomic locus (Barriere, Yang et al. 2009), and containing only a few genes (CGT, pers. obs.). Careful curation of candidate genes allowed them to be classified as single genes instead of several independent alleles of the same gene. In addition, mapping profiles allowed validation of a few candidates, and showed that the read coverage of gene loci were on average not uniformly spread throughout exons (Fig. 2) thus confirming in retrospect that the first analysis lacked precision in attributing reads to genes.

1.2. Phylogenetic conservation profiles

The list of 116 genes with highly sex-biased expression with which I started this study showed no significant pattern in the distribution of sex-bias. However, the 15 candidate genes corresponding to the selection criteria are all male-biased, consistent with previous observations of faster evolution and divergence of genes with expression levels biased towards the hetero-gametic sex in Caenorhabditis (Cutter and Ward 2005) and in other species (Torgerson, Kulathinal et al. 2002; Ranz, Castillo-Davis et al. 2003; Schultz, Hamra et al. 2003; Good and Nachman 2005;

Khaitovich, Hellmann et al. 2005; Zhang and Parsch 2005; Mank, Hultin-Rosenberg et al. 2007; Jiang and Machado 2009). Among the 15 genes we chose, most have a orthologue in *C. brenneri* and *C. japonica* (Table 2)*,* confirming loss in *C. elegans* and/or *C. briggsae*. A few do not have clear homologues in *C. japonica*, however given the status of the genome assembly and of the gene predictions of this species, other parameters redeemed these genes and made them interesting candidates.

The VRM family has members in all Caenorhabditis species surveyed, albeit different numbers and with varying numbers of repeats (Table 5). Most *C. remanei vrm* genes whose expression were detected were male-biased or male-specific in their expression. Genes expressed in sperm were shown to have the most abundant rate of species-specific loss, duplication and divergence between *C. elegans* and *C. briggsae* (Cutter and Ward 2005), consistent with the pattern observed in the VRM family across the genus. Their evolutionary history along with their expression patterns suggest that these proteins might be involved in sperm-related functions.

The *mss* genes displayed an interesting evolutionary dynamic with a recent duplication in *C. remanei* leading to the existence of 4 paralogues, one homologue in *C. brenneri* and no trace of homologues in *C. japonica*, *C. elegans*, *C. briggsae* and its gonochoristic sister species *C.* sp.9 (Cutter, Yan et al. 2010; Woodruff, Eke et al. 2010). While in the case of *C. elegans* and *C. briggsae* the BLAST results are rather reliable because their genome assemblies are, respectively, virtually perfect (Gerstein, Lu et al. 2010) and mostly complete (Ross, Koboldt et al. 2011), the answer is not so clear for *C. japonica* and *C.* sp. 9. As mentioned above, the *C. japonica* assembly is in rather poor shape, and that of *C.* sp. 9 is at an even earlier stage, and was reported recently to contain sequences from another species after cross-contamination of the samples used for sequencing (E. Schwarz, pers. comm.). The fact that micro-synteny is not conserved in that region in *C. japonica*, and the confirmation that there are no genes between the homologues of *twk-37* and *oac-9* in *C.* sp. 9 by direct sequencing from an uncontaminated source merely confirms the lack of a homologue at that locus in *C.* sp. 9 and *C. japonica*, but does not prove the lack of a homologue somewhere else in the genome.

The VRM family has more members in both *C. brenneri* and *C. remanei* than in *C. japonica*, *C. elegans* and *C. briggsae.* As explained above, the status of the *C. japonica* genome makes accurate estimation of the number of *C. japonica vrm* genes unreliable as of now. If there were more members than the detected 16, the decrease in size of the VRM family in selfing species could be more strongly correlated with mating system. Another explanation however is that the VRM family expanded after the split between *C. japonica* and the Elegans group, and shrunk again in *C. briggsae.*

CRE12278 and CRE12267 are recent tandem duplicates, both with orthologues in *C. brenneri*. No homologues could be identified in *C. briggsae* or *C. japonica*, and only one was found within the same syntenic location in *C. elegans*. In addition, the same location in the current contaminated (see above) *C.* sp. 9 assembly is hit significantly using tBLASTn with either protein sequence as queries, suggesting at least one potential homologue in *C.* sp. 9 (pending cross-contamination). Depending on the actual existence of a homologue in *C. japonica*, these genes either duplicated after the *C. elegans* / *C. brenneri* split, and were subsequently lost in *C. briggsae,* or their duplication originated before *C. elegans* and *C. brenneri* diverged, one copy was lost in *C. elegans* while both were lost in *C. briggsae*. Both interpretations are still consistent with loss in at least one selfing species and illustrate the rapid evolution of male-biased genes. Another interesting example is that of CRE01361 predicted to encode a Serine / Threonine Phosphatase (Fig. 2) which has orthologues in *C. japonica*, *C. brenneri* and *C. briggsae*, and whose orthologue became a pseudogene in *C. elegans,* indicating relaxation of selection in *C. elegans*.

Male-biased genes are expected to evolve at a faster rate than other genes in the Caenorhabditis genus (Cutter and Ward 2005). Thus a perceived loss in selfing species of a gene with male-biased expression in *C. remanei* could simply be because of sequence divergence. In this study however, I identified genes that are lost – as opposed to diverged – in selfing species and have clear orthologues in gonochoristic species, suggesting either a selection pressure in male/female species to maintain those genes and/or a relaxation of selection in selfing species. These genes were selected for further study to determine whether their role was associated with mating processes, in line with the hypothesis that they confer traits lost by selfing species as mating efficiency declined.

2. Expression and function of genes with male-biased expression and no homologues in selfing species

Caenorhabditis adult males develop specialized mating structures post-embryonically in the fourth larval (L4) stage (reviewed in Emmons 2005) and start spermatogenesis in late L4 (reviewed in L'Hernault 2006; Nishimura and

L'Hernault 2010). The spermatids are not activated – a process called spermiogenesis – until ejaculation, after which they mature in the uterus, crawl in the spermatheca and proceed to fertilization. The factor(s) triggering spermiogenesis are unknown as of now, but are thought to be contained in the seminal fluid (Ward and Carrel 1979). A variety of exogenous factors, including proteases (Ward, Hogan et al. 1983), are able to induce sperm activation in vitro. Mature spermatozoa are able to crawl by extending pseudopods thanks to polymerization and depolymerization of nematode-specific fibrous proteins called MSPs. Phosphorylation and de-phosphorylation have been shown to play a major role in sperm motility (Yi, Buttery et al. 2007; Yi, Wang et al. 2009).

Interestingly, in *C. elegans* male-derived sperm are bigger and faster than hermaphrodite self-sperm (LaMunyon and Ward 1998; Geldziler, Chatterjee et al. 2006). While the sperm-activating factor(s) are still unknown, spermatogenesis and fertilization-defective mutants have been extensively characterized in *C. elegans* (Latest review by Nishimura and L'Hernault 2010). Many genes involved in spermatogenesis are predicted to be kinases, soluble proteins or membrane proteins (Nishimura and L'Hernault 2010). Studying the function of these genes has been limited by the fact that genes expressed in spermatocytes, as well as neuronal genes, are refractory to RNAi (Tavernarakis, Wang et al. 2000; Kamath, Martinez-Campos et al. 2001; Kamath, Fraser et al. 2003; Reinke, Gil et al. 2004).

All candidate genes studied in this work were highly male-biased in their expression. In addition, temporal expression patterns showed that these genes were all expressed from the L4 stage on to adulthood, suggesting a lasting role after adult morphogenesis is completed. *In silico* predicted functions (kinase, phosphatase) or domains (signal peptide, transmembrane domain) together with the observed expression patterns would be consistent with potential roles in sperm activation, sperm motility or seminal fluid properties, supported by the lack of discernable RNAi phenotypes. A more thorough investigation would be required however to decipher the exact function and effect of any of these candidate genes. A first step would be to assess more precisely their spatial expression patterns in order to better predict their functions. Expression in sperm would prompt sperm motility and activation assay, while expression in the seminal vesicle would lead one to assay sperm activation, female behavior or male-male competition.

3. A novel male-associated candidate cis-regulatory element in Caenorhabditis

The characterization of sets of genes organized in cluster in *C. remanei* and presenting the same sex-biased expression provided an opportunity to explore the regulation of their expression. The novel site identified in this small set of genes was interestingly over-represented in the genome of *C. remanei* as well as in that of other Caenorhabditis species. In addition, this site was found to be more likely to be excluded from exons in all genomes surveyed, and in the case of *C. elegans* and *C. remanei*, was located preferentially in the putative promoters of genes male-biased in their expression. Lastly this site was partially palindromic, structural features allowing potential binding of dimeric factors.

Taken together these data suggest that this site may be a novel cis-regulatory element for a male-promoting factor. The pseudo-palindromic portion of this site – AGTTCGAAC – is also present in the human (38 occurences), mouse (73 occurences), fruit fly (2073 occurences) and plant (1059 occurences) genomes (BLASTn search). The higher number of occurences in the smaller genome of *D. melanogaster* compared to the bigger mammals genomes suggest that this element might be specific to Invertebrates, although more thorough searches have to be considered. While not related to the main topic of this study, this finding is worth pursuing by determining whether a factor does bind to this sequence and what effect it may have on the regulation of gene expression.

Conclusions

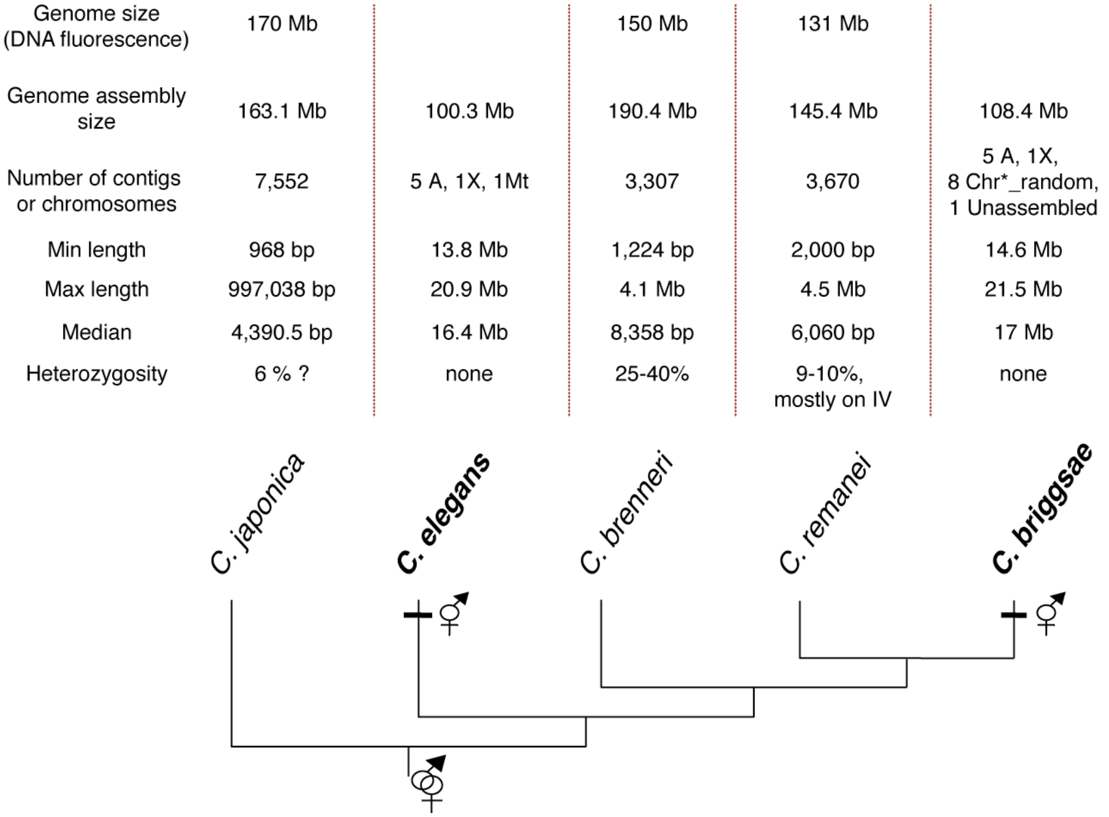
In this work I have identified and characterized *C. remanei* genes whose expression are strongly male-biased, and whose evolutionary profiles suggest that they have been lost in selfing species, perhaps as a consequence of decreased selection for efficient outcrossing. In addition, the temporal expression patterns of these genes are consistent with a role in mating-related processes, more specifically in functions associated with sperm, although spatial expression patterns and functional studies are required to confirm this hypothesis. The few genes that have been described here represent only a select fraction of genes lost in *C. elegans* and *C. briggsae* through emergence of selfing. Some of them also demonstrate that discovering genes missing in *C. elegans* is a way to find general attributes that cannot be studied in the "model species". Because the ways in which males and hermaphrodites of androdioecious species are poor at mating are varied, we expect that a deeper survey would allow the identification of many more such genes.

Chapter 3: Sex-biased expression in the Caenorhabditis genus

I. Summary

Mating systems shape genome structure over generations through their effects on population genetics (Lynch 2007) and reproductive traits. In the nematode genus Caenorhabditis, some species are androdioecious (male/self-fertile hermaphrodite), such as the model organism *C. elegans*, while most are gonochoristic (male/female) (Fig. 1, Cho, Jin et al. 2004; Kiontke, Gavin et al. 2004; Kiontke, Félix et al. in revision). Because *C. elegans* and *C. briggsae* propagate mainly through selfing, their effective population size is believed to be smaller than that of the extant gonochoristic species of the genus (Cutter, Baird et al. 2006). Lynch and Conery (2003) noted that genome sizes are generally inversely correlated with effective population size and proposed that this is because smaller populations accumulate repetitive DNA and genomic duplications. As a result, the genome size of gonochoristic species in the Caenorhabditis species were expected for a time to be smaller than that of *C. elegans* (Coghlan 2005). However, selfing imposes an extreme shift in population genetic dynamics. Under most scenarios for deleterious effects of transposable elements (TEs), high selfing rate is predicted to reduce their numbers (see Introduction, Wright and Schoen 1999). Under such models, genomes of outcrossing species should harbor more TEs than closely related selfing species. Consistent with this, there is more repetitive DNA in the genome sequence of *C. japonica, C. brenneri* and *C. remanei* than in those of *C. elegans* or *C. briggsae* (See Introduction, Table 1, Caenorhabditis

Repeat Libraries). Direct measurement of genome sizes have confirmed that among Caenorhabditis species in culture, selfing species have smaller genomes compared to gonochoristic species (Fig. 1, Thomas, Li et al. In prep.). The same is true in the Arabidopsis genus where self-fertile species *A. thaliana* has a smaller genome size (125Mb) than related obligately outcrossing species (Johnston, Pepper et al. 2005; Oyama, Clauss et al. 2008; Lysak, Koch et al. 2009). The recent sequencing of *A. lyrata*, which diverged 10 MYA from *A. thaliana* (Wright, Lauga et al. 2002; Beilstein, Nagalingum et al. 2010; Ossowski, Schneeberger et al. 2010) and whose genome size is 230Mb (Johnston, Pepper et al. 2005) led to the observation that the



	C. japonica	***C. elegans***	*C. brenneri*	*C. remanei*	***C. briggsae***
Genome size (DNA fluorescence)	170 Mb		150 Mb	131 Mb	
Genome assembly size	163.1 Mb	100.3 Mb	190.4 Mb	145.4 Mb	108.4 Mb
Number of contigs or chromosomes	7,552	5 A, 1X, 1Mt	3,307	3,670	5 A, 1X, 8 Chr*_random, 1 Unassembled
Min length	968 bp	13.8 Mb	1,224 bp	2,000 bp	14.6 Mb
Max length	997,038 bp	20.9 Mb	4.1 Mb	4.5 Mb	21.5 Mb
Median	4,390.5 bp	16.4 Mb	8,358 bp	6,060 bp	17 Mb
Heterozygosity	6 % ?	none	25-40%	9-10%, mostly on IV	none

Figure 1. **Caenorhabditis genome sizes and statistics.** Partial Molecular phylogeny of the Elegans group (Cho, Jin et al. 2004; Kiontke, Gavin et al. 2004; Kiontke, Félix et al. in revision) where androdioecious species are indicated in bold. Estimates of genome sizes for *C. japonica, C. brenneri,* and *C. remanei* were assessed by flow cytometry (Thomas, Li et al. In prep). Genome reference sequences statistics are based on the genome assemblies available on WormBase (Release WS224)

difference in size between their genomes was mainly due to TE and small deletions in non-coding DNA (Hu, Pattyn et al. 2011), although more genes are predicted in the *A. lyrata* genome sequence (Hu, Pattyn et al. 2011).

In Caenorhabditis, selfing species tend to lose ancestral traits related to mating, and are often much less efficient maters (see Introduction, Hodgkin and Doniach 1997; LaMunyon and Ward 1999; Chasnov and Chow 2002; Lipton, Kleemann et al. 2004; Geldziler, Chatterjee et al. 2006; Chasnov, So et al. 2007; Garcia, LeBoeuf et al. 2007; Kleeman and Basolo 2007; Palopoli, Rockman et al. 2008). Preliminary work has suggested that this is correlated with loss of genes in selfers whose expression is strongly sex-biased in a gonochorostic species, *C. remanei* (this study, Chapter 1 and 2). Here, I explore further the relationship between genome size reduction and loss of mating efficiency in a more complete study by comparing sex-specific gene expression genome-wide using RNA-seq in four species of the Caenorhabditis genus. I show that the adult transcriptome of *C. elegans* is markedly less complex than that of its gonochoristic relatives. Truly sex-specific quantifiable transcribed units (QTUs) are abundant in gonochoristic species but rare in *C. elegans*. Similarly, QTUs with strongly female-biased expression represent a smaller fraction of all QTUs in *C. elegans.* In addition, QTUs with a strong male-biased expression display weakened canalization – that is, ability to produce the same phenotype, regardless of genotype or environmental modifications - in the regulation of their expression. I also show that the homologues of genes displaying a strong sex-bias in their expression in *C. remanei* are more likely to be missing in selfing species than are expressed genes overall. This suggests that genes with high sex-biased expression contribute disproportionally to reduction in genome size in androdioecious Caenorhabditis species. These results show that the loss of genes expressed with a strong sex-bias is correlated both with loss of traits related to mating and genome size reduction and allow us to shed some light on the consequences of the emergence of a selfing on genome structure.

II. Material and Methods

1. Nematode culture and strains

C. remanei strains PB4641 and EM464, *C. japonica* DF5081, *C. brenneri* PB2801 and *C. elegans* CB4108 *fog-2(q71)* were maintained at 20˚C on NGM agar plates according to standard *C. elegans* methods (Wood 1988), using 2.2% agar to discourage burrowing. To maintain outbreeding and avoid inbreeding depression, large worm populations were maintained by transferring large chunks from two independent plates per passage.

2. RNA extraction

Worms were synchronized by NaOH / bleach treatment and washed off the plates in M9 buffer at various developmental stages, or individually picked in M9 buffer for males, females and pseudo-females preparations. Worms were washed three times in RNase free water and resuspended in 50 µL of RNase free water. 250µL of TRI-Reagent (Molecular Research Center) were added and the samples were frozen at -80˚C. The samples were thawed, pelleted and lysed using a plastic pestle. RNA was purified using phenol/chloroform extraction and subsequent isopropanol precipitation. All of the samples were treated with DNaseI (New England

Biolabs). Phenol/chloroform extraction followed by isopropanol precipitation was performed a second time and the RNA was resuspended in RNase free water.

3. cDNA library and Deep-sequencing

3 biological replicates per sex per species were collected generating a total of 24 samples. cDNA libraries were prepared according to Illumina's mRNA sample preparation protocol at NIDDK, NIH by Harold Smith. Libraries were sequenced on either an Illumina Genome Analyzer II or HiSeq following standard protocol at NIDDK, NIH by Harold Smith as well.

4. Reads alignment to genome and Gene expression quantification

In the case of predicted gene models analysis, the reads were aligned to the genome reference sequences available from WormBase.org (release WS224) using TopHat (Trapnell, Pachter et al. 2009) v. 1.2.0 and the expression level assessed using Cufflinks v. 1.0.2 (Trapnell, Williams et al. 2010) along with .gtf genome annotation files available at the time from ENSEMBL (CJ302 for *C. japonica*, WS220 for *C. elegans*, CB601 for *C. brenneri* and CR2 for *C. remanei*). Both TopHat and Cufflinks were run with the default parameters except for the minimum intron size which was set at 40bp to accommodate for the average reported intron length of *C. elegans* (Cutter, Dey et al. 2009). Gene expression level is expressed in FPKM (Fragments Per Kilobase of transcript per Million mapped reads, Trapnell, Williams et al. 2010), which in the case of paired-end sequencing theoretically allows to measure expression more accurately than RPKM (Reads Per Kilobase of transcript per Million mapped reads, Mortazavi, Williams et al. 2008). Prior to sequencing, cDNA libraries are fragmented, and the number of resulting fragments per cDNA should be proportional to the relative abundance of the cDNA. In the case of paired-end libraries, the second read might be of poor quality and not count towards expression level estimation. FPKM calculation takes into account potential biases resulting from mappability of reads by estimating the number of fragments per cDNA rather than number of reads mapping to a same gene. For all practical purposes, FPKM in this work is comparable to RPKM since we only generated single-end sequencing data.

