UNIVERSITY OF NOTTINGHAM

Left-right asymmetry variation in the pond snail, *Lymnaea stagnalis*: exploring patterns of gene expression

Harriet F Johnson

Thesis submitted for the degree of Doctor of Philosophy, November 2015

Thesis abstract

The establishment of left-right (LR) asymmetry in animal development remains an unanswered, fundamental question in biology. Many mechanisms of symmetry-breaking have been proposed and supported, although as yet no universal mechanism has been verified across bilaterian animals. Snails provide an invaluable study organism for understanding LR asymmetry, due to the prevalence of chirally variable species. In the pond snail *Lymnaea stagnalis* LR asymmetry and resulting shell-coiling direction is a well described genetically tractable trait, inherited through a maternal effect. However, the 'chirality gene' is still unknown.

In *L. stagnalis*, clockwise (dextral) coiling is the dominant genotype, therefore snails with homozygote genotype '*DD*' or heterozygote '*Dd*' both produce dextral offspring, whereas those with the homozygote recessive genotype '*dd*' have anticlockwise (sinistral) coiling offspring. To further the Davison research group's ongoing characterisation of the chirality gene in *L. stagnalis*, this project focussed on gene expression patterns exhibited between chiral genotypes. Differential gene expression was explored via a candidate gene approach, performing quantitative real-time PCR (qPCR) experiments on specific genes of interest, and also a transcriptomic sweep, utilising next generation sequencing.

To enable accurate quantification of gene expression by relative qPCR, first, stable endogenous control genes had to be established. In light of general failings of the previously published control genes to meet the criteria for appropriate use of qPCR, five genes were verified for use as stable endogenous controls in *L. stagnalis* embryo, ovotestis and foot tissue, for the accurate comparison of gene expression between and within chiral genotypes. These endogenous control genes will enable other researchers of *L. stagnalis* to rapidly identify stable controls for relative qPCR experiments.

qPCR experiments were performed to compare gene expression of 13 candidate genes between chiral genotypes in the single-cell embryo, ovotestis and foot tissues. Significant differential expression was observed between chiral genotypes only in the diaphanous related formin gene, *Ldia2*, and two actin-related protein genes, *Larp2/3 1a* and *Larp2/3 3*.

A frameshift mutation in the sinistral copy of *Ldia2*, discovered by the Davison research group, has identified *Ldia2* as the primary candidate for the causal gene in LR asymmetry determination in *L. stagnalis*. In support of this, *Ldia2* mRNA was found to be dramatically underrepresented in the sinistral one cell embryo and significantly reduced in the sinistral ovotestis tissue, yet not in the somatic foot tissue. *Ldia2* was also the only gene found to be overrepresented in the embryo tissue

relative to the ovotestis and foot tissue, providing further support for the functional importance the gene in early development. The expression level of *Ldia2* in the heterozygote genotype groups was calculated to be halfway between that of the homozygote groups, indicating equal expression dominance of the alleles at the chirality locus. The expression pattern observed in the actin-related proteins was less clear and will require further analysis to infer any true biological meaning. However due to the close interaction of actin-related proteins and formins the differential expression observed in the embryo tissue provides functional support for the role of *Ldia2* in chiral dimorphism.

Next generation transcriptome sequencing methods were employed to gain a transcriptome-wide scan of patterns of gene expression in the ovotestis tissue of snails of differing chiral genotype. A comparative analysis was initiated trialling a novel reduced-representation sequencing method, expression RAD sequencing (eRAD) and traditional RNA Seq. eRAD applies the method of restriction-site associated DNA Sequencing (RADSeq) to the transcriptome by utilising double-stranded complementary DNA (cDNA) in place of genomic DNA. Due to delays in sequencing, the RNA Seq data was not received in sufficient time to perform the comparative assessment within this thesis. Consequently, only the eRAD data is presented here.

The eRAD data failed to identify reliable differences in gene expression between chiral genotypes, although did provide a transcriptomic resource of *de novo* assembled contigs, which has been verified through further analyses. Overall the lack of differential expression identified between chiral genotypes in both the qPCR and eRAD analyses has indicated that the sinistral morph of *L. stagnalis* does not exhibit a large-scale loss of gene function and pleiotropic effects on gene expression. Therefore, the negative consequences of chiral reversal in *L. stagnalis*, such as the low hatch rate observed in sinistral broods, may all result from the single chirality gene polymorphism.

Declaration of own work

Except where clearly specified all practical experiments and data analyses were performed by me, although passive voice has been employed throughout this thesis in keeping with convention of scientific writing.