5. *de novo* Transcriptome Assembly

For each species, the single-end reads from both the 3 male replicates and the 3 female or pseudo-female replicates were used to assemble a transcriptome *de novo* using SOAPdenovo v. 1.05 (Li, Zhu et al. 2010), with a *k*-mer size of 23. In this study, the read-length was 36bp, and single-end sequencing did not allow scaffolding of the generated contigs. I therefore set the minimum contig length at 50bp. The assemblies were further improved as described in other studies (Chen, Yang et al. 2010; Li, Zhu et al. 2010) by removing the contigs at the tip of the de Bruijn graph whose lengths were below 2*k* and/or whose *k*-mer coverages were low. To define 'low coverage' several filtered assemblies were generated for each species with varying parameters.

For each of these assemblies, both the N50 length and contigs were defined (Table 1). The N50 length corresponds to the contig length such that contigs of this size or smaller represent half of the assembly size, and the N50th contig is the number of contigs smaller or equal to N50. Using a minimum *k*-mer coverage of 5 or 10 didn't seem to impact the quality of the assembly as the proportion of the total assembly being smaller than the N50th contig didn't vary. In comparison, increasing the minimum contig length to 100bp improved slightly the assemblies. However in this study rare transcripts are likely to be represented by shorter contigs due to the lack of paired-end information to scaffold contigs. In order to detect expression of such genes as well as that of abundant transcripts, I set the minimum contig length at 50bp, and the minimum *k*-mer coverage at 5. Subsequently, reads from each single replicates were mapped onto the *de novo* transcriptome reference using Bowtie v.0.12.7 (Langmead, Trapnell et al. 2009) and expression level for each contig expressed in RPKM. It should be noted that as a result of the algorithm used to build the assemblies, alternative isoforms might not be represented as full-length transcripts but rather as a set of a full length sequence and smaller sequences corresponding to alternative splice variants.

6. Sex-bias determination

For each species, consistency between replicates of each sex was verified by bootstrapping (Efron and Tibshirani 1993). Only genes or contigs expressed consistently in all three biological replicates were considered for further analyses. Differential expression was assessed using a fixed linear regression model taking into account the effects of sex and biological replication, as well as those of sequencing machine variation (GAII vs. HiSeq), using the *maanova* R package (Wu, Yang et al. 2011). The maximum p-value for significantly differentially expressed contigs or genes were determined by genome-wide permutation tests, and corresponded to a false discovery rate of less than 0.01 (qvalue R package, Dabney, Storey et al. 2011).

Table 1. N50 comparison of assemblies parameters. Numbers in parenthesis are fraction of the total number of contigs.

Species	Minimum contig length	Minimum *k*-mer coverage	Total number of contigs	N50 (bp)	N50th contig
C. japonica	50 bp	**5**	**73,739**	**346**	**62,752 (85.1 %)**
		10	67,081	263	57,158 (85.2 %)
	100 bp	5	29,545	387	23,551 (79.7 %)
		10	28,184	395	22,453 (79.6 %)
C. elegans	50 bp	**5**	**24,965**	**489**	**21,764 (87.2 %)**
		10	23,113	510	20,089 (86.9 %)
	100 bp	5	14,001	590	11,498 (82.1 %)
		10	13,480	602	11,053 (82 %)
C. brenneri	50 bp	**5**	**78,540**	**219**	**66,323 (84.4 %)**
		10	70,504	239	59,827 (84.8 %)
	100 bp	5	32,345	360	25,960 (80.3 %)
		10	29,990	377	24,092 (80.3 %)
C. remanei	50 bp	**5**	**50,841**	**331**	**43,889 (86.3 %)**
		10	46,007	354	39,676 (86.2 %)
	100 bp	5	23,181	457	18,691 (80.6 %)
		10	22,081	467	17,785 (80.5 %)

III. Results

1. RNA-seq analysis

We deep-sequenced 3 biological replicates per sex in *C. japonica, C. elegans, C. brenneri* and *C. remanei* L4 to adults nematodes (Table 2). We used strains used for genome sequencing projects except for *C. elegans* for which the strain used carried the *fog-2(q71)* allele in an otherwise wild-type background. *fog-2(q71)* leads to defective spermatogenesis in XX hermaphrodites, XO males developing normally (Schedl and Kimble 1988). The use of this strain allowed in this study for biologically relevant comparison between *C. elegans* pseudo-females and females of other species.

Table 2. RNA-seq raw data. The number of reads obtained per library are indicated. One library per lane was sequenced. Runs are indicated by numbers in parentheses. Asterisks indicate run on the GAII sequencer, others were run on a HiSeq2000 machine.

		C. japonica	*C. elegans*	*C. brenneri*	*C. remanei*
Female libraries	1	25,090,601 (1)*	26,708,038 (1)*	64,639,719 (1)*	26,354,382 (1)*
	2	52,085,305 (2)	28,711,730 (3)	34,674,541 (3)	51,186,141 (3)
	3	44,162,755 (3)	20,714,660 (2)	34,649,591 (2)	34,100,950 (4)
Male libraries	1	27,788,961 (1)*	23,534,532 (1)*	26,413,000 (4)	21,527,896 (1)*
	2	46,328,649 (2)	27,405,842 (2)	27,910,679 (3)	36,784,213 (3)
	3	40,806,157 (3)	24,957,082 (4)	40,248,635 (2)	26,044,610 (2)

1.1. Typical analytical pipeline

There are several softwares and programs available as of today to analyze RNA-seq data, most of which require a reference sequence onto which the short reads are mapped, as well as gene annotation files to determine the expression level of genes. Both genome sequences and gene annotations are available for several Caenorhabditis species and allowed us to use such an analytical pipeline (Fig. 2a). I used TopHat (Trapnell, Pachter et al. 2009) to map the reads from individual replicates to the appropriate reference sequence and Cufflinks (Trapnell, Williams et

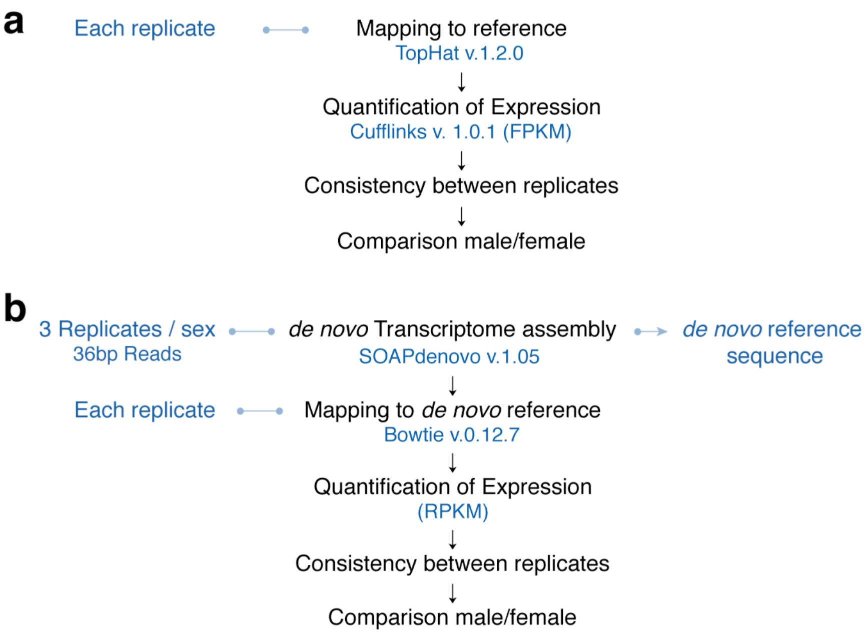

a
Each replicate
Mapping to reference
TopHat v.1.2.0
Quantification of Expression
Cufflinks v. 1.0.1 (FPKM)
Consistency between replicates
Comparison male/female
b
3 Replicates / sex
36bp Reads
de novo Transcriptome assembly
SOAPdenovo v.1.05
de novo reference sequence
Each replicate
Mapping to de novo reference
Bowtie v.0.12.7
Quantification of Expression
(RPKM)
Consistency between replicates
Comparison male/female

Figure 2. Analysis pipeline. **a.** Typical analytical pipeline for RNA-seq data analysis. **b**. Pipeline devised for analyzing RNA-seq data with *de novo* transcriptome assembly.

al. 2010) to determine the expression value of gene predictions. Using this method, the expression of 75.5%, 92.2%, 80.7% and 75.2% of the gene predictions of *C. japonica, C. elegans, C. brenneri* and *C. remanei* respectively was detected (Table 3), and the expression values in the male and female or pseudo-female datasets were compared to determine differential expression (see below). The averages of expression value in the male and female or pseudo-female datasets varied drastically between species, although the ratios of male and female averages were comparable, both for genes whose expression was detected and for the subset significantly differentially expressed between sexes (Table 4).

Table 3. Sex-biased expression in Caenorhabditis gene predictions. The percentages in parentheses are fraction of consistently detected gene predictions.

	C. japonica	*C. elegans*	*C. brenneri*	*C. remanei*
Total number of gene predictions	25,870	20,668	26,037	31,518
Consistently detected gene predictions	19,533	19,050	21,020	23,713
Differentially expressed gene predictions	9,750 (49.9%)	8,534 (44.8%)	6,609 (28.9%)	11,619 (49%)
Female-biased	6,346 (32.5%)	3,444 (18.1%)	2,271 (11.3%)	2,369 (22.6%)
Including strongly female-biased	*622* *(3%)*	*439* *(2.3%)*	*797* *(4.1%)*	*751* *(3.2%)*
Including Female-specific	*24* *(0.1%)*	*18* *(0.1%)*	*122* *(0.6%)*	*29* *(0.1%)*
Male-biased	3,404 (17.4%)	5,090 (26.7%)	3,700 (17.6%)	6,252 (26.4%)
Including strongly male-biased	*2,051* *(9.8%)*	*2,140* *(11.2%)*	*1,775* *(9.1%)*	*2,553* *(10.8%)*
Including Male-specific	*71* *(0.4%)*	*271* *(1.4%)*	*321* *(1.5%)*	*358* *(1.5%)*

The *C. japonica*, *C. brenneri* and *C. remanei* genomes are not assembled or annotated as well as the *C. elegans* genome (Fig. 1, The *C. elegans* consortium 1998;

Hillier, Coulson et al. 2005; Gerstein, Lu et al. 2010). The resulting incomplete transcriptome-wide patterns of expression make direct comparisons between species inaccurate. In order to meaningfully compare transcriptomes and sex-biased expression between these four species, we developed an alternative pipeline (Fig. 2b). I assembled transcriptomes *de novo* using short reads, and after assessing their quality and accuracy, I used them as references to assess transcribed units expression levels in males and females of each species.

Table 4. Distribution of expression value in the TopHat / Cufflinks analysis. The average and median expression values in males and females or pseudo-females of genes whose expression was detected in FPKM and which were differentially expressed are indicated for each species, as well as the ratio of averages.

			C. japonica	*C. elegans*	*C. brenneri*	*C. remanei*
All genes whose expression was detected	Male	Average	249.7	100.9	63.9	131.3
		Median	8.2	10.5	6.7	6.5
	Female	Average	166.6	64.8	48.4	57.9
		Median	9.1	4.6	2.3	2.7
	Ratio Male average /Female average		1.5	1.5	1.3	2.3
Genes with significant differential expression	Male	Average	367.6	135.7	89.6	143.1
		Median	13.8	18.0	10.9	13.0
	Female	Average	174.0	54.9	35.7	54.9
		Median	23.8	7.0	1.2	7.7
	Ratio Male average /Female average		2.1	2.5	2.5	2.6

1.2. SOAPdenovo *de novo* transcriptome assembly

There are a number of ways to assemble transcriptomes *de novo* without a genomic reference (Velvet and Oases (Zerbino and Birney 2008), Trinity (Grabherr, Haas et al. 2011), ABySS (Simpson, Wong et al. 2009), SOAPdenovo (Li, Zhu et al. 2010)). Though SOAPdenovo was designed to assemble genomes, it has been used successfully to assemble transcriptomes without genome reference (Chen, Yang et al.

2010; Adamidi, Wang et al. 2011; Hao, Ge et al. 2011; Wong, Cannon et al. 2011). All these studies used paired-end libraries, for which the read-lengths varied from 50 to 76 bp, and the minimum length of scaffold was set at 100 (Chen, Yang et al. 2010; Adamidi, Wang et al. 2011) or 200 bp (Hao, Ge et al. 2011; Wong, Cannon et al. 2011). In this work, single-ends were sequenced and generated 36bp reads. To try and compensate for the lack of paired-end information, each transcriptome was sequenced at an estimated depth of 185-203X depending on species (Fig. 3a, dotted lines). The number of contigs assembled (Fig. 3a) as well as assembly sizes (Fig. 3b) depended both on the species and on the number of reads used for assembly. The assembly sizes of the final assembly seemed to correlate with genome size, *C. japonica* having the largest genome and the largest assembly size, and *C. elegans* the smallest of both (Compare Fig. 1 and Table 5).

Table 5. SOAPdenovo transcriptome assembly statistics. Transcriptome size is estimated from *C. elegans* exon content (The *C. elegans* consortium 1998)

	C. japonica	*C. elegans*	*C. brenneri*	*C. remanei*
Total number of reads	236,262,428	152,031,884	228,536,165	195,998,192
Estimated transcriptome size	45.9 Mb	27 Mb	40.5 Mb	35.37 Mb
Estimated sequencing depth	185X	203X	203X	199X
Total number of contigs	73,739	24,965	78,540	50,841
Assembly size (bp)	11,694,030	6,161,094	12,204,742	9,473,757
N50 (bp)	246	489	219	331
N50th contig	62,752	21,764	66,323	43,889

The strains of the gonochoristic species used in this study are known to have retained heterozygosity after several generations of inbreeding (Barriere, Yang et al. 2009). In order to assess the extent of heterozygosity in the *de novo* transcriptome assemblies, I used BLASTn to compare each contig sequence to the whole assembly.

While this method does not allow clear discrimination between contigs representing processing variants of the same transcript, alternative alleles or paralogs, it does give an indication of the proportion of the *de novo* assembly representing unique transcript sequences. The sizes of the assemblies without repeated sequences are also consistent with genome sizes, the largest one being *C. japonica* and the smallest *C. elegans* (Table 6).

Table 6. Residual heterozygosity, repeats and gene families. In each assembly, each contig was queried against the assembly using BLASTn. Contigs sharing significant hits (E-value < 10^{-10}) were grouped together, each of these groups representing transcripts from the same allele, derived from different alleles of the same gene and derived from genes of the same gene family. The number of contigs representing unique sequences in each assembly is indicated, as well as the length of the assembly represented by unique sequences.

	C. japonica	*C. elegans*	*C. brenneri*	*C. remanei*
Total number of contigs	73,739	24,965	78,540	50,841
Number of significant hits to other contigs	29,691	5,729	41,682	20,526
Number of groups	13,119	2,207	17,120	8,238
Number of unique sequences	60,620	22,758	61,420	42,603
Total length of unique sequences (bp)	9,433,962	5,558,824	9,009,373	7,840,799

1.3. Exactness of *de novo* assemblies

A near-perfect genome assembly now exists for *C. elegans* (The *C. elegans* consortium 1998; Hillier, Coulson et al. 2005; Gerstein, Lu et al. 2010), and extensive

Table 7. *C. elegans* SOAPdenovo transcriptome BLASTn statistics.

E-value	Number of SOAPdenovo contigs with a hit in the *C. elegans* predicted genes	Fraction of total number of contigs	Number of *C. elegans* predicted CDS hit
0	4,528	18.1%	3,143
10^{-10}	20,647	82.7%	6,481
10^{-5}	21,234	85.05%	6,570
10^{-2}	22,047	88.31%	7,272
10^{-1}	22,843	91.5%	8,058

curated annotations are available. As such, it provides a good reference to both test the accuracy of the *de novo* assemblies and estimate the performance of using a *de novo* transcriptome assembly to that of using existing genome annotations. The *de novo* assembled contigs were aligned to the *C. elegans* predicted CDS database using BLASTn. 82.7% of the contigs were attributed to predicted *C. elegans* CDS by BLASTn, irrespective of their sizes (E-value < 10^{-10}, Fig. 4a, also see Table 7). In addition, the BLASTn alignments of significant hits (E-value < 10^{-10}) encompassed most of the contig lengths (Fig. 4b). The *C. elegans* SOAPdenovo *de novo* assembly hence represented at least 6,481 *C. elegans* gene predictions (Table 7)

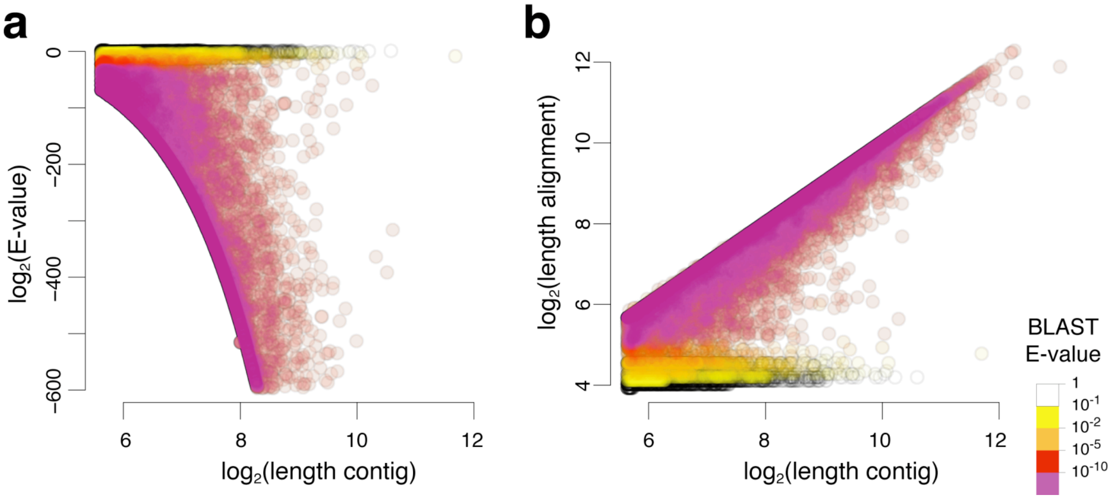


Figure 4. *C. elegans de novo* assembly accuracy. *C.elegans* SOAPdenovo contigs were aligned against the set of *C. elegans* predicted genes using BLASTn. SOAPdenovo contig length is plotted against BLASTn alignment E-value (a**)**, or against BLASTn alignment length (**b**). Hits for which the BLASTn E-value was below 10^{-10} are plotted in purple, between 10^{-10} and 10^{-5} in red, between 10^{-5} and 10^{-2} in orange and between 10^{-2} and 10^{-1} in yellow.

1.4. Detecting expression in *de novo* transcriptome assemblies

The reads from each replicate were mapped to the corresponding species transcriptome *de novo* assembly using Bowtie (Langmead, Trapnell et al. 2009) and the expression values for each contig were calculated in RPKM. Only contigs expressed consistently in all replicates of at least one sex were considered for further analyses. 97 to 99% of the *de novo* assemblies were expressed consistently in one or both sexes (Table 8). Their expression values were compared between sexes to determine which were significantly sex-biased. As in the case of expression levels of gene predictions described above, the averages of expression values of contigs in males and females or pseudo-females was variable depending on species but the ratio of averages was similar across species, both for contigs whose expression was detected and differentially expressed (Table 9).

Table 8. Sex-biased expression in Caenorhabditis *de novo* transcriptome assemblies. The percentages in parentheses are fraction of consistently detected SOAPdenovo contigs.

	C. japonica	*C. elegans*	*C. brenneri*	*C. remanei*
Total number of contigs	73,739	24,965	78,540	50,841
Consistently detected contigs	72,956	24,399	76,496	50,427
Differentially expressed contigs	30,802 (42.2%)	11,394 (46.7%)	22,930 (30%)	23,767 (47.1%)
Female-biased	17,781 (24.4%)	5,061 (20.7%)	8,926 (11.7%)	11,817 (23.4%)
Including strongly female-biased	*3,375* *(4.6%)*	*825* *(3.4%)*	*3,756* *(4.9%)*	*2,664* *(5.3%)*
Including Female-specific	*500* *(1.6%)*	*138* *(1.2%)*	*530* *(2.3%)*	*570* *(1.1%)*
Male-biased	13,021 (17.8%)	6,333 (26%)	14,004 (18.3%)	11,950 (23.7%)
Including strongly male-biased	*5,989* *(8.2%)*	*3,011* *(12.3%)*	*5,532* *(7.2%)*	*4,184* *(8.3%)*
Including Male-specific	*3,238* *(4.4%)*	*836* *(3.4%)*	*6,675* *(8.7%)*	*4,601* *(9.1%)*

Table 9. Distribution of expression value in the SOAPdenovo assemblies analysis. The average and median expression values in males and females or pseudo-females of QTUs whose expression was detected and which were differentially expressed are indicated for each species, as well as the ratio of averages.

			C.japonica	*C. elegans*	*C. brenneri*	*C. remanei*
All genes whose expression was detected	Male	Average	85.8	209.4	76.6	112.5
		Median	25.0	95.0	25.9	34.8
	Female	Average	74.6	182.4	69.0	97.0
		Median	31.5	88.8	22.4	34.6
	Ratio Male average /Female average		1.2	1.1	1.1	1.2
Genes with significant differential expression	Male	Average	107.5	219.5	90.1	121.9
		Median	27.7	94.5	29.6	37.3
	Female	Average	81.0	161.8	60.4	92.1
		Median	33.4	60.7	1.5	25.9
	Ratio Male average /Female average		1.3	1.4	1.5	1.3

1.5. Comparison of analytical pipelines

I used the *C. elegans* data to compare the performances of the TopHat / Cufflinks or SOAPdenovo/Bowtie analytical pipelines. This former method allows one to use or not use a genome annotation file, which contains information pertaining to exons coordinates within the genome sequence. Without the gene annotation file, gene models are build by Cufflinks based on FPKM values of the different expressed fragments. Both possibilities were explored, and compared to the performance of the SOAPdenovo / Bowtie pipeline. While the actual numbers of contigs, Cufflinks gene models or annotated genes varied, the proportions of consistently detected Quantifiable Transcribed Units (QTUs) they represented were comparable to each other. No significant differences in numbers and proportions of QTUs with significant expression differentials were found between using SOAPdenovo / Bowtie or TopHat / Cufflinks in *C. elegans* (Table 10). However, the numbers of gene models defined by

Cufflinks, as well as the proportion of consistently detected QTUs they represented did not correspond to current *C. elegans* gene predictions, nor to the numbers obtained with SOAPdenovo and Bowtie (Table 10) suggesting that Cufflinks gene models are inaccurate.

Table 10. Comparison of analysis pipelines in *C. elegans*. The numbers represent numbers of QTUs for each analysis pipeline. The percentages in parentheses are fraction of consistently detected QTUs.

Species	*C. elegans*	*C. elegans*	*C. elegans*
Analytical pipeline	SOAPdenovo / Bowtie	TopHat / Cufflinks	TopHat / Cufflinks
Genome annotation file	N/A	No	Yes
Consistently detected	24,399	23,353	19,050
Differentially expressed	11,394 (46.7%)	7,072 (30.3%)	8,534 (44.8%)
Female-biased	5,061 (20.7%)	2,693 (11.5%)	3,444 (18.1%)
Including Female-specific	*138 (0.6%)*	*30 (0.1%)*	*18 (0.1%)*
Male-biased	6,333 (26%)	4,379 (18.7%)	5,090 (26.7%)
Including Male-specific	*836 (3.4%)*	*278 (1.2%)*	*271 (1.4%)*

In contrast, 97.1% of the SOAPdenovo contigs whose expression is consistently detected have a significant BLASTn hit whose expression is consistently detected as well (Fig. 5). Further, among those pairs in which expression of both SOAPdenovo contig and BLASTn hit are detected, 89.9% of the SOAPdenovo contigs differentially expressed between males and females were also found to be significantly differentially expressed when using the TopHat / Cufflinks analysis pipeline (Table 11). Altogether, this indicated that the vast majority of the SOAPdenovo *C. elegans* transcriptome assembly was accurate and represented legitimate QTUs. Moreover it validated the use of a similar analysis pipeline for the other Caenorhabditis species in this study for which a genome reference sequence is not available, rather than using Cufflinks gene models which depend on incomplete genome reference sequences and inaccurate gene predictions capacity.

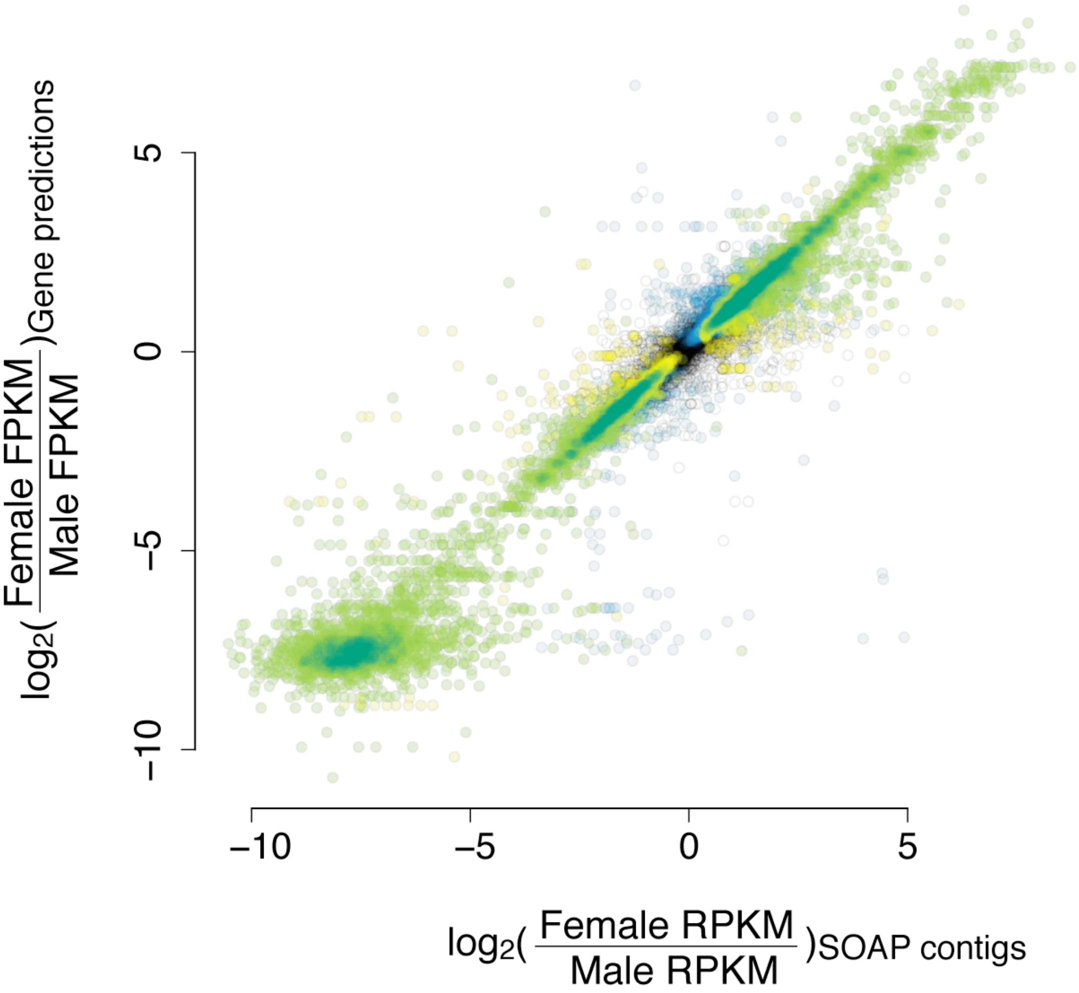


Figure 5. Sex-biased expression of SOAPdenovo contig vs gene predictions in *C. elegans*. For each contig assembled by SOAPdenovo with a significant BLASTn hit (E-value < 10^{-10}), the value of the expression differential between the female and male datasets is plotted against the value of the expression differential of its corresponding best BLASTn hit in the *C. elegans* set of predicted genes as defined by theTopHat / Cufflinks analysis. Yellow dots represent QTU pairs in which the expression of the SOAPdenovo contig is significantly sex-biased, blue dots those for which the expression of the predicted gene is sex-biased. Transparency settings were used such that green (and greener) dots represent pairs for which both QTUs are significantly sex-biased.

Table 11. *C. elegans* SOAPdenovo transcriptome and TopHat / Cufflinks comparison. The number of pair of contigs/corresponding BLASTn hits in each category are indicated

BLASTn E-value Threshold	1	10^{-10}
Detected in both pipelines	23,091	20,060
Total in SOAPdenovo pipeline	23,662	20,331
Total in Cufflinks/Tophat pipeline	23,612	20,372
Differentially expressed in both pipelines	9,646	8,991
Total differentially expressed in SOAPdenovo pipeline	11,180	10,003
Total differentially expressed in Tophat / Cufflinks pipeline	13,555	11,802

2. Sex-biased expression in Caenorhabditis

In the case of gene predictions, 29 to 50% of the detected genes displayed a significant expression differential between sexes depending on the species. Most of these were male-biased, except in *C. japonica* where more genes had a female-biased expression (Table 3). The same trends were observed with the *de novo* assemblies, where the significantly sex-biased fraction of the transcriptomes of *C. japonica, C. elegans, C. brenneri* and *C. remanei* represented 30%, 46.7%, 42.2% and 47.1% of the detected contigs, respectively (Table 8).

Similarly, in both analyses QTUs with strong male-biased expression (defined as greater than 10-fold) displayed a wider range of expression values than those with a strong female-biased expression in all four species (gene predictions, Fig. 6 and *de novo* assemblies, Fig. 7). QTUs exhibiting strong expression differentials between sexes corresponded to different proportions of QTUs whose expression was detected depending both on sex and species. Those with a strong male-bias represented 9.5 to 12.7 % of genes (Fig. 6) and 12.7 to 17.4% of contigs (Fig. 7) while 2.3 to 3.6% of genes (Fig. 6) and 4 to 6.4% of contigs (Fig. 7) showed a strong female-biased expression.

C. elegans displayed the lowest fraction of QTUs with strong female-biased expression in both analyses. In addition, the pattern of male-biased QTUs was strikingly different between *C. elegans* and the other species irrespective of the way the analysis was performed (Fig. 6 and 7). The range of expression value ratios of the highly male-biased QTUs was more limited in *C. elegans* than in the other species (Fig. 8). In order to define whether this was due to higher expression in pseudo-females or lower expression in males, I compared the expression values in males and pseudo-females or females of *C. elegans* and *C. remanei* respectively of a subset of strict orthologous genes with a strong male-biased expression in both species. 399 such genes were identified, and the average of their expression values is lower in *C. elegans* males than in *C. remanei* males. In addition, the average of their expression values is higher in *C. elegans* pseudo-females than in *C. remanei* females (Table 12).

Table 12. Expression values of genes with a strong male-bias. 399 strict orthologous pairs of *C. elegans* and *C. remanei* genes with a strong male-biased expression in both species were identified. The average and median expression value in males and pseudo-females or females of these genes are indicated, as well as the ratio of averages.

		C. elegans	*C. remanei*
Male	Average FPKM	190.6	353.6
	Median FPKM	63.6	75.4
Female	Average FPKM	2.3	0.6
	Median FPKM	0.5	0.2
Ratio Male/Female averages		82.9	589.3

Lastly, in the SOAPdenovo / Bowtie analysis, *C. elegans* displayed a marked decrease in numbers of contigs with sex-biased expression as well as those both with male- and female-specific expression compared to *C. japonica, C. brenneri* and *C. remanei*. These trends did not hold when considering fractions of consistently detected contigs, although the fraction of contigs with male-specific expression was the lowest in *C. elegans* (Table 8).

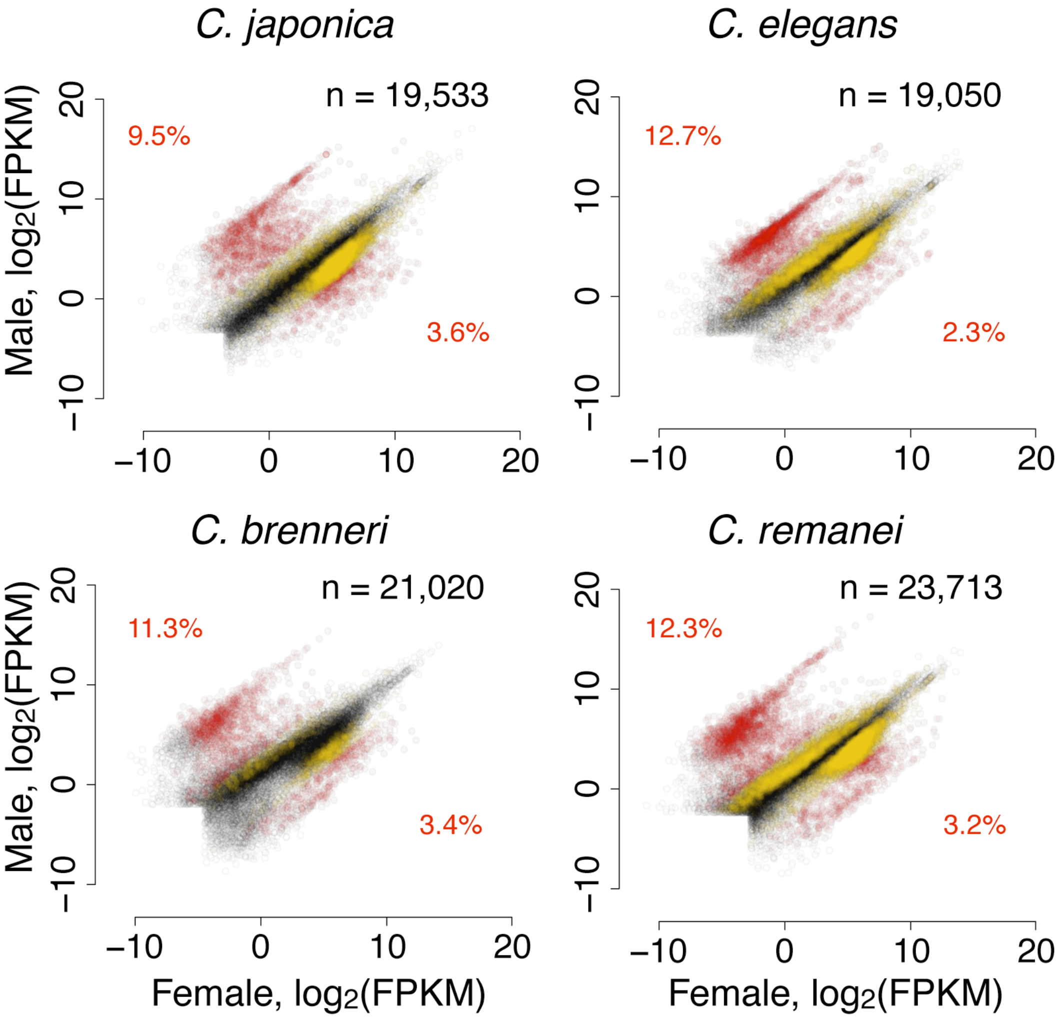


Figure 6. Sex-biased expression as measured using TopHat and Cufflinks. Plot of the expression values for each detectably expressed predicted gene in male against female datasets in *C. japonica*, *C. elegans*, *C. brenneri* and *C. remanei*. Expression values are expressed as $\log_2$ of FPKM. Significantly differentially expressed predicted genes are plotted in red, those displaying a difference over 10-fold between sexes in yellow. The total number of predicted genes whose expression was detected is indicated on the top right corner, and the fractions of consistently expressed predicted genes that are highly male or female-biased, including those that were male- or female-specific, are indicated respectively in the top left and bottom right corner.

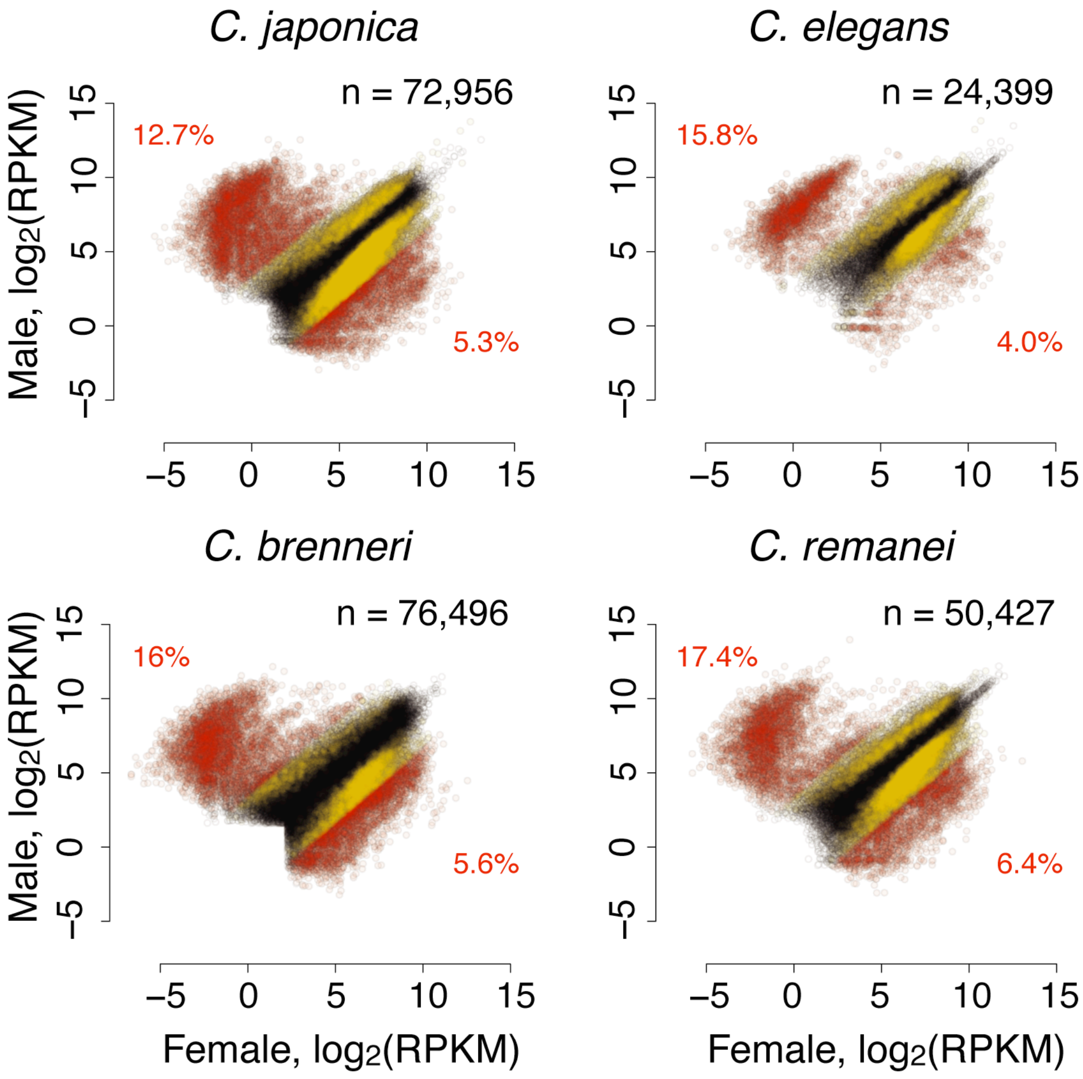


Figure 7. Sex-biased expression as measured with cDNA contigs. Plot of the expression values for each SOAPdenovo contig in male against female datasets in *C. japonica*, *C. elegans*, *C. brenneri* and *C. remanei*. Expression values are expressed as $\log_2$ of RPKM. Significantly differentially expressed contigs are plotted in red, those displaying a difference over 10-fold between sexes in yellow. The total number of contigs which were detected is indicated on the top right corner, and the fractions of consistently expressed contigs that are highly male or female-biased, including those male- or female-specific, are indicated respectively in the top left and bottom right corner.

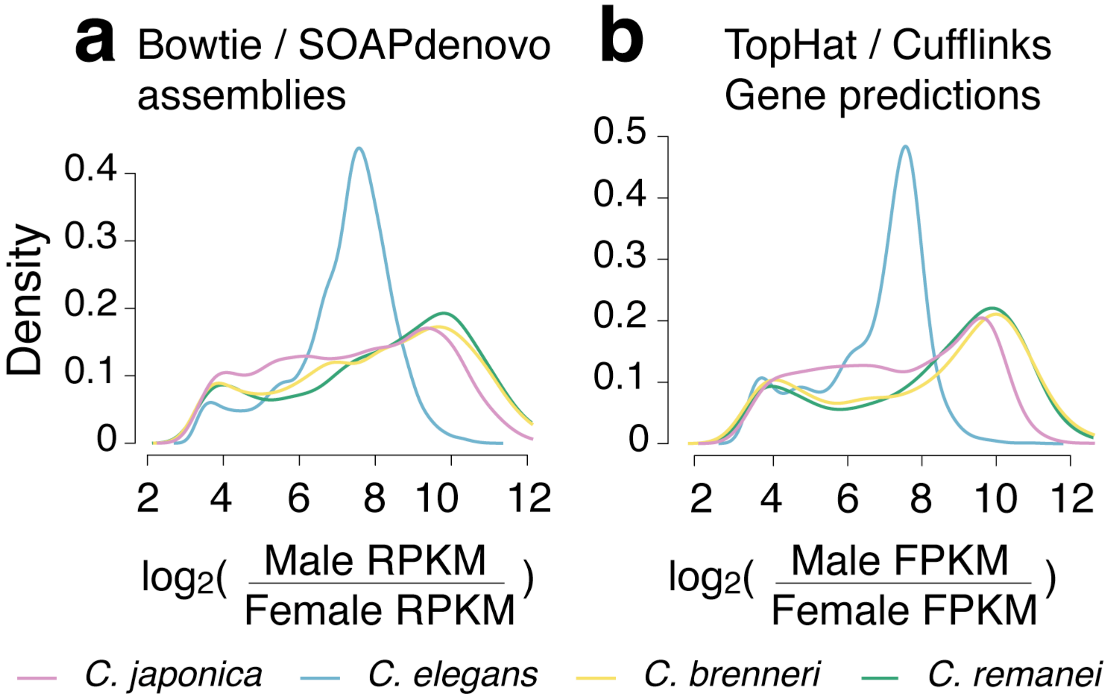


Figure 8. Degree of sex bias in highly male-biased subset. Density plots corresponding to the distributions of the ratios of male expression values to female expression values for *C. japonica* (purple), *C. elegans* (blue), *C. brenneri* (yellow) and *C. remanei* (green) for the SOAPdenovo assembly analysis (**a**) as well as for the TopHat / Cufflinks analysis of gene predictions (**b**).

3. Conservation of sex-biased expression

The decreased number of QTUs with sex-biased expression observed in *C. elegans* could be explained by disproportionate loss during genome shrinkage or by massive gene expression pattern modification of conserved genes. Testing either of these hypotheses requires defined homology relationships between genes of the Caenorhabditis species, something *de novo* transcriptome assembly would not allow. Given the phylogeny of the genus and the status of current gene predictions of Caenorhabditis genomes, *C. remanei* was chosen as a representative of gonochoristic species to explore conservation in selfing species of genes expressed in a sex-biased fashion in *C. remanei*.