Acknowledgements

Firstly, I would like to express my gratitude to the BBSRC and the University of Nottingham for funding this PhD project, without which my doctoral training would not have been possible.

I would also like to thank my supervisor Dr Angus Davison and other members of the Davison research group past and present for stimulating discussions and invaluable experience. I would like to thank especially Dr Maureen (Mengning) Liu who initiated the eRAD sequencing trial, which was adopted by me in her absence, and of course the Genome Analysis Centre (TGAC) for carrying out the Illumina sequencing runs.

Kind acknowledgements go to Dr Teri Evans who provided the transcriptome annotation analysis via Blast2GO, of which the program would likely still be running now with the free version.

Although the *in situ* experiments did not produce many biologically informative results during the exchange, my time spent at the Jackson lab in Göttingen was a progressive opportunity to learn a new experimental method (and I do love methods). Additionally, the embryo manipulation skills I developed proved essential to the major findings of this project. The lab exchange itself provided invaluable insight into the German academic system and Germany itself, and I am incredibly grateful for being welcomed so warmly by all at the research department. Thank you to the Jackson group members: Dr Jenny Hohagen, Juliane Germer and especially to Ines Herlitze who dedicated an immeasurable amount of time helping me, and of course to Dr Daniel Jackson for hosting and the generous funding from the German Academic Exchange Service (DAAD) and Boehringer Ingelheim.

I have been fortunate during my time spent at the University of Nottingham to partake in a number of additional ventures including the Biotechnology YES! scheme gaining insight into entrepreneurship in biology and an incredible two-month internship at the Naked Scientists, kindly funded by the Genetics Society, developing indispensable science communication skills not to mention live radio presenting.

For a first introduction to qPCR I would like to thank Amy Hall, and for further debates about analysis methods Shaun Robertson. Furthermore, PrimerDesign Ltd has been notably helpful in the design

and implementation of qPCR experiments and also financially in awarding the 'silver sponsorship' every year.

Thank you also to the members of the evolution and ecology group within the school of Life Sciences at the University of Nottingham (now undoubtedly under a different name but you know who you are, the ones with the cake), for generally making the department a positive place to be.

Finally, I would like to express my endless gratitude to Oliver Page for being just the nicest guy ever. The last few months of writing this thesis have felt like a long time but forever is a long, long time.

"*All scientific work is incomplete - whether it be observational or experimental. All scientific work is liable to be upset or modified by advancing knowledge. That does not confer upon us a freedom to ignore the knowledge we already have, or to postpone the action that it appears to demand at a given time.*" – Sir Austin Bradford Hill, 1965

"*Je n'ai fait celle-ci plus longue que parce que je n'ai pas eu le loisir de la faire plus courte*" – Blaise Pascal, 1656

Table of Contents

Chapter 1:
General Introduction

Left-Right asymmetry

Defining Chirality

Left-right (LR) asymmetry occurs frequently in nature, observed on varying scales ranging from the directional spiralling of galaxies to the structure of the molecules they are composed of.

The term 'chiral' was coined to describe any geometrical figure as having 'chirality' if its mirror-image cannot be superimposed onto itself (Kelvin 1904, Thompson 1910, cited in McManus 2002). A relatable example of a chiral object is the human hand. When looking at the pair of hands together they reflect mirror images of each other (with the exception of a few environmental differences/fluctuating asymmetries), as such one hand cannot be put on top of the other without inversion.

Chirality is frequently discussed in molecular chemistry. Pasteur in the 1840s recognised that many organic molecules are found in two mirror-image structural forms (McManus 2002) (Figure 1a). It was found that laboratory synthesised chiral molecules would produce equal amounts of left-handed (L- for laevo) and right-handed (D- for dextro) forms, creating a 50/50 racemic mixture. Biosynthetic processes however only produced one of the two chiral forms. This chiral consistency is observed in all living things, which are made of entirely L-amino acids and D-sugars (Fischer 1894, cited in Mason 1991). The dominance of L-amino acids has been observed to extend to those found on extra-terrestrial asteroids (Engel and Macko 1997). The question of whether this chiral consistency on a molecular level leads to limitations on larger level asymmetries remains debated.

Chiral organisms

The bilataria represent 99% of animal species, most often described as having originated from a radially symmetrical common ancestor (Ruppert, Fox et al. 2004). This however, is still a subject of debate, as is the overall phylogeny of the origins of the metazoa.

The bilateral body plan, exhibits external bilateral symmetry, which is present when only one plane of bisection, produces left and right mirror images (Figure 1c). Although externally bilaterally symmetrical, the vast majority of bilatarians exhibit LR asymmetry in their internal organisation and

of the organs themselves (Figure 1d). This internal asymmetry represents a fundamental feature of multicellular organisms and is believed to further date back to the earliest life on Earth (Babcock 2005). It is still open to discussion whether bilateral symmetry was superimposed onto an originally asymmetrical body plan, ie. an asymmetrical common ancestor, or vice versa (Wolpert, 1991).

There is interesting variation in the early developmental body plans within the metazoa. Some examples would indicate an initially symmetrical body plan, such as the equal radially cleaving embryos of the deuterostomes, which later develop LR asymetries in body plan. Conversely there are also initially asymmetrical embryos which later aquire bilateral symmetry, for example within nematodes and ciliates (Wolpert, 1991). Cnidarians provide an insightful reference point in this debate as they are 'pre-bilaterians'. Interestingly, the most radially symmetrical species, are found within the most derived group of cnidarians, the Hydrozoa (Martindale, 2005). This supports asymmetry as the primitive state in the common ancestor prior to cnidarians and bilatarians. Corbalis and Morgan (1978) argued that lateral asymmetries result from a LR maturation gradient, in which there is earlier and more rapid development on the left side compared to the right. The resulting fundamental LR asymmetry is no different to the mechanisms required for the determination of the AP or DV axis specification. Many big questions remain regarding the origins of axial patterning, largely as a consequence of, yet also a contributing factor to, the remaining uncertainty within the phylogeny of the early metazoa.

It is important to define the difference between primary asymmetry and secondary asymmetry. The former refers to the polarity of LR patterning in early development, which corresponds to subsequent visceral asymmetry of the heart and lungs, for example. Secondary asymmetry, such as the handedness of lobster claws (Govind 1989) or the mouth of scale eating fish (Hori 1993), develops independently of primary asymmetry and exhibits much higher levels of variation (Sutcharit, Asami et al. 2007).

Asymmetry also occurs at incrementing scales, ranging from organ positioning and brain lateralisation, to hair whorls and behavioural lateralisation such as handedness, all of which show varied and perplexing associations to visceral asymmetry (Neville 1976, McManus 2002). However, these lateralisation relationships are beyond the scope of this thesis. The focus here surrounds primary LR asymmetry in animal body plans, and more specifically how LR asymmetry can vary during early development of the animal body plan. Hereafter, the term asymmetry refers to primary/visceral asymmetry.

Potential benefits to asymmetrical internal organisation have been proposed, such as fluid dynamics in the heart (Kilner, Yang et al. 2000), however there is no apparent reason for the overwhelming consistency in the sidedness of LR asymmetry exhibited in the majority of animals.

Deviations from normal LR patterning can result in serious clinical consequences. The main classes of laterality defects that occur during early development in humans are described here. In the rare condition '*situs inversus totalis'*, present in 0.01% of the population (Burn 1991) the entire body plan is reversed, displaying a mirror-image of the normal internal organisation, '*situs solitus'*. No impairment to function is specifically caused by this condition (Torgersen 1950). However, the group of individuals with *situs inversus totalis* includes those that suffer from Kartagener's syndrome, which is linked to ciliary dysfunction (Kartagener and Stucki 1962, Afzelius 1976). Another condition; *situs ambiguus*, occurs from the failure to control asymmetry, resulting in independent LR patterning of organs. This condition is often associated with cases of *isomerism*, in which the individual displays LR symmetry across the midline, resulting in two left sides or two right sides (Burn 1991, Peeters and Devriendt 2006).

Due to the diversity and overlap of the morphological effects resulting from variant forms of heterotaxia, gaining accurate estimates of the occurrence of each condition in humans proves difficult, yet it is estimated that combined, LR laterality malformations effect 1 in 5000 births (Casey and Hackett 2000). *Situs ambiguus* and *isomerism* frequently result in cardiac and gastrointestinal defects. With 80% of *situs ambiguus* cases presenting complex congenital heart disease (Peeters and Devriendt 2006), LR laterality defects represent an important area of developmental research.

There is an unmistakable necessity for LR patterning to be conserved for the functioning of organs, yet in light of the lack of pathologies associated with *situs inversus totalis*, there is no clear indication for a need for directional asymmetry, especially between species (Wood 1997). True enantiomorphs; 'mirror-image' organisms displaying reversed primary asymmetry, have only been observed in gastropods and nematodes (Vermeij 1975, Robertson 1993, Wood 1997, Okumura, Utsuno et al. 2008) and are therefore key in exploring potential selection of chiral morphs and the evolutionary dynamics of chirality variation.