Presence or absence of homologues of *C. remanei* sex-biased genes were assessed in *C. japonica, C. elegans, C. brenneri* and *C. briggsae* using a set of orthologues defined by the modENCODE group (Brian Oliver, pers. comm.), based on the WS220 release of WormBase (Wormbase.org). *C. remanei* genes with sex-biased and strongly sex-biased expression were compared to all *C. remanei* genes whose expression was detected. Genes categorized as missing in one or both selfing species had a homologue in at least one other gonochoristic species genome, and thus were lost in at least one hermaphrodite lineage (see Fig. 1). In addition, the expression of 2,293 of the 3,503 *C. remanei* genes missing in both selfing species was detected, and 1,056 of them are significantly differentially expressed. The average of their expression values corresponded to that of *C. remanei* detected genes overall, with the exception of the average expression value in males of genes missing in both selfing species and differentially expressed which was higher (Table 13). Genes defined as *C. remanei* specific genes had no defined homologues in neither species. Genes with sex-biased expression tended to be more conserved compared to overall detected genes, whereas homologues of highly-sex biased genes were significantly

Table 13. Expression values of *C. remanei* genes without homologues in both selfing species. The average and median expression value in males and females of *C. remanei* genes with no homologues in neither *C. elegans* nor *C. briggsae* and whose expression were detected or significantly sex-biased are indicated, as well as the ratio of averages.

		Detected *(n = 2,293)*	Differentially expressed *(n = 1,056)*
Male	Average FPKM	151.9	265.8
	Median FPKM	4.2	10.0
Female	Average FPKM	49.3	42.1
	Median FPKM	1.3	3.4
Ratio Male/Female averages		3.0	6.3

more likely to be missing in one or both selfing species (Fig. 9a, Fisher's exact test, $p < 0.001$). Both highly female- and male-biased transcripts seemed to contribute to the loss of genes whose expression was highly sex-biased in one (Fig. 9a, Fisher's exact test, $p < 0.001$ and $p < 0.05$ respectively) or both selfing species (Fig. 9a, Fisher's exact test, $p < 0.001$ and $p < 0.01$ respectively). Genes specific to the *C. remanei* lineage displayed no significant patterns of over-representation except for genes whose expression was highly male- or female-biased (Fig. 9a, Fisher's exact test, $p < 0.05$ and $p < 0.001$ respectively, see below).

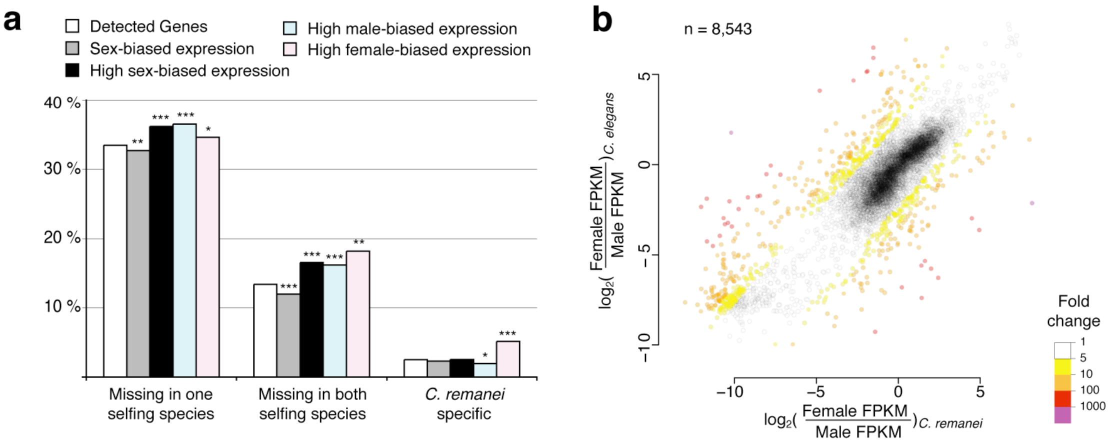


Figure 9. Conservation of *C. remanei* genes with sex-biased expression. a. Comparison of the patterns of conservation of *C. remanei* genes whose expression is detected (white), sex-biased (grey), highly sex-biased (over 10 fold, black), highly male-biased (blue) or highly female-biased (pink). The fraction of genes with a homologue in at least one other gonochoric species but none in one or both of the selfing species are represented, as well as the *C. remanei* specific genes. Significance of difference to detected number of genes was assessed by Fisher's exact test (* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$). **b**. Plot of the expression differential of *C. elegans* genes against that of their *C. remanei* homologues. Only genes for which both homologues are detected in both sexes are represented. Genes for which the fold change between expression differential of each homologue is between 5 and 10 are plotted in yellow, between 10 and 100 in orange, between 100 and 1000 in red and over 1000 in purple.

In order to address whether modification of conserved gene expression contributed to the differences in sex-bias observed between *C. elegans* and

C. remanei, the expression of *C. remanei* and *C. elegans* homologous genes. Strict orthologues (1:1) were identified using the modENCODE homology set. 8,543 such genes were identified, 94.3% of which displayed the same sex-bias in their expression in both species and this to a comparable degree (Fig. 9b).

4. GO term analysis

56.4% of the genes whose expression was detected in *C. remanei* were associated to GO terms (Ashburner, Ball et al. 2000, Fig. 10). *C. remanei* genes whose homologues were missing in both selfing species, as well as those with sex-biased expression tended to be significantly more associated with GO terms than expected by chance (Fig. 10, Fisher's exact test, $p < 0.01$ and $p < 0.001$ respectively), as opposed to genes with highly sex-biased expression (Fig. 10, Fisher's exact test, $p < 0.001$). Two GO categories represented at a high frequency were integral membrane proteins and those related to retrotransposons functions. Genes encoding predicted membrane proteins were under-represented in genes whose expression was sex-biased (Fig. 10, Fisher's exact test, $p < 0.001$) or whose homologues were missing in both selfing species (Fig. 10, Fisher's exact test, $p < 0.001$) but were over-represented in genes whose expression was highly sex-biased (Fig. 10, Fisher's exact test, $p < 0.001$). This over-representation was mostly due to high male-biased expression (Fig. 10, Fisher's exact test, $p < 0.001$). The over-representation of GO terms related to retrotransposons was observed only for genes lost in selfing species, including those whose expression was sex-biased (Fig. 10, Fisher's exact test, $p < 0.001$ for both).

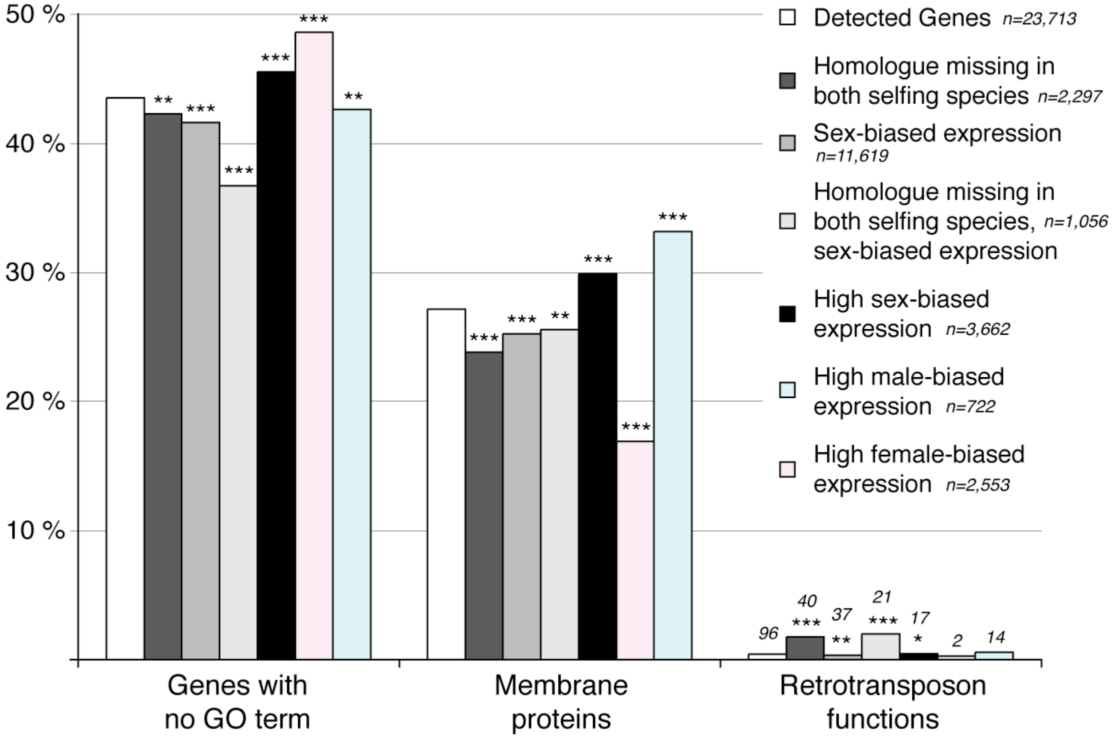


Figure 10. GO term analysis. The fraction of genes associated to GO terms, predicted to be integral to membrane and whose GO terms pertained to retrotransposon functions are indicated for all *C. remanei* genes whose expression was detected (white), sex-biased (grey), highly sex-biased (black), highly female-biased (pink) and highly male-biased (blue), as well as for genes whose homologues were missing in both selfing species (dark grey) and sex-biased (light grey). Significance of difference to detected number of genes was assessed by Fisher's exact test (* p < 0.05; ** p < 0.01; *** p < 0.001). Number of genes aare indicated in italic.

5. Sex-bias in lineage-specific genes

The modENCODE set of homologues identifies 424, 283, 548 and 426 genes found only in *C. japonica, C. elegans, C. brenneri* and *C. remanei* respectively. The distribution of their sex-bias was compared to that of genes whose expression was detected and which were part of the modENCODE homology dataset. *C. elegans*-specific genes did not seem to be significantly different than *C. elegans* genes overall, while in the other species the species-specific genes tended to have strongly female-biased expression (Fig. 11, χ^2 test, p < 0.05 in *C. japonica* and p < 0.001 in *C. brenneri* and *C. remanei*). In addition, *C. japonica*-specific genes were less likely to be male-specific, strongly male-biased or male-biased in their expression (Fig. 11, χ^2 test, $p < 0.05$, $p < 0.05$ and $p < 0.01$ respectively), genes with a strong male-biased expression were under-represented among genes specific to *C. brenneri* (Fig. 11, χ^2 test, $p < 0.01$) and those with male-biased expression among genes specific to *C. remanei* (Fig. 11, χ^2 test, $p < 0.001$).

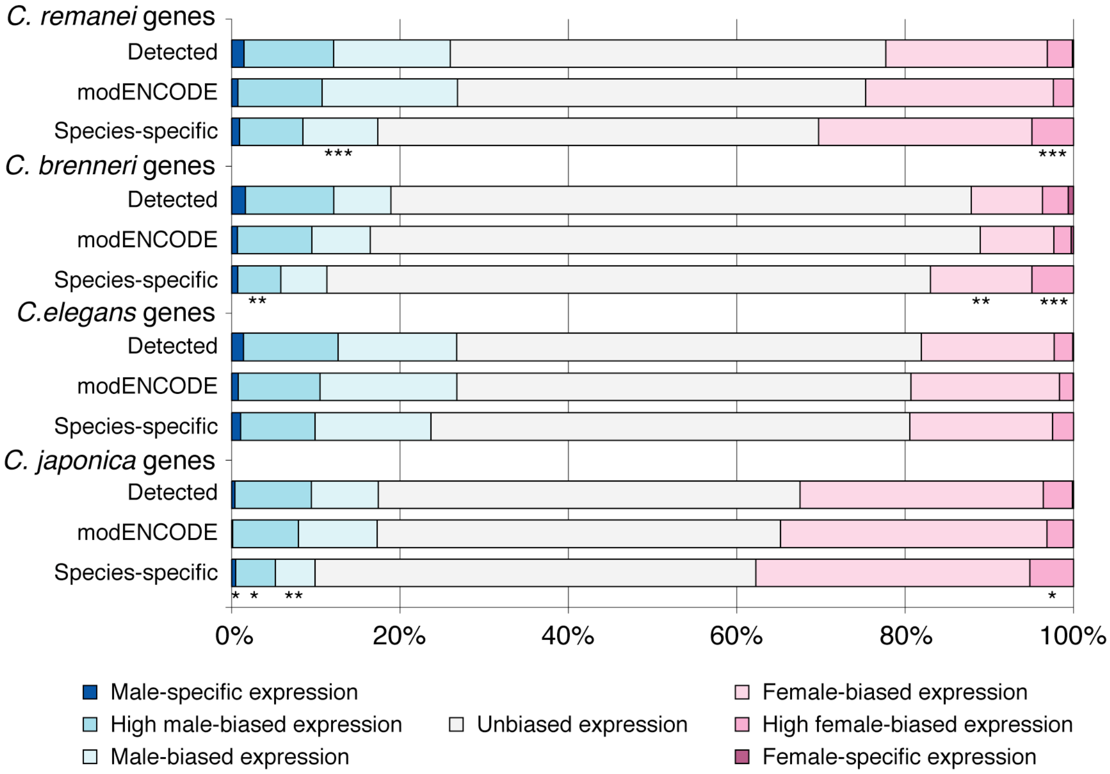


Figure 11. Distribution of species-specific genes expression. The fraction of genes male-specific (dark blue), highly male-biased (blue), male-biased (light blue), female-biased (light pink), highly female-biased (pink), female-specific (dark pink) or not biased (grey) in their expression are indicated for all genes whose expression was detected, the subset of genes part of the modENCODE homology dataset and genes among those found to be species-specific. Significance of difference between all detected genes in the modENCODE dataset and lineage-specific genes was assessed by χ^2 test (* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$).

6. Characterization of previsouly identified candidate genes

In Chapters 1 and 2, a set of *C. remanei* candidate genes was identified based on their expression bias and phylogenetic conservation pattern in the Caenorhabditis genus and/or predicted functions. Analysis of the larger dataset confirmed that all candidate genes but one, CRE23534, were expressed in a significantly male-biased fashion in *C. remanei*. CRE23534 expression was male-biased but the p-value associated with its expression values was slightly above the threshold of significance ($p = 0.037$, for $p_{\text{threshold}} = 0.032$). The expression of 91.5% of the VRM genes was detected in *C. remanei*, and 90.7% of them are significantly male-biased in their expression. In addition, the expression of most of the homologues of the candidate genes in *C. japonica*, *C. elegans*, and *C. brenneri* could be determined and all of them were significantly male-biased. In the case of the VRM family, the expression of all 15 *C. japonica* paralogues, 8 out of 10 *C. elegans* paralogues and 75 out of 97 *C. brenneri* paralogues whose expression was significantly male-biased.

IV. Discussion

1. Transcriptome analysis

1.1. Gene predictions and Genome sequences

Thanks to next-generation technologies, transcriptomes can be sequenced at depths allowing complete surveys of gene expression in a sample, or accurate assessment of differential expression between several (reviewed in Wang, Gerstein et al. 2009; Oshlack, Robinson et al. 2010; Salzberg 2010). Accurate interpretation of RNA-seq data however depends immensely on the reference sequence used to map the reads, as well as on the completeness of the gene annotations used to estimate gene expression levels. This study illustrated this point, as the expression of most of the (high quality) *C. elegans* gene predictions was detected, whereas only 75 to 80 %

of *C. japonica, C. brenneri* and *C. remanei* predictions were. Insofar as the RNA sequenced represented L4 and adults, genes expressed at earlier stages of development were not expected to be detected.

The genome of *C. elegans* has been sequenced completely and is virtually gap-free (Hillier, Coulson et al. 2005). In addition, most gene predictions are supported by expression data (Gerstein, Lu et al. 2010). In contrast, genome sequences of gonochoristic species not only retained allelic variants (Barriere, Yang et al. 2009), they are still at the draft stage and gene models are mostly unsupported *in silico* predictions. Because all four species were sampled at similar depth, the fact that the expression of proportionally fewer gene predictions is detected in gonochoristic species could reflect the inaccuray and incompleteness of both their genome sequences and corresponding gene annotations. As a result, while the interpretation of gene expression in *C. japonica, C. brenneri* and *C. remanei* using this method would be exact, it would not represent accurately the whole transcriptome.

1.2. *de novo* transcriptome assemblies

In order to circumvent the need for genome-dependent analysis, alternative approaches have been developed that rely on very deep sequencing of cDNA libraries. These use RNA-seq raw data to generate assemblies that can be used as references to assess expression levels of transcribed units. The most powerful approaches exploit paired-end reads to assemble contigs into scaffolds (reviewed in Martin and Wang 2011). In the present study, the use of single-end reads limited the efficacy of assembling algorithms, but the depth of sequencing (180-203X) allowed the generation of *de novo* transcriptomes for each species.

The inclusion of *C. elegans* as a best-case control in this analysis enabled a thorough comparison of expression data derived from the *de novo* transcriptome assembly to that of curated reliable gene predictions. The proportion of QTUs differentially expressed were similar for each species in both analyses, confirming the biological relevance of the *de novo* assembly. However, one undesirable feature was that highly expressed genes were more likely to be represented by contigs in the *de novo* transcriptome assemblies (data not shown). This limits the otherwise reference-independent perspective gained by using *de novo* assemblies to those 20% of genes with the highest expression levels. While unfortunate, the same artefact applies to all four species, and the partial transcriptomic expression patterns obtained are still comparable between species. In addition, the TopHat / Cufflinks analysis offers an alternative view on the transcriptomes, which albeit incomplete, allows estimation of expression level of most predicted genes.

Among *de novo* transcriptomes of different species, contigs in *C. elegans* tended to be longer than those in gonochoristic species, as demonstrated by a larger N50 length. The high sequencing depth, and near-complete homozygosity in *C. elegans* are both likely contributors to this effect. Consistent with this, the average expression level of *C. elegans* contigs was invariably higher than in gonochoristic species (Table 9). However, the ratio of average expression in males to that of average expression in females or pseudo-females was very similar across species. This indicated that while absolute expression values were highly dependent on sequencing depth and *de novo* transcriptome fragmentation, ratio of expression values were comparable between species, and much less variable than in the case of gene predictions (compare Table 9 to Table 3).

1.3. Genome and transcriptomes sizes

In selfing *A. thaliana* and obligate outcrosser *A. lyrata*, with genome sizes of respectively 125Mb and 230Mb (Johnston, Pepper et al. 2005), most of the difference in genome size is due to small indels in non-coding DNA and transposons (Hu, Pattyn et al. 2011). In addition, TEs are less numerous in *A. thaliana*, and were found to be more active and of more recent origins in *A. lyrata*.. The number of gene predictions is also smaller in *A. thaliana* than in *A. lyrata*, as is the case in the Caenorhabditis.

The genome sizes of a number of Caenorhabditis nematodes have been measured directly by flow cytometry (Thomas, Li et al. In prep.), and the three selfing species in this genus have consistently smaller genomes than their gonochoristic congeners. Interestingly both the Caenorhabditis and the Arabidopsis genera are contradicting the theory that genome sizes of selfing species should increase overtime because of accumulation of repeats and selfish DNA (Lynch and Conery 2003). The repeat content of genomes of androdioecious Caenorhabditis species is consistently smaller than that of gonochoristic species (see Introduction, Table 1, Caenorhabditis Repeat Libraries), and accounts for about 1 Mb of the difference in genome size between *C. elegans* and *C. remanei*.

Interestingly, the average intergenic distances are similar across Caenorhabditis species (Table 14), as are average gene lengths (*C. elegans* 1.9kb (Lynch 2007), *C. remanei* 2.2kb, *C. brenneri* 1.9kb (CGT, pers. obs.), suggesting that the difference in genome size might be mainly due to number of genes and TEs. In addition, according to modENCODE data (B. Oliver, pers. comm.) 4,153 among the *C. remanei* genes whose expression was detected are missing or lost in *C. elegans* – that is, they have no homologues in *C. elegans* and have one in either *C. japonica* or *C. brenneri*. Given an average gene length in *C. remanei* of 2,242 bp (based on 795 genes, CGT, pers. obs.) and the average intergenic distances, these 4,153 genes explain around 19.4Mb more of the difference in genome sizes between *C. elegans* and *C. remanei*.

Table 14. Intergenic distances in Caenorhabditis. The number of genes upon which calculations were based are indicated.

	C. elegans	*C. brenneri*	*C. remanei*
Number of genes	314	571	795
Average intergenic distance (bp)	2,444	2,115	1,791
Median intergenic distance (bp)	1,115	922	704

I note, however, that because genome assemblies and annotations for gonochoristic species are more preliminary than those of selfing species, the current numbers of gene predictions in these species are likely to be wrong. It is therefore rather risky to compare the total predicted haploid gene count across the genus. With this work, I provide an alternative approach. More gene predictions were expressed in males and females of gonochoristic species than in *C. elegans* (Table 2). In addition, the sizes of gonochoristic *de novo* transcriptome assemblies were larger than that of *C. elegans* both in terms of number of contigs and in terms of base pairs. Further, this relationship held true for the unique portions of the *de novo* assemblies, confirming that *C. japonica, C. brenneri* and *C. remanei* have larger transcriptomes, and susbsequently more genes than *C. elegans,* and indicating that the genome shrinkage observed in selfing species was in part due to loss of genes.

The fraction of the *de novo* cDNA contigs falling into families of highly similar sequences was much smaller in *C. elegans* than in the other species, reflecting higher heterozygosity and potentially larger gene families in the gonochoristic species. Another source of repeated sequences are transposable elements, and expression data as well as GO term analysis suggested in this study that there might be more TEs expressed in gonochoristic species than in *C. elegans*. However, this trend should be considered with extreme caution until further verification. While gene predictions corresponding to TEs are systematically removed from the *C. elegans* genome annotation, the effort has not been extended to those of other species yet. Better genome sequence information would be required in order to sort through those hypotheses, and would also allow better understanding of TE content in Caenorhabditis genomes.

2. Sex-biased expression in Caenorhabditis

2.1. General trends

The advent of technologies allowing genome-scale characterization of gene expression has enabled powerful comparisons between samples and led to complete studies of sex-biased expression in several organisms (see review by Ellegren and Parsch 2007). Most of these studies made use of micro-arrays, and often experimental designs and statistical analyses used to define the threshold for sex-bias vary drastically from study to study. Nevertheless, differential expression between sexes has been measured in a variety of organisms and tissues and was found to range from 4 to 90% of detected genes (Rinn, Rozowsky et al. 2004; Ayroles, Carbone et al. 2009). Among closely related species of Drosophila, the fractions of detected genes with sex-biased expression were mostly similar across species (Zhang, Sturgill et al. 2007). In this study the same trend was true in four closely related Caenorhabditis species : 47 % of the QTUs on average had a significantly sex-biased expression in *C. japonica, C. elegans* and *C. remanei* (Table 2 and 7). Surprisingly this figure fell to 30% in *C. brenneri* and seemed to be due to an increase in the number of genes with unbiased expression rather than a decrease in the number of those with sex-biased expression. In the case of *C. remanei* the proportion of differentially expressed QTUs was higher than observed in our preliminary survey (Chapter 1). However the preliminary *C. remanei* transcriptome analysis was not as thorough as this study, using only one biological replicate and lower coverage of the transcriptome.

Interestingly, the patterns of sex-biased expression were also somewhat inconsistent between studies. While the preliminary survey (Chapter 1) and most studies found more genes with female-biased expression in *C. elegans* (Reinke, Gil et al. 2004), several species of Drosophila (Ayroles, Carbone et al. 2009), mosquitoes (Hahn and Lanzaro 2005), sticklebacks (Leder, Cano et al. 2010), and mice (Yang, Schadt et al. 2006), I detected opposite trends – that is more male-biased expression overall - in *C. brenneri*, *C. elegans* and *C. remanei* regardless of the analytical method used, but confirmed it for *C. japonica*. This might solely be due to the statistical treatment of the data and the significance threshold, as the spread of expression values in each sex were very similar across species in this study and others: QTUs with male-biased expression displayed a much wider range of expression values than those with female-biased expression in all four species using either analytical process.

2.2. Lineage-specific genes

Genes with expression biased towards the hetero-gametic sex are often found to evolve faster than other genes (mammals (Torgerson, Kulathinal et al. 2002; Schultz, Hamra et al. 2003; Good and Nachman 2005; Khaitovich, Hellmann et al. 2005), flies (Clark, Eisen et al. 2007), flies and mammals (Meisel 2011), nematodes (Cutter and Ward 2005), Chicken (Mank, Hultin-Rosenberg et al. 2007)). In addition genomic comparative studies between *C. elegans* and *C. briggsae* showed that sperm genes were more divergent between species and had a greater number of species-specific paralogues. In line with this, I found that more species-specific genes had strong male-biased expression than strong female-biased in *C. elegans, C. brenneri* and *C. remanei*. However in all three gonochoristic species there are more species-specific genes with female-biased expression than with male-biased expression. There are also more genes with strong female-biased expression than expected by chance, while there were no differences in the distribution of sex-biased expression among genes specific to *C. elegans*. This indicates that the trends observed by comparing *C. briggsae* to *C. elegans* might not reflect the overall trend in the Caenorhabditis genus. Alternatively, these trends could result from sampling bias in the set of homologous genes used to perform this analysis, both because the gene predictions are not necessarily accurate in the genomes of gonochoristic species and because not all genes whose expression were detected were also part of the modENCODE dataset used to define homology. Better genome assemblies and annotations would be required in order to establish absolute trends.

2.3. Effect of mating system

Transcriptional variance is under strong stabilizing selection in *C. elegans* (Denver, Morris et al. 2005). This trend should be conserved in other Caenorhabditis at least for genes conserved across the genus. Consistent with this and as observed previously, regulation of gene expression seemed to be mostly conserved between *C. elegans* and *C. remanei,* and overall expression patterns were mostly similar between *C. elegans* and gonochoristic species. Nevertheless there were notable differences between selfing and gonochoristic species. Compared with gonochoristic species, *C. elegans* genes with strongly male-biased had lower expression levels in males and higher expression levels in pseudo-females. This suggests that in *C. elegans*, "male gene" expression is de-repressed in *C. elegans* pseudo-females, and that *C. elegans* males are less able to initiate the expression levels of their ancestors. In addition, there were consistently fewer QTUs observed with high female-biased expression in either analytical pipeline in *C. elegans* compared to gonochoristic species. In the light of the distribution of sex-biased expression among lineage-specific genes (see above) this could be due to a higher rate of creation or divergence of genes with strong female-biased expression in gonochoristic species compared to *C. elegans.* Lastly, there were fewer QTUs displaying sex-biased expression overall in *C. elegan*s compared to gonochoristic species. Although this is in part a consequence of *C. elegans* smaller genome size, genes with strong sex-biased expression in *C. remanei* were more likely to be missing in one or both selfing species: *C. remanei* genes with a strong sex-biased expression and missing in *C. elegans* or in *C. briggsae* represent respectively about 3.4Mb and 3.9Mb of coding sequences and intergenic DNA. While this accounts for a small fraction of the difference in sizes between genomes of androdioecious and gonochoristic species, genes with strong sex-biased expression are lost disproportionately in selfing species compared to genes overall.

3. Function of genes with sex-biased expression missing in selfing species

As anticipated from the preliminary survey (Chapter 1), *C. remanei* genes with strong sex-biased expression were disproportionately absent from the genomes of both selfing species *C. elegans* and *C. briggsae*. Surprisingly these genes were more likely to be associated with GO terms (Ashburner, Ball et al. 2000), in large part due to their enrichment for genes predicted to encode retroelement-associated proteins. Because transposable element genes are systematically removed from genome annotations in *C. elegans* and to a lesser extent from that of *C. briggsae*, this overrepresentation might be an artefact due to faulty genome annotations. A greater expression of retroelements in gonochoristic species would be consistent however with higher numbers of TE in gonochoristic species, in line with the observed higher number of repeats in the genomes of gonochoristic species (see Introduction, Table 1, Caenorhabditis Repeat Libraries) and illustrating the importance of sex and outcrossing for maintenance of TE in populations as modeled by Wright and Schoen (1999). In addition, in the only other comparable study to date, more TEs were identified in obligate outcrossing *A. lyrata* compared to close congener selfing species *A. thaliana* (Hu, Pattyn et al. 2011). A previous genome analysis in *C. remanei* and

C. elegans identified less TEs in the *C. remanei* genome than in that of *C. elegans* (Dolgin, Charlesworth et al. 2008) indicating that the widespread belief that genome sizes should increase with smaller effective population size (Lynch and Conery 2003) may have somewhat limited investigations in the Caenorhabditis genus. In the light of our findings, a more thorough examination of TE expression and genome content should be contemplated.

Conclusions

Taken together, these results suggest that the decrease in QTUs with strongly sex-biased expression observed in *C. elegans* is mainly due to a reduction in the number of genes whose expression is sex-biased rather than to a reattribution of expression pattern from one sex to the other. Further, drastic genome size reduction in both *C. briggsae* and *C. elegans* is correlated with loss of genes whose homologues are expressed in a sex-biased fashion in other Caenorhabditis species. We propose that the emergence and ongoing evolution of selfing in Caenorhabditis species leads to desexualization of their transcriptomes in addition to loss of traits related to mating processes.

Conclusions : A model for the consequences of the emergence of selfing on gene expression

This study investigated the effect of the evolution of a new mating system on sex-biased expression in Caenorhabditis. Two mating systems exists in this genus, gonochorism with males and females, and androdioecy with rare males and self-fertile hermaphrodites. I have shown that *C. elegans* is drastically different from the gonochorists as a group in both the numbers and expression levels of genes or QTUs with highly sex-biased expression. More specifically, genes with highly female-biased expression are disproportionately rarer in *C. elegans* than expected from the gonochoristic species, and conserved genes with highly male-biased expression are not as canalized. Lastly, genes consistently male-biased in their expression and lost in selfing species displayed expression patterns consistent with roles in mating processes. Overall, our observation suggest strongly that genome size reduction in selfing species is mainly due to loss of genes (9.3 Mb of coding genes and 10.1 Mb of inergenic DNA between *C. remanei* and *C. elegans*) – as opposed to loss of TEs (1 Mb between *C. remanei* and *C. elegans*). Among genes missing in selfing species, loss of sex-biased QTUs is higher than expected by chance, and may therefore contribute to the observed decrease in mating ability.

The recently discovered process of indel segregation distortion (ISD) in Caenorhabditis promotes disproportionate transmission of shorter chromosomes through X-bearing gametes (Wang, Chen et al. 2010). Androdioecious Caenorhabditis preferentially reproduce by selfing (Dolgin, Charlesworth et al. 2007). I propose that as selfing arose in the Caenorhabditis genus, selective pressure on genes conferring better mating relaxed. As a result some of these genes may have gradually disappeared from the genomes of XX animals through the combined action of unequal chromosomal recombination and ISD. Over time, the loss of such mating-related genes (and most likely many others) would contribute to the decrease in genome size generation after generation. In addition to loss of genes, the regulation of conserved genes with male-biased expression was modified both in males and in hermaphrodites, such that they were expressed at a lower level in androdioecious males and partially de-repressed in hermaphrodites (formerly females). These conserved genes are likely to be involved in male-specific phenotypes absolutely required for mating, such as male anatomy, seminal fluid or sperm production. The modification of the regulation of their expression could be explained by either relaxed selection on male traits, or on a cryptic maleness required for the evolution of the ability of hermaphrodites to produce sperm. In either case, “male” genes are derepressed in hermaphrodites and their expression levels slightly reduced in males.

Future Directions

In order to study the evolution of sex-biased expression in the Caenorhaditis genus, large RNA-seq datasets were generated, representing the transcriptomes of *C. japonica, C. elegans, C. brenneri* and *C. remanei* males and females or pseudo-females at the L4 and adult stages. These datasets could be used to confirm current *in silico* gene predictions, improve gene annotations, discover new transcripts and explore alternative splicing in each of these species. Moreover, they could be used to characterize differential expression and transcriptional variance between species.

I determined sex-biased expression of existing gene predictions for each of four Caenorhabditis species, and assessed patterns of conservation of *C. remanei* genes with sex-biased expression. To complete the understanding of how sex-biased expression changes during evolution, codon usage as well as traces of selection in genes with sex-biased expression should be measured and compared both to that of genes unbiased in their expression, and species to species. This will allow us to establish the rate at which those genes are evolving compared to each other, and the strength of selection they are under. Genes with male-biased expression are expected to evolve faster in XX/XY systems, and little is known about genes with female-biased expression so far (reviewed in Ellegren and Parsch 2007).

The impact of inbreeding on genome structure is thought to be substantial (Lynch 2007). The only instance of direct comparison of genome structure in two closely related species with alternative mating systems was done in Arabidopsis and showed that both coding gene number and the number of transposable elements were major contributors to the genome size difference between *A. thaliana* and *A. lyrata* (Hu, Pattyn et al. 2011). While the same sort of assessement was beyond the scope of this study, we did identify an unexpectedly high proportion of sex-biased transcripts missing in both selfing species and related to retrotransposon in a gonochoristic species. Whether this overrepresentation was due to genome annotation discrepancies or real biological phenomena remains to be determined. However, our datasets provide a unique opportunity to examine retrotransposon expression in gonochorisic species compared to selfing species.

In addition to TE dynamics, indel segregation distortion (ISD) at meiosis (Wang, Chen et al. 2010) is very likely to play a major role in rapid changes in Caenorhabditis genome sizes reduction and in the evolution of genes with sex-biased expression. Since the X chromosomes are exceptionally well conserved between species in the Caenorhabditis species (Hillier, Miller et al. 2007; Ross, Koboldt et al. 2011, CGT pers. obs.) ISD is expected to primarily affect autosomes. Therefore, genes sex-biased in their expression and with no homologues in selfing species should be predominantly located on high recombination domains of autosomes (Barnes, Kohara et al. 1995; Hillier, Miller et al. 2007; Rockman and Kruglyak 2009; Ross, Koboldt et al. 2011). In addition, genes on the X chromosomes would be expected to be less likely to be male-biased in gonochoristic species, as is the case in *C. elegans* (Reinke et al 2000). Locating the position of genes with sex-biased expression in their respective genomes should allow to test these predictions.

A third androdioecious species, *C.* sp. 11, has recently been discovered in the Caenorhabditis genus, and its genome is even smaller than that of *C. elegans* or

C. briggsae (Thomas, Li et al. In prep.). It would be interesting to assay different parameters related to mating behavior in *C.* sp 11 males and hermaphrodites from this species, and in parallel compare the content of their genome to that of other Caenorhabditis species. Since the genome size of *C.* sp. 11 is much smaller, careful characterization of is content in relation with behavior assays would provide an additional point of comparison and allow us to test the prediction that genome shrinkage in Caenorhabditis is partly due to the loss of genes conferring mating efficiency-related traits, and partly due to loss of TE.

Efforts are underway to generate more complete genome sequences for *C. japonica, C. brenneri, C. remanei* (H. Smith, pers. comm.) and *C.* sp. 11, 9 and 7 (E. Schwarz, pers. comm.). Another avenue to consider pertains to genomic DNA-level comparisons between closely related Caenorhabditis species pairs, such as androdieocious *C. briggsae* and gonochorist *C.* sp. 9. The genome of the latter is twice the size of the former, yet these species can still produce fertile hybrids (Woodruff, Eke et al. 2010). Direct comparison of genomic sequences would allow a more precise identification of lost sequences.

The characterization of some *C. remanei* genes with high sex-biased expression led to the identification of candidate genes whose homologues were lost in one or both of the selfing species. Further functional studies of these genes, facilitated by the assessment of their expression pattern should shed some light on the molecular processes affected in *C. elegans* and *C. briggsae* as selfing arose in the Caenorhabditis genus. Similarly, an equally careful curation of genes with strong female-biased expression, less abundant in *C. elegans,* would allow a better understanding of how selfing arose and measure the consequences on *C. elegans* gene expression and genome structure.

Lastly, this study led to the identification of a novel candidate cis-regulatory element associated with male-biased expression. Further studies would be required in order to assess the exact role of this site through the identification of putative binding factors.

Appendices

Appendix 1.

List and description of scripts. Authors are indicated when not CGT

PERL scripts (Alphabetical order)

Chapter 1

10fold.pl
input: 2 files output from mapstatcounts.pl
output: list of genes over 10 fold in both input lists

adapterfreq.pl
input: RNA-seq FASTA file
output: list of reads starting with AAGCAGTGGT (adapter contamination)

all_fpkm_rem.pl
input: output from Cufflinks : FPKM values for each set for each gene prediction, for each of the 8 RNA-seq sets
output: one table, 8 FPKM values per gene prediction, for every gene prediction (0 if no FPKM value), and 8 files for input in R

avge_mal_fem.pl
input: output from R script: table with 8 FPKM values per gene prediction
output: 1 file per dataset (35bp or 50bp) with average FPKM in males and females and chi square values per gene prediction

bias.pl
input: two files with gene counts from each dataset (35bp and 50bp) and one gene ID
output: read counts in males and females for gene ID of interest

blast_filter2.pl
author: Ian Korf (Korf lab), modified by CGT
input: output from qstaq.pl
output: BLAST hits per reads, score above 16, percent identity over 95% and either 3 first or last bp matching to hit

chi.pl
input: output from stat_counts.pl
output: table with female and male counts and chi square value per gene prediction for those with significant chi square values (Bonferroni corrected)

chi_comp.pl
input: output from stat_counts.pl
output: table of male and female pseudo-counts, chi square values, and WS195 homologues per gene prediction

compare_list.pl
input: list of gene predictions with male and female counts, chi square values, expression differential value and bias for gene within Bonferroni correction threshold
output: three tables - genes over 10 fold differential in each dataset (35bp and 50bp), those with same bias in both datasets and those with opposite bias

compdataset_corr.pl
input: outputs from chi_comp.pl
output: same table but with pseudo-counts and chi squares from both datasets

compl_extractloc.pl
input: read counts in males and females for each dataset (35bp and 50bp) and gene name
output: table with counts per location (input for mapping profile)

coor_lookup.pl
author: Ian Korf (Korf lab)
input: output from blast_filter.pl
output: table of location in genome sequence reference (gene prediction, near gene or unknown) per read alignment

m.pl
input: output from mapstatcounts.pl
output: table with read counts in males and females, and chi square values per gene prediction

mapstats.pl
author: Ian Korf (Korf lab)
input: output from coor_lookup.pl
ouptut: statistics about alignments (nb of total reads, unique reads, number of matches, number of genes hit, uniqueness of matches and location)

mapstatcounts.pl
input: output from coor_lookup.pl
ouptut: read counts per gene prediction

overlap_comp.pl
input: list of genes and read counts in males and females from each dataset (35 and 50 bp)
output: expression differential in each dataset per gene prediction for all detected gene predictions

pipeline.pl
author: Ian Korf (Korf lab)
input: N/A
output: runs uniquer.pl, qstaq.pl and blast_filter.pl in a pipeline

polycheck.pl
author: Ian Korf (Korf lab)
input: RNA-seq FASTA file
output: table with homopolymer frequencies

prctbase_control.pl
input: RNA-seq FASTA file
output: table with frequencies of each nucleotide per position on the reads

qstaq.pl
input: output from uniquer.pl
output: BLASTn result of unique reads against *C. remanei* genome assembly

stat_counts.pl
input: output from tablecounts.pl
output: table of male and female counts, pseudo-counts and chi square values per gene prediction

readcount.pl
author: Ian Korf (Korf lab)
input: RNA-seq FASTA file
output: distribution of number of reads

tablecounts.pl
input: output from mapstatcounts.pl
output: tables with read counts in each dataset per gene prediction and per genomic location

uniquer.pl
author: Ian Korf (Korf lab)
input: RNA-seq FASTA file
output: unique reads from the input FASTA file

Chapter 2

nsite.pl
input: list of sequences, FASTA format (i.e. assembly)
output: locations of site defined by
C(A|T)(C|G)[ACGT](A|T)(A|C)AGTTCGAA[ACG](T|A) in sequences

Chapter 3
10_20_contigs.pl
input: list of contigs, FASTA format (i.e. assembly)
output: list of bottom 10 and 20% is size (2 files, contigs ID and sizes in bp)

2cvge_ctg.pl
input: bowtie .map output
output: table of contig IDs, length, total number of mapped reads, number of unique mapped reads, double, triple and quadruple mappers, number of mapped reads per bp, length of contig/genome length, GC content and expected value of mapped reads based on window size of 200bp.

3merge.pl
input: output from R script : RPKM values per contig for each RNA-seq set of same sex
output: table of RPKM values per contig, only for contigs with a value in 3 sets from same sex

3merge_fpkm.pl
input: output from Cufflinks : FPKM values per gene prediction for each RNA-seq set of same sex
output: table of FPKM values per gene prediction, only for those with a value in 3 sets from same sex

3merge_over1_fpkm.pl
input: output from R script : FPKM values per contig for each RNA-seq set of same sex
output: table of FPKM values per contig, only for contigs with a value in 3 sets from same sex and FPKM > 1

add_length_to_DE.pl
input: statistics of list of contigs (ID, length and GC content) and list of contigs with expression values in each of 6 datasets and p-value for expression differential
output: list of contigs with expression values in each of 6 sets, p-value and length of contig

add_length_to_DE1.pl
input: statistics of list of contigs (ID, length and GC content) and list of contigs with expression values in males and females, and p-value for expression differential
output: list of contigs with expression values in males and females, p-value and length of contig

all_ctgs_rpkm.pl
input: output from R : RPKM values for each set for each contig, for each of the 6 RNA-seq sets
output: one table, 6 RPKM values per contig, for every contig from soap output (0 if no RPKM value)

all_fpkm.pl
input: output from Cufflinks : FPKM values for each set for each gene prediction, for each of the 6 RNA-seq sets
output: one table, 6 FPKM values per gene prediction, for every gene prediction (0 if no FPKM value), and 6 files for input in R

bin_sorting.pl
input: list of contigs with expression values in each of 6 datasets, p-values and length
output: three files for short (up to 48bp), medium (49-199bp) and long contigs (longer than 200bp) with expression values, p-values and lengths

blast_stuff.pl
input: list of sequences, FASTA format (i.e. assembly)
output: BLASTn results of BLASTing list of sequences against itself

bre_only_ort.pl
input: output from worm_ort.pl, generated from modENCODE orthologue table
output: table of *C. brenneri* gene predictions modENCODE orthologues

comp_2_lists_sb.pl
input: list of remanei and elegans 1:1 orthologues
output: table of remanei genes and their bias, elegans orthologue and their bias

Compte_rem_with_homologue.pl
input: list of C. remanei gene predictions and attributed *C. elegans* homologues
output: number of C. remanei detected gene prediction with a homologue in *C. elegans*

count_go.pl
input: list of GO terms per gene predictions
output: number of genes per GO term of interest

count_groups.pl
input: assembly statistics (ID, length and GC content) and BLAST output from blast_stuff.pl
output: list and number of contigs corresponding to the same biological sequence

count_in_geneIDs.pl
input: geneID file from wormbase
output: number of gene predictions per species

count_loci_in_gtf.pl
input: .gtf file
output: number of protein coding gene predictions

count_locus_in_expr.pl
input: .expr file from Cufflinks
output: number of loci

count_ort_modencode.pl
input: list of genes prediction of interest
output: list of homologues from modENCODE per gene prediction of interest, and number of genes in each category

cts_dbl_in_blast.pl
input: output from blast_stuff.pl
output: number of different gene predictions hit

cvge_ctg.pl
input: bowtie .map output
output: table of contig IDs, length, total number of mapped reads, number of unique mapped reads, double, triple and quadruple mappers, number of mapped reads per bp, length of contig/genome length, GC content and expected value of mapped reads based on contig length.

cvge_male_female.pl
input: bowtie .map outputs
output: table of length, GC content, mapped reads per contigs

ele_bias.pl
input: *C. elegans* gene predictions hits from BLASTn SOAPdenovo contigs and expression values for contigs and gene predictions
output: table of expression value for each contig and corresponding *C. elegans* gene prediction

ele_only_ort.pl
input: output from worm_ort.pl, generated from modENCODE orthologue table
output: table of *C. elegans* gene predictions modENCODE orthologues

filter_1-1_ort.pl
input: table of *C. remanei* and *C. elegans* homologues
output: table of *C. remanei* and *C. elegans* 1:1 orthologues

final_merge.pl
input: two sets of expression values from 3merge.pl or 3merge_fpkm.pl
output: merged table of expression values for RPKM or FPKM per QTU ID

get_CREids.pl
input: c_remanei.WS225.orthologs.txt, list of homologues of all *C. remanei* gene predictions, wormbase
output: table of homologues per *C. remanei* gene prediction

get_bre_ort.pl
input: modENCODE homology table
output: homologues in 12 species of *C. brenneri* gene predictions

get_ele_ort.pl
input: modENCODE homology table
output: homologues in 12 species of *C. elegans* gene predictions

get_fc_elevsrem.pl
input: lists of homologous genes between *C. elegans* and *C. remanei* and their expression values
output: table with expression differential of each gene per homologous pair

get_jap_ort.pl
input: modENCODE homology table
output: homologues in 12 species of *C. japonica* gene predictions

get_rem_ort.pl
input: modENCODE homology table
output: homologues in 12 species of *C. remanei* gene predictions

get_seq_from_soap.pl
input: list of contig IDs
output: list of corresponding contig sequences, FASTA format

go_stuff.pl
input: gene_ontology.WS226.obo
output: GO term for each *C. remanei* gene prediction

go_stuff_ele.pl
input: gene_ontology.WS226.obo
output: GO term for each *C. elegans* gene prediction

jap_only_ort
input: output from worm_ort.pl, generated from modENCODE orthologue table
output: table of *C. japonica* gene predictions modENCODE orthologues

log2.pl
input: output from R script, table of RPKM values
output: log2 value of input, same format

log2_2sets.pl
input: output from R script, table of FPKM values
output: log2 value of input, same format

look_for_go.pl
input: list of *C. remanei* gene
output: list of GO term per gene prediction

look_for_go_ele.pl
input: list of *C. elegans* gene
output: list of GO term per gene prediction

mean_and_median.pl
input: table of expression values in males and females per QTU
output: mean and median RPKM or FPKM value per sex

merge_list.pl
input: two sets of expression values in males, females and chi square per gene prediction
output: merged table of values in each set for gene predictions in both sets

merge_rpkm.pl
input: tables from each of 6 datasets with RPKM values
output: table with RPKM values per contig for those present in all 6 RNA-seq datasets

n50.pl
input: assembly statistics table (ID, length and GC content)
output: assembly size (bp and number of contigs), N50 (bp), N50 contig

new_assembly.pl
input: SOAPdenovo output, FASTA format
output: filtered assembly (tips out, coverage > 5 and contigs > 49bp)

ort_count.pl
input: list of C. remanei gene predictions
output: corresponding homologues from Wormbase file

over1_all_fpkm.pl
input: Cufflinks output, 6 datasets
output: table of FPKM values in each dataset per gene prediction, values < 1 changed to 0

parse_blast.pl
input: output from blast_stuff.pl
output: files with self-hits and significant hits

soap_stat_contigs.pl
input: SOAPdenovo output (list of sequences, FASTA format)
output: table of contig ID and length

stat_counts.pl
input: table with read counts per gene prediction
output: table with read counts, corrected read counts and chi square values per gene predictions

stats_contigs.pl
input: list of sequences, FASTA format (i.e. assembly)
output: table with all contig IDs, length and GC content

volc_prep.pl
input: table with RPKM or FPKM values in each of 6 datasets and p-values per QTU
output: table with average value in males and females, p-values per QTUs

worm_ort.pl
input: modENCODE homology table
output: modENCODE homology table only with Caenorhabditis gene predictions

WB_to_protID_modencode.pl
input: output from worm_ort.pl
output: same table with geneID names instead of WB numbers

Misc.
comp_2_lists.pl
input: two list of genes
output: list of genes in common to two lists

count_gns_in_fa.pl
input: list of sequences, FASTA format (i.e. assembly)
output: number of sequences

extract_expression_values.pl
input: list of genes and list of all genes and expression values
output: list of expression values for genes of interest

nt_freq.pl
input: list of sequences, FASTA format (i.e. assembly)
output: mono- and di-nucleotide frequencies

revcomp_fasta.pl
input: list of sequences, FASTA format
output: reverse complement of sequences

PERL libraries

FAlite.pm
library, Ian Korf
FASTA files manipulation

libcris.pm
library, CGT
various subs (GC content, nucleotide and aa compositions)

DataBrowser.pm
library, Ian Korf
Browsing tables

R scripts

* developed with Renhua Li (then in the Oliver lab)

Chapter 1
Rlogs_rem_deepseq1.r
Analysis of TopHat / Cufflinks outputs and figures

Chapter 3
R_logs_brenn_no.r
Analysis of SOAPdenovo / Bowtie outputs

*R_logs_bre_new.r**
Analysis of SOAPdenovo / Bowtie outputs

Rlogs_bre_gtf.r
Analysis of TopHat / Cufflinks with GTF file outputs

Rlogs_bre_nogtf.r
Analysis of TopHat / Cufflinks without GTF file outputs

*R_logs_jap_new.r**
Analysis of SOAPdenovo / Bowtie outputs

Rlogs_jap_gtf.r
Analysis of TopHat / Cufflinks with GTF file outputs

Rlogs_jap_nogtf.r
Analysis of TopHat / Cufflinks without GTF file outputs

Rlogs_remgtf.r
Analysis of TopHat / Cufflinks with GTF file outputs

Rlogs_remnogtf.r
Analysis of TopHat / Cufflinks without GTF file outputs

*R_logs_rem_new.r**
Analysis of SOAPdenovo / Bowtie outputs outputs

*R_logs_ele_new_soap.r**
Analysis of SOAPdenovo / Bowtie outputs

Rlogs_ele_gtf.r
Analysis of TopHat / Cufflinks with GTF file outputs

Rlogs_ele_nogtf.r
Analysis of TopHat / Cufflinks without GTF file outputs

Figures :
soap_contig_stats.r
More_figures.r
R_script_figures.r
Addon_script_figures.

Appendix 2. Primers used in this study. T7 promoter sequences as well as attB sequences are indicated in bold. Primers used exclusively for PCR and sequencing are in black, primers used in addition for qRT-PCR or dsRNA synthesis or both are respectively indicated in blue, brown and grey. Primers used only for qRT-PCR are in green, those used only for dsRNA synthesis in purple. Pirmers used for Gateway cloning are in pink.

Gene ID	Primer ID		Primer sequence (3'-5')
CRE01361	CGT243	F	**taatacgactcactatagggaga**ATGAgtagagagcaagtttctgtggc
CRE01361	CGT244	R	**taatacgactcactatagggaga**TGGTGGATCATGCAACGGAC
CRE01361	CGT189	F	ATGAgtagagagcaagtttctgtggc
CRE01361	CGT190	R	TGGTGGATCATGCAACGGAC
CRE01361	CGT192	R	CTAGGTGATTGTTGGTGGGAGG
CRE01361	CGT191	F	CGTTGCATgatccaccaaac
CRE01361	CGT251	R	GAGAGTGACAAGTCTTCTGTTGGc
CRE01558	CGT096	F	cgtggcacgacaagatactgc
CRE01558	CGT097	R	GGGCATCGATGGATAAATCAC
CRE04599	CGT261	F	**taatacgactcactatagggaga**atggaagaaattcaagcgatcac
CRE04599	CGT262	R	**taatacgactcactatagggaga**CGAACCATTCCGTAGTCAATG
CRE04599	CGT177	F	atggaagaaattcaagcgatcac
CRE04599	CGT179	F	**taatacgactcactatagggaga**CAATCAAACAGtcacacttaaggtttc
CRE04599	CGT235	F	gaagttggatacacacatcggg
CRE04599	CGT250	F	cattgactacgGAATGGTTCg
CRE04599	CGT180	R	**taatacgactcactatagggaga**GTTCAAAATTTTCTTCTTGAAGTGAGC
CRE04599	CGT248	R	AAGCTTTGAGAAGATAAGCTTCGAG
CRE04599	CGT249	R	cGATTAACATGTAAAGCCACGACC
CRE04599	CGT256	F	GGCTCATAAAAGAGAAGCAAATATCC
CRE04599	CGT178	R	CGAACCATTCCGTAGTCAATG
CRE05483	CGT171	F	GGGGCTctctgtttttaataaaaaacc
CRE05483	CGT173	F	tggcggagAACCACGGAAC
CRE05483	CGT172	R	GGAGTTACAGCTAACATCTTGGACC
CRE05483	CGT236	R	GTCAAAACCGTCTCGAATGC
CRE05483	CGT175	F	**taatacgactcactatagggaga**cGGCCAGATTCCAGTCAAGG
CRE05483	CGT176	R	**taatacgactcactatagggaga**TTATTTCAGAATGGTCATCGTCGA
CRE05483	CGT260	F	CAATAATCGACGTGTGGAGACAG
CRE05483	CGT174	R	CCTTTGCAATTTCCAAAAAGTTGAG
CRE05585	CGT095	F	GCTCGGAAACTCAATTacataaagc
CRE05585	CGT094	R	CGGAATAGATTAACATTCCAATCG
CRE08297	CGT027	F	**taatacgactcactatagggaga**AGCTGCAAAATCGAAGAACG

CRE08297	CGT088	R	GATGAGCACACAGAATCCACTCac
CRE11138	CGT155	F	ATGCGCTTCCGAAGGTTGC
CRE11138	CGT157	F	aagttctgGTTACCAAATCGTCG
CRE11138	CGT159	F	GAGTCTTTCCATCTGGGGTCAGG
CRE11138	CGT161	F	GACGATAGCAgctgggttcC
CRE11138	CGT163	F	gagttatcgaCCGATCAAATAGAATTC
CRE11138	CGT165	F	tcaaggagatcTCAGTGTGGC
CRE11138	CGT156	R	GTTCTTCAACCTGGGATTTGAGG
CRE11138	CGT158	R	ACATTACTTTGTTCCTGACCCCA
CRE11138	CGT160	R	CCGCATCCGATTAAAGCATACAC
CRE11138	CGT162	R	GGAGTCATGTCGTCTTCTAGAGATGG
CRE11138	CGT164	R	GTAAGTGGAAACTCCCCTTCTTTC
CRE11138	CGT166	R	CTAATTACCTCGAACAGTCATTTCG
CRE12267	CGT187	F	**taatacgactcactatagggaga**ATGAGAATCGTCGCCTTCTCC
CRE12267	CGT188	R	**taatacgactcactatagggaga**CTAGAACACCTTCACAGGGCACttc
CRE12267	CGT258	F	TTCCTGGCCCTCATCTCGTC
CRE12267	CGT259	R	TTGGAAGTGCTTACTCTTCACCG
CRE12278	CGT185	F	**taatacgactcactatagggaga**ATGAGAATTCCGATTCTTTTCTTCG
CRE12278	CGT186	R	**taatacgactcactatagggaga**TCAGTCGAACACCTTTTCAGGAC
CRE12278	CGT263	F	TTCAAAGAGTTCCTTGCGATGC
CRE12278	CGT264	R	CCACTTGTTATCGGCCACGTT
CRE14475	CGT238	R	CGAACATGTGTATCAGGATGGaaag
CRE23534	CGT181	F	AACGTATTTGTAATAATCCTTGCTGC
CRE23534	CGT247	F	ATGTGGTGACTCCAATCTTCACg
CRE23534	CGT182	R	GATAATTCGGAAACTGTATTTCACG
CRE23534	CGT184	R	**taatacgactcactatagggaga**TCAAACAATTTTCTTAATCAACTCCG
CRE23534	CGT183	F	**taatacgactcactatagggaga**GGTCTTGGTGAGatcaactcgtac
CRE24268	CGT092	F	caatcaatcaattgttgaaatccg
CRE24268	CGT093	R	GGCTTTCAGATTTTGCTCATCC
CRE28795	CGT167	F	atgtttccagttgtcgatggtg
CRE28795	CGT169	F	atggagaaCAGTACAAATGGAGGC
CRE28795	CGT168	R	CCATTTGTACTGTTCTCCATCATGc
CRE28795	CGT170	R	TCAACGACGGAAATTGAAAAAGAGTAc
CRE28796	CGT206-B4	F	**GGGGACAACTTTGTATAGAAAAGTTG**ttGACCCAGGAGTTTGTACTTTG
CRE28796	CGT-208-B2R	F	**GGGGACAGCTTTCTTGTACAAAGTGGGA**TAAACAAATTAGGAGGCGCTATATCGa
CRE28796	CGT205-B1R	R	**GGGGACTGCTTTTTTGTACAAACTTG**tgttgtttcggcagatgagg
CRE28796	CGT207-B3	R	**GGGGACAACTTTGTATAATAAAGTTG**ttatgtgaagggatcggagtaaac

CRE28796	CGT241	F	ctGAAAATGGGTATGAGTCAGTGG
CRE28796	CGT252	F	aaGGAAATTACCAATCTCAACAAGC
CRE28796	CGT242	R	gGTTTGGATTTTGGATTCTCTCG
CRE28796	CGT253	R	GGAGCATCTTCATATTCAAACTGG
CRE28796	CGT098	F	**taatacgactcactatagggaga**GTTCATGCTAGCAgggcattgc
CRE28796	CGT099	R	**taatacgactcactatagggaga**GAATAATATTAAAATGGCACACAATGAC
CRE28796	CGT100	R	**taatacgactcactatagggaga**GCATTATACGAGCCATTTCATCAAC
CRE28796	CGT109	F	GTTCATGCTAGCAgggcattgc
CRE28796	CGT149	R	ggGGACGAACAGAGACTCATCAAG
Cre-mss-1	CGT193-B4	F	**GGGGACAACTTTGTATAGAAAAGTTG**GTGTCGAGGATTGAACCTTGATC
Cre-mss-1	CGT-195-B2R	F	**GGGGACAGCTTTCTTGTACAAAGTGGGA**TAAACTACAGTATCCATCAACAACGgc
Cre-mss-1	CGT194-B1R	R	**GGGGACTGCTTTTTTGTACAAACTTG**GGTTTATCCCAACTCGGTTGAA
Cre-mss-1	CGT196-B3	R	**GGGGACAACTTTGTATAATAAAGTTG**GGTTTATCCCAAGTCGGTTGAAG
Cre-mss-1	CGT056	F	CCCAAAGACACTCAACCTGATGC
Cre-mss-1	CGT035	R	**taatacgactcactatagggaga**TTGGAGGTGCAATGCCGTTG
Cre-mss-1,2,4	CGT033	F	**taatacgactcactatagggaga**ATGATCCAGAAGACGACCATTCT
Cre-mss-2	CGT195-B4	F	**GGGGACAACTTTGTATAGAAAAGTTG**ACTACAGTATCCATCAACAACGgc
Cre-mss-2	CGT-197-B2R	F	**GGGGACAGCTTTCTTGTACAAAGTGGGA**TAAACTACAGTATCCATCAACAACGGC
Cre-mss-2	CGT196-B1R	R	**GGGGACTGCTTTTTTGTACAAACTTG**GGTTTATCCCAAGTCGGTTGAAG
Cre-mss-2	CGT198-B3	R	**GGGGACAACTTTGTATAATAAAGTTG**GGTTGATGTCAGCGAATACTGttc
Cre-mss-2	CGT058	F	GGATTTCCGACTCCACCATCTG
Cre-mss-2	CGT057	R	GGATCTTCTGGGGCTTTCGG
Cre-mss-3	CGT037	R	**taatacgactcactatagggaga**CAAATGATGCTTTGAGTGTGATG
Cre-mss-3	CGT197-B4	F	**GGGGACAACTTTGTATAGAAAAGTTG**ACTACAGTATCCATCAACAACGGC
Cre-mss-3	CGT-199-B2R	F	**GGGGACAGCTTTCTTGTACAAAGTGGGA**TAAACTACAGAATCCATCAAAAAATGCatc
Cre-mss-3	CGT198-B1R	R	**GGGGACTGCTTTTTTGTACAAACTTG**GGTTGATGTCAGCGAATACTGttc
Cre-mss-3	CGT200-B3	R	**GGGGACAACTTTGTATAATAAAGTTG**CGGTCTCTATAATTACCAATCGCc
Cre-mss-3	CGT034	F	**taatacgactcactatagggaga**TTTTCTCTTCATGACCTTCGcac
Cre-mss-3	CGT059	R	CGTTTCTTCTATTGGTACTTCCGC
Cre-mss-4	CGT201-B4	F	**GGGGACAACTTTGTATAGAAAAGTTG**TTGTGCACATGCTTTGAATCc
Cre-mss-4	CGT-203-B2R	F	**GGGGACAGCTTTCTTGTACAAAGTGGGA**TAATCTCCAGTCTCCAAAAATTGACttg
Cre-mss-4	CGT202-B1R	R	**GGGGACTGCTTTTTTGTACAAACTTG**GGTTTATCCCAAGTCGGTGg
Cre-mss-4	CGT204-B3	R	**GGGGACAACTTTGTATAATAAAGTTG**GAAATTCGAAAAATTGTGGGGag
Cre-mss-4	CGT060	F	CCAAAACAATACCGCACAGGG
Cre-mss-4	CGT036	R	**taatacgactcactatagggaga**GTTGAAATTCTCAGCAAGTCAATTTT
Cre-nol-1	CGT153	F	gaagaaggagaaggctgctgagg
Cre-nol-1	CGT154	R	CCAATGTTTTTCATTGGTGCAGTC

Cre-vrm-1	CGT218-B4	F	**GGGGACAACTTTGTATAGAAAAGTTG**CGCAACTCGATCGAATGACT
Cre-vrm-1	CGT-220-B2R	F	**GGGGACAGCTTTCTTGTACAAAGTGGGA**TAAGATGTTCTGTTAAGGTAGGCAGCTG
Cre-vrm-1	CGT217-B1R	R	**GGGGACTGCTTTTTTGTACAAACTTG**AATTACTCGATCGAATAATTTCGAATT
Cre-vrm-1	CGT219-B3	R	**GGGGACAACTTTGTATAATAAAGTTG**CAAATCGGACTTATTTGAATTACGG
Cre-vrm-1	CGT135	F	atgaaattgctctcattgctcacc
Cre-vrm-1	CGT137	F	cacaagctgACTTGGAACACGg
Cre-vrm-1	CGT139	F	CCAAGAAGTTCAACGTGCCG
Cre-vrm-1	CGT141	F	**taatacgactcactatagggaga**atgaaattgctctcattgctcacc
Cre-vrm-1	CGT239	F	caaCGATATATCTCCATGTCTGGC
Cre-vrm-1	CGT136	R	cGTGTTCCAAGTCAGCTTGTGC
Cre-vrm-1	CGT140	R	TTACTCAGAAAGCGAACAAAGCc
Cre-vrm-1	CGT142	R	**taatacgactcactatagggaga**cGTGTTCCAAGTCAGCTTGTGC
Cre-vrm-1	CGT240	R	TTACCAGCTCATGCATGTCgg
Cre-vrm-1	CGT138	R	CCGAAGTGAAGTCCAGTTCATGC
Cre-vrm-1	CGT152	F	agaccaatccggaaattgtgacc
Cre-vrm-2	CGT-226-B2R	F	**GGGGACAGCTTTCTTGTACAAAGTGGGA**TAAAATCACACTCCCGAATCAATGGA
Cre-vrm-2	CGT228-B4	F	**GGGGACAACTTTGTATAGAAAAGTTG**CGCTACTCGATCGAATGACTTCG
Cre-vrm-2	CGT225-B3	R	**GGGGACAACTTTGTATAATAAAGTTG**CACGTTACCTAGCCTATGAGTGCTG
Cre-vrm-2	CGT227-B1R	R	**GGGGACTGCTTTTTTGTACAAACTTG**CGCTTCTTGATCGAATAATTTCGAA
Cre-vrm-2	CGT115	F	GCATTCGCCGCTCTTCACC
Cre-vrm-2	CGT119	F	atGCATAAATTGGCATGGAAGG
Cre-vrm-2	CGT121	F	**taatacgactcactatagggaga**GCATTCGCCGCTCTTCACC
Cre-vrm-2	CGT148	F	caacattgtcgtttgtttgattgg
Cre-vrm-2	CGT237	F	CGTAACTGATTCACATCGGAGAG
Cre-vrm-2	CGT254	F	GTATTCCAATGCATATTGACGTCG
Cre-vrm-2	CGT116	R	cTTGAAATCTCCACTTGTTCCAATC
Cre-vrm-2	CGT118	R	attCCTTCCATGCCAATTTATGC
Cre-vrm-2	CGT120	R	ggGAGTGTGATTCTATTTTCCATCAG
Cre-vrm-2	CGT122	R	**taatacgactcactatagggaga**cTTGAAATCTCCACTTGTTCCAATC
Cre-vrm-2	CGT145	R	AAGTCAATTTGTGCATGTTTGGc
Cre-vrm-2	CGT255	R	GCATGTTTGAAATTTGATATACCCTG
Cre-vrm-2	CGT117	F	CTTTTGGGAACAGAgggtgg
Cre-vrm-2	CGT147	R	CCAGGTCTTCGCTCCAAATCAG
Cre-vrm-3	CGT230-B4	F	**GGGGACAACTTTGTATAGAAAAGTTG**TTGTGGCTTTGGTAGATCGGCTC
Cre-vrm-3	CGT-232-B2R	F	**GGGGACAGCTTTCTTGTACAAAGTGGGA**TAAAGTCATCAGCTTTCACATGTCTTGTG
Cre-vrm-3	CGT229-B1R	R	**GGGGACTGCTTTTTTGTACAAACTTG**CGCAACTTGATCAAATTTCTTCGA
Cre-vrm-3	CGT231-B3	R	**GGGGACAACTTTGTATAATAAAGTTG**AACGTCCCCAGAACAGTGTCC

Cre-vrm-3	CGT129	F	cctccatttgtcgcttctactcc
Cre-vrm-3	CGT131	F	caaattgACTTGGAGCAACGAGc
Cre-vrm-3	CGT133	F	**taatacgactcactatagggaga**cctccatttgtcgcttctactcc
Cre-vrm-3	CGT150	F	gacatggaagcagtcccttgc
Cre-vrm-3	CGT132	R	ATTCGCTTTGGTTGGAAGTGAGC
Cre-vrm-3	CGT134	R	**taatacgactcactatagggaga**ttCCATATGAAGAGTGTTTCTGGtc
Cre-vrm-3	CGT151	R	GCAAGGGACTGCTTCCATGTC
Cre-vrm-3	CGT130	R	ttCCATATGAAGAGTGTTTCTGGtc
Cre-vrm-3	CGT144	F	gatGAGTTTATGAAGAAAAGTGACCG
Cre-vrm-4	CGT221-B4	F	**GGGGACAACTTTGTATAGAAAAGTTG**TGGTATCTCCAGAAAAGAAACGAGC
Cre-vrm-4	CGT-223-B2R	F	**GGGGACAGCTTTCTTGTACAAAGTGGGA**TAAAAAAGAAAATTCGCTTCTCGCG
Cre-vrm-4	CGT222-B1R	R	**GGGGACTGCTTTTTTGTACAAACTTG**CGCAACTTGATCAAATTTCTTCG
Cre-vrm-4	CGT224-B3	R	**GGGGACAACTTTGTATAATAAAGTTG**CAGGGTTCTGAGTGTCGTCGTC
Cre-vrm-4	CGT125	F	CGAGTTGacatggaagcagtcg
Cre-vrm-4	CGT127	F	**taatacgactcactatagggaga**aacgagttgATCTGGTCGGaag
Cre-vrm-4	CGT124	R	CATGTCAACTCGTGCATATTGG
Cre-vrm-4	CGT126	R	CATTCGATATGACTGGAAGTGAGC
Cre-vrm-4	CGT128	R	**taatacgactcactatagggaga**CATGTCAACTCGTGCATATTGG
Cre-vrm-4	CGT123	F	aacgagttgATCTGGTCGGaag
Cre-vrm-4	CGT143	R	GCTTTCATTTCCCATCCTTTTCG
Cre-pie-1	CGT-213-B2R	F	**GGGGACAGCTTTCTTGTACAAAGTGGGA**TAATTGTATTGTCTATATTGTATTATTCTCAGATACT
Cre-pie-1	CGT215-B4	F	**GGGGACAACTTTGTATAGAAAAGTTG**gtctgtgatcgggtgactgtc
Cre-pie-1	CGT214-B3	R	**GGGGACAACTTTGTATAATAAAGTTG**GCATAAATTCGTGTATACATATCAGtg
Cre-pie-1	CGT216-B1R	R	**GGGGACTGCTTTTTTGTACAAACTTG**TTTTGCTGGAGAATTGAGAATTATTC
Cre-spe-11	CGT209-B4	F	**GGGGACAACTTTGTATAGAAAAGTTG**CTGAAATAAATTATTCGTATTGGGGaa
Cre-spe-11	CGT-211-B2R	F	**GGGGACAGCTTTCTTGTACAAAGTGGGA**TAAAGAAAAAGCAGAAAATGACTACTAGAA
Cre-spe-11	CGT210-B1R	R	**GGGGACTGCTTTTTTGTACAAACTTG**GTTCTGACGGTTCTGGCGG
Cre-spe-11	CGT212-B3	R	**GGGGACAACTTTGTATAATAAAGTTG**ATTGTTTGATCCTAGTCGAAAATTC
Cre-unc-94	CGT091	F	cgacactcgtcacacagtcg
Cre-unc-94	CGT022	R	**taatacgactcactatagggaga**GGAGACGCACTCGTTCATAATTC
Cre-vha-8	CGT089	F	CGAGCAAGAAGCCAACGAGAAG
Cre-vha-8	CGT090	R	TTCACGAGCCTTCAAACAACGG
oac-9	CGT085	R	AAAATGACCCAGAAAGAGCTGC
twk-37	CGT084	F	TCATTCAGCACGTTTTCAACAG

Bibliography

Adamidi, C., Y. Wang, et al. (2011). "*De novo* assembly and validation of planaria transcriptome by massive parallel sequencing and shotgun proteomics." Genome Res **21**(7): 1193-200.

Anderson, J. L., L. T. Morran, et al. (2010). "Outcrossing and the maintenance of males within *C. elegans* populations." J Hered **101 Suppl 1**: S62-74.

Artieri, C. G., W. Haerty, et al. (2008). "Sexual selection and maintenance of sex: evidence from comparisons of rates of genomic accumulation of mutations and divergence of sex-related genes in sexual and hermaphroditic species of Caenorhabditis." Mol Biol Evol **25**(5): 972-9.

Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-9.

Ayroles, J. F., M. A. Carbone, et al. (2009). "Systems genetics of complex traits in *Drosophila melanogaster*." Nat Genet **41**(3): 299-307.

Baer, C. F., J. Joyner-Matos, et al. (2010). "Rapid decline in fitness of mutation accumulation lines of gonochoristic (outcrossing) Caenorhabditis nematodes." Evolution **64**(11): 3242-53.

Bailey, T. L. and C. Elkan (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, Menlo Park, California, AAAI Press.

Baldi, C., S. Cho, et al. (2009). "Mutations in two independent pathways are sufficient to create hermaphroditic nematodes." Science **326**(5955): 1002-5.

Barker, D. M. (1994). "Copulatory plugs and paternity assurance in the nematode *Caenorhabditis elegans*." Animal Behaviour **48**(1): 147-156.

Barr, M. M. and L. R. Garcia (2006). "Male mating behavior." WormBook: 1-11.

Barriere, A. and M. A. Felix (2005). "High local genetic diversity and low outcrossing rate in *Caenorhabditis elegans* natural populations." Curr Biol **15**(13): 1176-84.

Barriere, A. and M. A. Felix (2007). "Temporal dynamics and linkage disequilibrium in natural *Caenorhabditis elegans* populations." Genetics **176**(2): 999-1011.

Barriere, A., S. P. Yang, et al. (2009). "Detecting heterozygosity in shotgun genome assemblies: Lessons from obligately outcrossing nematodes." Genome Res **19**(3): 470-80.

Baudry, E., C. Kerdelhue, et al. (2001). "Species and recombination effects on DNA variability in the tomato genus." Genetics **158**(4): 1725-35.

Beadell, A. V., Q. Liu, et al. (In revision). "Independent recruitments of a translational regulator in the evolution of self-fertile nematodes." PNAS.

Beilstein, M. A., N. S. Nagalingum, et al. (2010). "Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*." Proc Natl Acad Sci U S A **107**(43): 18724-8.

Bendtsen, J. D., H. Nielsen, et al. (2004). "Improved prediction of signal peptides: SignalP 3.0." J Mol Biol **340**(4): 783-95.

Brodie, R., R. L. Roper, et al. (2003). "JDotter: A Java Interface to Multiple DotPlots Generated by Dotter." Bioinformatics 20(2): 279-281.

Bryne, J. C., E. Valen, et al. (2008). "JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update." Nucleic Acids Res **36**(Database issue): D102-6.

Caenorhabditis Repeat Libraries Release R1.0 (2008). http://genome.sfu.ca/projects/worm_genomes/REPEATS/RELEASE1.0/readme

Carr, D. E. and M. R. Dudash (2003). "Recent approaches into the genetic basis of inbreeding depression in plants." Philos Trans R Soc Lond B Biol Sci **358**(1434): 1071-1084.

Cassada, R. C. and R. L. Russell (1975). "The dauerlarva, a post-embryonic developmental variant of the nematode *Caenorhabditis elegans*." Dev Biol **46**(2): 326-42.

Charlesworth, B. and D. Charlesworth (1978). "A Model for the Evolution of Dioecy and Gynodioecy." The American Naturalist **112**(988): 975-997.

Charlesworth, B., A. Lapid, et al. (1992). "The distribution of transposable elements within and between chromosomes in a population of *Drosophila melanogaster.* II. Inferences on the nature of selection against elements." Genet Res **60**(2): 115-30.

Charlesworth, D. (1984). "Androdioecy and the evolution of dioecy." Biological Journal of the Linnean Society **22**(4): 333-348.

Charlesworth, D. (2006). "Evolution of plant breeding systems." Curr Biol **16**(17): R726-35.

Charlesworth, D. and S. I. Wright (2001). "Breeding systems and genome evolution." Curr Opin Genet Dev **11**(6): 685-90.

Chasnov, J. R. and K. L. Chow (2002). "Why are there males in the hermaphroditic species *Caenorhabditis elegans*?" Genetics **160**(3): 983-94.

Chasnov, J. R., W. K. So, et al. (2007). "The species, sex, and stage specificity of a Caenorhabditis sex pheromone." Proc Natl Acad Sci U S A **104**(16): 6730-5.

Chen, S., P. Yang, et al. (2010). "*De novo* analysis of transcriptome dynamics in the migratory locust during the development of phase traits." PLoS One **5**(12): e15633.

Cho, S., S. W. Jin, et al. (2004). "A phylogeny of caenorhabditis reveals frequent loss of introns during nematode evolution." Genome Res **14**(7): 1207-20.

Choi, K. Y., Y. J. Ji, et al. (2003). "Vacuolar-type H+-ATPase E subunit is required for embryogenesis and yolk transfer in *Caenorhabditis elegans*." Gene **311**: 13-23.

Clark, A. G., M. B. Eisen, et al. (2007). "Evolution of genes and genomes on the Drosophila phylogeny." Nature **450**(7167): 203-18.

Cloonan, N., A. R. Forrest, et al. (2008). "Stem cell transcriptome profiling via massive-scale mRNA sequencing." Nat Methods **5**(7): 613-9.

Coghlan, A. (2005). "Nematode genome evolution." WormBook: 1-15.

Coghlan, A. and K. H. Wolfe (2002). "Fourfold faster rate of genome rearrangement in nematodes than in Drosophila." Genome Res **12**(6): 857-67.

The *C. elegans* consortium. (1998). "Genome sequence of the nematode *C. elegans*: a platform for investigating biology." Science **282**(5396): 2012-8.

Cutter, A. D. (2008). "Reproductive evolution: symptom of a selfing syndrome." Curr Biol **18**(22): R1056-8.

Cutter, A. D., S. E. Baird, et al. (2006). "High nucleotide polymorphism and rapid decay of linkage disequilibrium in wild populations of *Caenorhabditis remanei*." Genetics **174**(2): 901-13.

Cutter, A. D., A. Dey, et al. (2009). "Evolution of the *Caenorhabditis elegans* genome." Mol Biol Evol **26**(6): 1199-234.

Cutter, A. D. and B. A. Payseur (2003). "Rates of deleterious mutation and the evolution of sex in Caenorhabditis." J Evol Biol. **16**:812-822

Cutter, A. D. and S. Ward (2005). "Sexual and temporal dynamics of molecular evolution in *C. elegans* development." Mol Biol Evol **22**(1): 178-88.

Cutter, A. D., J. D. Wasmuth, et al. (2006). "The evolution of biased codon and amino acid usage in nematode genomes." Mol Biol Evol **23**(12): 2303-15.

Cutter, A. D., J. D. Wasmuth, et al. (2008). "Patterns of molecular evolution in Caenorhabditis preclude ancient origins of selfing." Genetics **178**(4): 2093-104.

Cutter, A. D., W. Yan, et al. (2010). "Molecular population genetics and phenotypic sensitivity to ethanol for a globally diverse sample of the nematode *Caenorhabditis briggsae*." Mol Ecol **19**(4): 798-809.

Dabney, A., J. D. Storey, et al. (2011). *qvalue*: Q-value estimation for false discovery rate control.

de Castro, E., C. J. Sigrist, et al. (2006). "ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins." Nucleic Acids Res **34**(Web Server issue): W362-5.

de Hoon, M. and Y. Hayashizaki (2008). "Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference." Biotechniques **44**(5): 627-8, 630, 632.

del Castillo-Olivares, A., M. Kulkarni, et al. (2009). "Regulation of sperm gene expression by the GATA factor ELT-1." Dev Biol **333**(2): 397-408.

Denver, D. R., K. Morris, et al. (2005). "The transcriptional consequences of mutation and natural selection in *Caenorhabditis elegans*." Nat Genet **37**(5): 544-8.

Denver, D. R., K. Morris, et al. (2003). "Phylogenetics in *Caenorhabditis elegans*: an analysis of divergence and outcrossing." Mol Biol Evol **20**(3): 393-400.

Dolgin, E. S., B. Charlesworth, et al. (2007). "Inbreeding and outbreeding depression in Caenorhabditis nematodes." Evolution **61**(6): 1339-52.

Dolgin, E. S., B. Charlesworth, et al. (2008). "Population frequencies of transposable elements in selfing and outcrossing Caenorhabditis nematodes." Genet Res (Camb) **90**(4): 317-29.

Drummond AJ, A. B., Buxton S, Cheung M, Cooper A, Duran C, Field M, Heled J, Kearse M, Markowitz S, Moir R, Stones-Havas S, Sturrock S, Thierer T, Wilson A. (2011). "Geneious v5.4." from http://www.geneious.com/.

Efron, B. and R. Tibshirani (1993). An introduction to the Bootstrap, Chapman and Hall/CRC; 1 edition (May 15, 1994).

Ellegren, H., L. Hultin-Rosenberg, et al. (2007). "Faced with inequality: chicken do not have a general dosage compensation of sex-linked genes." BMC Biol **5**: 40.

Ellegren, H. and J. Parsch (2007). "The evolution of sex-biased genes and sex-biased gene expression." Nat Rev Genet **8**(9): 689-98.

Emmons, S. W. (2005). "Male development." WormBook: 1-22.

Felix, M. A. (2008). "RNA interference in nematodes and the chance that favored Sydney Brenner." J Biol **7**(9): 34.

Finn, R. D., J. Mistry, et al. (2010). "The Pfam protein families database." Nucleic Acids Res **38**(Database issue): D211-22.

Fraser, A. G., R. S. Kamath, et al. (2000). "Functional genomic analysis of *C. elegans* chromosome I by systematic RNA interference." Nature **408**(6810): 325-30.

Fukuoka, Y., H. Inaoka, et al. (2004). "Inter-species differences of co-expression of neighboring genes in eukaryotic genomes." BMC Genomics **5**(1): 4.

Garcia, L. R., B. LeBoeuf, et al. (2007). "Diversity in mating behavior of hermaphroditic and male-female Caenorhabditis nematodes." Genetics **175**(4): 1761-71.

Garg, R., R. K. Patel, et al. (2011). "Gene discovery and tissue-specific transcriptome analysis in chickpea with massively parallel pyrosequencing and web resource development." Plant Physiol **156**(4): 1661-78.

Geldziler, B., I. Chatterjee, et al. (2006). "A comparative study of sperm morphology, cytology and activation in *Caenorhabditis elegans, Caenorhabditis remanei* and *Caenorhabditis briggsae*." Dev Genes Evol **216**(4): 198-208.

Gerstein, M. B., Z. J. Lu, et al. (2010). "Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project." Science **330**(6012): 1775-87.

Gibson, G. and B. Weir (2005). "The quantitative genetics of transcription." Trends Genet **21**(11): 616-23.

Good, J. M. and M. W. Nachman (2005). "Rates of protein evolution are positively correlated with developmental timing of expression during mouse spermatogenesis." Mol Biol Evol **22**(4): 1044-52.

Grabherr, M. G., B. J. Haas, et al. (2011). "Full-length transcriptome assembly from RNA-Seq data without a reference genome." Nat Biotechnol **29**(7): 644-52.

Graustein, A., J. M. Gaspar, et al. (2002). "Levels of DNA polymorphism vary with mating system in the nematode genus Caenorhabditis." Genetics **161**(1): 99-107.

Guo, Y., S. Lang, et al. (2009). "Independent recruitment of F box genes to regulate hermaphrodite development during nematode evolution." Curr Biol **19**(21): 1853-60.

Hahn, M. W. and G. C. Lanzaro (2005). "Female-biased gene expression in the malaria mosquito Anopheles gambiae." Curr Biol **15**(6): R192-3.

Hao, D. C., G. Ge, et al. (2011). "The first insight into the tissue specific Taxus transcriptome via Illumina second generation sequencing." PLoS ONE **6**(6).

Hill, R. C., C. E. de Carvalho, et al. (2006). "Genetic flexibility in the convergent evolution of hermaphroditism in Caenorhabditis nematodes." Dev Cell **10**(4): 531-8.

Hillier, L. W., A. Coulson, et al. (2005). "Genomics in *C. elegans*: so many genes, such a little worm." Genome Res **15**(12): 1651-60.

Hillier, L. W., R. D. Miller, et al. (2007). "Comparison of *C. elegans* and *C. briggsae* genome sequences reveals extensive conservation of chromosome organization and synteny." PLoS Biol **5**(7): e167.

Hodgkin, J. (1983). "Male Phenotypes and Mating Efficiency in CAENORHABDITIS ELEGANS." Genetics **103**(1): 43-64.

Hodgkin, J. and T. Doniach (1997). "Natural variation and copulatory plug formation in Caenorhabditis elegans." Genetics **146**(1): 149-64.

Hodgkin, J., H. R. Horvitz, et al. (1979). "Nondisjunction Mutants of the Nematode CAENORHABDITIS ELEGANS." Genetics **91**(1): 67-94.

Hu, T. T., P. Pattyn, et al. (2011). "The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change." Nat Genet **43**(5): 476-81.

Hunt-Newbury, R., R. Viveiros, et al. (2007). "High-throughput in vivo analysis of gene expression in *Caenorhabditis elegans*." PLoS Biol **5**(9): e237.

Iorizzo, M., D. A. Senalik, et al. (2011). "*De novo* assembly and characterization of the carrot transcriptome reveals novel genes, new markers, and genetic diversity." BMC Genomics **12**: 389.

Itoh, Y., E. Melamed, et al. (2007). "Dosage compensation is less effective in birds than in mammals." J Biol **6**(1): 2.

Itoh, Y., K. Replogle, et al. (2010). "Sex bias and dosage compensation in the zebra finch versus chicken genomes: general and specialized patterns among birds." Genome Res **20**(4): 512-8.

Jiang, M., J. Ryu, et al. (2001). "Genome-wide analysis of developmental and sex-regulated gene expression profiles in *Caenorhabditis elegans*." Proc Natl Acad Sci U S A **98**(1): 218-23.

Jiang, Z. F. and C. A. Machado (2009). "Evolution of sex-dependent gene expression in three recently diverged species of Drosophila." Genetics **183**(3): 1175-85.

Johnson, T. E. and E. W. Hutchinson (1993). "Absence of strong heterosis for life span and other life history traits in *Caenorhabditis elegans*." Genetics **134**(2): 465-74.

Johnson, T. E. and W. B. Wood (1982). "Genetic analysis of life-span in *Caenorhabditis elegans*." Proc Natl Acad Sci U S A **79**(21): 6603-7.

Johnston, J. S., A. E. Pepper, et al. (2005). "Evolution of genome size in Brassicaceae." Ann Bot **95**(1): 229-35.

Jovelin, R., B. C. Ajie, et al. (2003). "Molecular evolution and quantitative variation for chemosensory behaviour in the nematode genus Caenorhabditis." Mol Ecol **12**(5): 1325-37.

Jovelin, R., J. P. Dunham, et al. (2009). "High nucleotide divergence in developmental regulatory genes contrasts with the structural elements of olfactory pathways in Caenorhabditis." Genetics **181**(4): 1387-97.

Kall, L., A. Krogh, et al. (2004). "A combined transmembrane topology and signal peptide prediction method." J Mol Biol **338**(5): 1027-36.

Kamath, R. S., A. G. Fraser, et al. (2003). "Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi." Nature **421**(6920): 231-7.

Kamath, R. S., M. Martinez-Campos, et al. (2001). "Effectiveness of specific RNA-mediated interference through ingested double-stranded RNA in *Caenorhabditis elegans*." Genome Biol **2**(1):

Khaitovich, P., I. Hellmann, et al. (2005). "Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees." Science **309**(5742): 1850-4.

Kimura, M. and T. Ota (1971). "Theoretical aspects of population genetics." Monogr Popul Biol **4**: 1-219.

Kiontke, K., M.-A. Félix, et al. (in revision). "A phylogeny and molecular barcodes for Caenorhabditis, with numerous new species from rotting fruits." BMC Evolutionary Biology.

Kiontke, K. and D. H. Fitch (2005). "The phylogenetic relationships of Caenorhabditis and other rhabditids." WormBook: 1-11.

Kiontke, K., N. P. Gavin, et al. (2004). "Caenorhabditis phylogeny predicts convergence of hermaphroditism and extensive intron loss." Proc Natl Acad Sci U S A **101**(24): 9003-8.

Kleeman, G. A. and A. L. Basolo (2007). "Facultative decrease in mating resistance in hermaphroditic *Caenorhabditis elegans* with self-sperm depletion." Animal Behaviour **74**(5): 1339-1347.

L'Hernault, S. W. (2006). "Spermatogenesis." WormBook: 1-14.

LaMunyon, C. W. and S. Ward (1998). "Larger sperm outcompete smaller sperm in the nematode *Caenorhabditis elegans*." Proc Biol Sci **265**(1409): 1997-2002.

LaMunyon, C. W. and S. Ward (1999). "Evolution of sperm size in nematodes: sperm competition favours larger sperm." Proc Biol Sci **266**(1416): 263-7.

LaMunyon, C. W. and S. Ward (2002). "Evolution of larger sperm in response to experimentally increased sperm competition in *Caenorhabditis elegans*." Proc Biol Sci **269**(1496): 1125-8.

Langmead, B., C. Trapnell, et al. (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." Genome Biol **10**(3): R25.

Leder, E. H., J. M. Cano, et al. (2010). "Female-biased expression on the X chromosome as a key step in sex chromosome evolution in threespine sticklebacks." Mol Biol Evol **27**(7): 1495-503.

Lercher, M. J., T. Blumenthal, et al. (2003). "Coexpression of neighboring genes in *Caenorhabditis elegans* is mostly due to operons and duplicate genes." Genome Res **13**(2): 238-43.

Li, H., A. Coghlan, et al. (2006). "TreeFam: a curated database of phylogenetic trees of animal gene families." Nucleic Acids Res **34**(Database issue): D572-80.

Li, R., W. Fan, et al. (2010). "The sequence and *de novo* assembly of the giant panda genome." Nature **463**(7279): 311-7.

Li, R., Y. Li, et al. (2008). "SOAP: short oligonucleotide alignment program." Bioinformatics **24**(5): 713-4.

Li, R., H. Zhu, et al. (2010). "*De novo* assembly of human genomes with massively parallel short read sequencing." Genome Res **20**(2): 265-72.

Lints, R. and S. W. Emmons (2002). "Regulation of sex-specific differentiation and mating behavior in *C. elegans* by a new member of the DM domain transcription factor family." Genes Dev **16**(18): 2390-402.

Lipton, J., G. Kleemann, et al. (2004). "Mate searching in *Caenorhabditis elegans:* a genetic model for sex drive in a simple invertebrate." J Neurosci **24**(34): 7427-34.

Lister, R., R. C. O'Malley, et al. (2008). "Highly integrated single-base resolution maps of the epigenome in Arabidopsis." Cell **133**(3): 523-36.

Liu, X., D. L. Brutlag, et al. (2001). "BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes." Pac Symp Biocomput: 127-38.

Lynch, M. (2007). The origins of genome architecture, Sinauer Associates Inc.

Lynch, M. and J. S. Conery (2003). "The origins of genome complexity." Science **302**(5649): 1401-4.

Lysak, M. A., M. A. Koch, et al. (2009). "The dynamic ups and downs of genome size evolution in Brassicaceae." Mol Biol Evol **26**(1): 85-98.

Mable, B. K. (2008). "Genetic causes and consequences of the breakdown of self-incompatibility: case studies in the Brassicaceae." Genet Res (Camb) **90**(1): 47-60.

Maeda, I., Y. Kohara, et al. (2001). "Large-scale analysis of gene function in *Caenorhabditis elegans* by high-throughput RNAi." Curr Biol **11**(3): 171-6.

Malone, J. H., D. L. Hawkins, Jr., et al. (2006). "Sex-biased gene expression in a ZW sex determination system." J Mol Evol **63**(4): 427-36.

Mank, J. E. and H. Ellegren (2009). "All dosage compensation is local: gene-by-gene regulation of sex-biased expression on the chicken Z chromosome." Heredity **102**(3): 312-20.

Mank, J. E., L. Hultin-Rosenberg, et al. (2007). "Rapid evolution of female-biased, but not male-biased, genes expressed in the avian brain." Mol Biol Evol **24**(12): 2698-706.

Marioni, J. C., C. E. Mason, et al. (2008). "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays." Genome Res **18**(9): 1509-17.

Martin, J. A. and Z. Wang (2011). "Next-generation transcriptome assembly." Nat Rev Genet **12**(10): 671-82.

McKay, S. J., R. Johnsen, et al. (2003). "Gene expression profiling of cells, tissues, and developmental stages of the nematode *C. elegans*." Cold Spring Harb Symp Quant Biol **68**: 159-69.

Meisel, R. P. (2011). "Towards a more nuanced understanding of the relationship between sex-biased gene expression and rates of protein-coding sequence evolution." Mol Biol Evol **28**(6): 1893-900.

Merritt, C., D. Rasoloson, et al. (2008). "3' UTRs are the primary regulators of gene expression in the *C. elegans* germline." Curr Biol **18**(19): 1476-82.

Merritt, C. and G. Seydoux (2010). "Transgenic solutions for the germline." WormBook: 1-21.

Morin, R., M. Bainbridge, et al. (2008). "Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing." Biotechniques **45**(1): 81-94.

Morran, L. T., B. J. Cappy, et al. (2009). "Sexual partners for the stressed: facultative outcrossing in the self-fertilizing nematode *Caenorhabditis elegans*." Evolution **63**(6): 1473-82.

Morran, L. T., M. D. Parmenter, et al. (2009). "Mutation load and rapid adaptation favour outcrossing over self-fertilization." Nature **462**(7271): 350-2.

Mortazavi, A., B. A. Williams, et al. (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq." Nat Methods **5**(7): 621-8.

Nagalakshmi, U., Z. Wang, et al. (2008). "The transcriptional landscape of the yeast genome defined by RNA sequencing." Science **320**(5881): 1344-9.

Naurin, S., B. Hansson, et al. (2011). "The sex-biased brain: sexual dimorphism in gene expression in two species of songbirds." BMC Genomics **12**: 37.

Nayak, S., J. Goree, et al. (2005). "*fog-2* and the evolution of self-fertile hermaphroditism in Caenorhabditis." PLoS Biol **3**(1): e6.

Newberg, L. A., W. A. Thompson, et al. (2007). "A phylogenetic Gibbs sampler that yields centroid solutions for cis-regulatory site prediction." Bioinformatics **23**(14): 1718-27.

Nigon, V. (1951). "Polyploidie experimentale chez un nematode libre, Rhabditis elegans maupas." Bull. Biol. Fr. Belg. **85**: 187–255.

Nisbet, A. J., P. A. Cottee, et al. (2008). "Genomics of reproduction in nematodes: prospects for parasite intervention?" Trends Parasitol **24**(2): 89-95.

Nishimura, H. and S. W. L'Hernault (2010). "Spermatogenesis-defective (spe) mutants of the nematode *Caenorhabditis elegans* provide clues to solve the puzzle of male germline functions during reproduction." Dev Dyn **239**(5): 1502-14.

Oshlack, A., M. D. Robinson, et al. (2010). "From RNA-seq reads to differential expression results." Genome Biol **11**(12): 220.

Ossowski, S., K. Schneeberger, et al. (2010). "The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*." Science **327**(5961): 92-4.

Oyama, R., M. Clauss, et al. (2008). "The shrunken genome of *Arabidopsis thaliana*." Plant Systematics and Evolution **273**(3): 257-271.

Palopoli, M. F., M. V. Rockman, et al. (2008). "Molecular basis of the copulatory plug polymorphism in *Caenorhabditis elegans*." Nature **454**(7207): 1019-22.

Pannell, J. R. (2002). "The evolution and maintenance of androdioecy." Annual Review of Ecology and Systematics **33**: 397-425.

Parisi, M., R. Nuttall, et al. (2003). "Paucity of genes on the Drosophila X chromosome showing male-biased expression." Science **299**(5607): 697-700.

Pavesi, G., P. Mereghetti, et al. (2006). "MoD Tools: regulatory motif discovery in nucleotide sequences from co-regulated or homologous genes." Nucleic Acids Res **34**(Web Server issue): W566-70.

Pavesi, G., F. Zambelli, et al. (2007). "WeederH: an algorithm for finding conserved regulatory motifs and regions in homologous sequences." BMC Bioinformatics **8**: 46.

Perez-Enciso, M., A. L. Ferraz, et al. (2009). "Impact of breed and sex on porcine endocrine transcriptome: a bayesian biometrical analysis." BMC Genomics **10**: 89.

Pollak, E. (1987). "On the theory of partially inbreeding finite populations. I. Partial selfing." Genetics **117**(2): 353-60.

Praitis, V. (2006). "Creation of transgenic lines using microparticle bombardment methods." Methods Mol Biol **351**: 93-107.

Ramakers, C., J. M. Ruijter, et al. (2003). "Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data." Neurosci Lett **339**(1): 62-6.

Ranz, J. M., C. I. Castillo-Davis, et al. (2003). "Sex-dependent gene expression and evolution of the Drosophila transcriptome." Science **300**(5626): 1742-5.

Reinke, V., I. S. Gil, et al. (2004). "Genome-wide germline-enriched and sex-biased expression profiles in *Caenorhabditis elegans*." Development **131**(2): 311-23.

Reinke, V., H. E. Smith, et al. (2000). "A global profile of germline gene expression in *C. elegans*." Mol Cell **6**(3): 605-16.

Rinn, J. L., J. S. Rozowsky, et al. (2004). "Major molecular differences between mammalian sexes are involved in drug metabolism and renal function." Dev Cell **6**(6): 791-800.

Roberts, A., C. Trapnell, et al. (2011). "Improving RNA-Seq expression estimates by correcting for fragment bias." Genome Biol **12**(3): R22.

Rockman, M. V. and L. Kruglyak (2009). "Recombinational landscape and population genomics of Caenorhabditis elegans." PLoS Genet **5**(3): e1000419.

Rose, A. M. and D. L. Baillie (1979). "Effect of temperature and parental age on recombination and nondisjunction in *Caenorhabditis elegans.*" Genetics **92**:409-418

Ross, J. A., D. C. Koboldt, et al. (2011). "*Caenorhabditis briggsae* recombinant inbred line genotypes reveal inter-strain incompatibility and the evolution of recombination." PLoS Genet **7**(7): e1002174.

Ruan, J., H. Li, et al. (2008). "TreeFam: 2008 Update." Nucleic Acids Res **36(**Database issue): D735-40.

Ruijter, J. M., C. Ramakers, et al. (2009). "Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data." Nucleic Acids Res **37**(6): e45.

Rumble, S. M., P. Lacroute, et al. (2009). "SHRiMP: accurate mapping of short color-space reads." PLoS Comput Biol **5**(5): e1000386.

Ruvinsky, I. and G. Ruvkun (2003). "Functional tests of enhancer conservation between distantly related species." Development **130**(21): 5133-42.

Salzberg, S. L. (2010). "Recent advances in RNA sequence analysis." F1000 Biol Rep **2**: 64.

Sandelin, A., W. Alkema, et al. (2004). "JASPAR: an open-access database for eukaryotic transcription factor binding profiles." Nucleic Acids Res. **32**(Suppl 1): D91-D94.

Sassaman, C. and S. C. Weeks (1993). "The genetic mechanism of sex determination in the conchostracan shrimp *Eulimnadia texana*." Am Nat **141**(2): 314-28.

Savolainen, O., C. H. Langley, et al. (2000). "Contrasting patterns of nucleotide polymorphism at the alcohol dehydrogenase locus in the outcrossing

Arabidopsis lyrata and the selfing *Arabidopsis thaliana*." Mol Biol Evol **17**(4): 645-55.

Schedl, T. and J. Kimble (1988). "*fog-2*, a germ-line-specific sex determination gene required for hermaphrodite spermatogenesis in *Caenorhabditis elegans*." Genetics **119**(1): 43-61.

Schug, J. and G. C. Overton (1997). TESS: Transcription Element Search Software on the WWW, Computational Biology and Informatics Laboratory
School of Medicine
University of Pennsylvania.

Schultz, N., F. K. Hamra, et al. (2003). "A multitude of genes expressed solely in meiotic or postmeiotic spermatogenic cells offers a myriad of contraceptive targets." Proc Natl Acad Sci U S A **100**(21): 12201-6.

Semple, J. I., R. Garcia-Verdugo, et al. (2010). "Rapid selection of transgenic *C. elegans* using antibiotic resistance." Nat Methods **7**(9): 725-7.

Sigrist, C. J., L. Cerutti, et al. (2010). "PROSITE, a protein domain database for functional characterization and annotation." Nucleic Acids Res **38**(Database issue): D161-6.

Simon, J. M. and P. W. Sternberg (2002). "Evidence of a mate-finding cue in the hermaphrodite nematode *Caenorhabditis elegans*." Proc Natl Acad Sci U S A **99**(3): 1598-603.

Simpson, J. T., K. Wong, et al. (2009). "ABySS: A parallel assembler for short read sequence data." Genome Res **19**(6): 1117-1123.

Sivasundar, A. and J. Hey (2005). "Sampling from natural populations with RNAI reveals high outcrossing and population structure in *Caenorhabditis elegans*." Curr Biol **15**(17): 1598-602.

Sonnichsen, B., L. B. Koski, et al. (2005). "Full-genome RNAi profiling of early embryogenesis in *Caenorhabditis elegans*." Nature **434**(7032): 462-9.

Srinivasan, J., F. Kaplan, et al. (2008). "A blend of small molecules regulates both mating and development in *Caenorhabditis elegans*." Nature **454**(7208): 1115-8.

Stein, L. D., Z. Bao, et al. (2003). "The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics." PLoS Biol **1**(2): E45.

Stevenson, T. O., K. B. Mercer, et al. (2007). "*unc-94* encodes a tropomodulin in *Caenorhabditis elegans*." J Mol Biol **374**(4): 936-50.

Stewart, A. D. and P. C. Phillips (2002). "Selection and maintenance of androdioecy in *Caenorhabditis elegans*." Genetics **160**(3): 975-82.

Sultan, M., M. H. Schulz, et al. (2008). "A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome." Science **321**(5891): 956-60.

Tavernarakis, N., S. L. Wang, et al. (2000). "Heritable and inducible genetic interference by double-stranded RNA encoded by transgenes." Nat Genet **24**(2): 180-3.

Teotonio, H., D. Manoel, et al. (2006). "Genetic variation for outcrossing among *Caenorhabditis elegans* isolates." Evolution **60**(6): 1300-5.

Thoemke, K., W. Yi, et al. (2005). "Genome-wide analysis of sex-enriched gene expression during *C. elegans* larval development." Dev Biol **284**(2): 500-8.

Thomas, C. G., R. Li, et al. (In prep.). "Genome shrinkage and desexualisation in self-fertile Caenorhabditis nematodes."

Thomas, J. H. (2008). "Genome evolution in Caenorhabditis." Brief Funct Genomic Proteomic **7**(3): 211-6.

Thompson, W., M. J. Palumbo, et al. (2004). "Decoding human regulatory circuits." Genome Res **14**(10A): 1967-74.

Thompson, W., E. C. Rouchka, et al. (2003). "Gibbs Recursive Sampler: finding transcription factor binding sites." Nucleic Acids Res **31**(13): 3580-5.

Thompson, W. A., L. A. Newberg, et al. (2007). "The Gibbs Centroid Sampler." Nucleic Acids Res **35**(Web Server issue): W232-7.

Torgerson, D. G., R. J. Kulathinal, et al. (2002). "Mammalian sperm proteins are rapidly evolving: evidence of positive selection in functionally diverse genes." Mol Biol Evol **19**(11): 1973-80.

Trapnell, C., L. Pachter, et al. (2009). "TopHat: discovering splice junctions with RNA-Seq." Bioinformatics **25**(9): 1105-11.

Trapnell, C., B. A. Williams, et al. (2010). "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation." Nat Biotechnol **28**(5): 511-5.

Wang, J., P. J. Chen, et al. (2010). "Chromosome size differences may affect meiosis and genome size." Science **329**(5989): 293.

Wang, Z., M. Gerstein, et al. (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nat Rev Genet **10**(1): 57-63.

Ward, S. and J. S. Carrel (1979). "Fertilization and sperm competition in the nematode *Caenorhabditis elegans*." Dev Biol **73**(2): 304-21.

Ward, S., E. Hogan, et al. (1983). "The initiation of spermiogenesis in the nematode *Caenorhabditis elegans*." Dev Biol **98**(1): 70-9.

Weeks, S. C. (2004). "Levels of inbreeding depression over seven generations of selfing in the androdioecious clam shrimp, *Eulimnadia texana*." J Evol Biol **17**(3): 475-84.

Weeks, S. C., C. Benvenuto, et al. (2006). "When males and hermaphrodites coexist: a review of androdioecy in animals." Integr Comp Biol **46**(4): 449-64.

Weeks, S. C., E. G. Chapman, et al. (2009). "Evolutionary transitions among dioecy, androdioecy and hermaphroditism in limnadiid clam shrimp (Branchiopoda: Spinicaudata)." J Evol Biol **22**(9): 1781-99.

Weeks, S. C., V. Marcus, et al. (1999). "Inbreeding depression in a self-compatible, androdioecious crustacean, *Eulimnadia texana*." Evolution **53**(2): 472-483.

Wegewitz, V., H. Schulenburg, et al. (2008). "Experimental insight into the proximate causes of male persistence variation among two strains of the androdioecious *Caenorhabditis elegans* (Nematoda)." BMC Ecol **8**: 12.

Wilhelm, B. T., S. Marguerat, et al. (2008). "Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution." Nature **453**(7199): 1239-43.

Winston, W. M., M. Sutherlin, et al. (2007). "*Caenorhabditis elegans* SID-2 is required for environmental RNA interference." Proc Natl Acad Sci U S A **104**(25): 10565-70.

Wolf, J. B. and J. Bryk (2011). "General lack of global dosage compensation in ZZ/ZW systems? Broadening the perspective with RNA-seq." BMC Genomics **12**: 91.

Wong, M. M., C. H. Cannon, et al. (2011). "Identification of lignin genes and regulatory sequences involved in secondary cell wall formation in *Acacia auriculiformis* and *Acacia mangium* via *de novo* transcriptome sequencing." BMC Genomics **12**: 342.

Wood, W. B. (1988). The nematode *Caenorhabditis elegans*. NY, Cold Spring Harbor Laboratory Press.

Woodruff, G. C., O. Eke, et al. (2010). "Insights into species divergence and the evolution of hermaphroditism from fertile interspecies hybrids of Caenorhabditis nematodes." Genetics **186**(3): 997-1012.

WormBase (release WS224). http://www.WormBase.org.

Wright, S. I., B. Lauga, et al. (2002). "Rates and patterns of molecular evolution in inbred and outbred Arabidopsis." Mol Biol Evol **19**(9): 1407-20.

Wright, S. I., Q. H. Le, et al. (2001). "Population dynamics of an Ac-like transposable element in self- and cross-pollinating Arabidopsis." Genetics **158**(3): 1279-88.

Wright, S. I. and D. J. Schoen (1999). "Transposon dynamics and the breeding system." Genetica **107**(1-3): 139-48.

Wu, H., H. Yang, et al. (2011). *maanova*: Tools for analyzing micro array experiments.

Yang, X., E. E. Schadt, et al. (2006). "Tissue-specific expression and regulation of sexually dimorphic genes in mice." Genome Res **16**(8): 995-1004.

Yi, K., S. M. Buttery, et al. (2007). "A Ser/Thr kinase required for membrane-associated assembly of the major sperm protein motility apparatus in the amoeboid sperm of Ascaris." Mol Biol Cell **18**(5): 1816-25.

Yi, K., X. Wang, et al. (2009). "Dephosphorylation of major sperm protein (MSP) fiber protein 3 by protein phosphatase 2A during cell body retraction in the MSP-based amoeboid motility of Ascaris sperm." Mol Biol Cell **20**(14): 3200-8.

Yi, W. and D. Zarkower (1999). "Similarity of DNA binding and transcriptional regulation by *Caenorhabditis elegans* MAB-3 and Drosophila melanogaster DSX suggests conservation of sex determining mechanisms." Development **126**(5): 873-81.

Young, J. M. and I. A. Hope (1993). "Molecular markers of differentiation in *Caenorhabditis elegans* obtained by promoter trapping." Dev Dyn **196**(2): 124-32.

Zarkower, D. and J. Hodgkin (1993). "Zinc fingers in sex determination: only one of the two *C. elegans* TRA-! proteins binds DNA in vitro." Nucleic Acids Res **21**(16): 3691-8.

Zerbino, D. R. and E. Birney (2008). "Velvet: algorithms for de novo short read assembly using de Bruijn graphs." Genome Res **18**(5): 821-9.

Zha, X., Q. Xia, et al. (2009). "Dosage analysis of Z chromosome genes using microarray in silkworm, *Bombyx mori*." Insect Biochem Mol Biol **39**(5-6): 315-21.

Zhang, Y., D. Sturgill, et al. (2007). "Constraint and turnover in sex-biased gene expression in the genus Drosophila." Nature **450**(7167): 233-7.

Zhang, Z. and J. Parsch (2005). "Positive correlation between evolutionary rate and recombination rate in Drosophila genes with male-biased expression." Mol Biol Evol **22**(10): 1945-7.