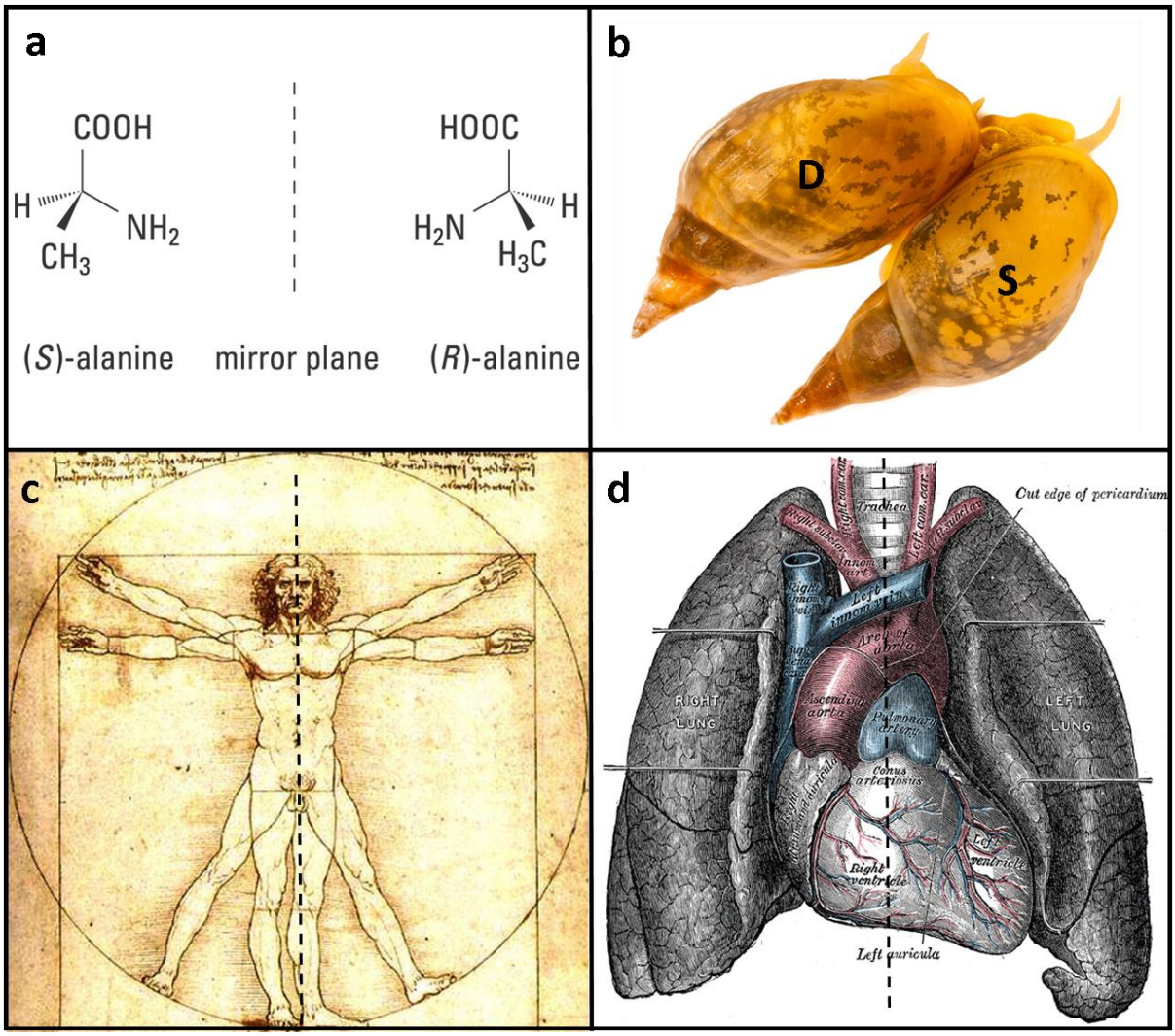


Figure 1 Examples of left-right patterning. a: The chiral amino acid alanine and it's mirror image counterpart, image courtesy of: The Nobel Prize in Chemistry 2001 - Popular Information (Nobelprize.org). b: Enantiomorphs of the pond snail *Lymnaea stagnalis*, the dextral form is indicated by a 'D' and the sinistral form by an 'S', photo credit: Ester de Roij (esterderoij@gmail.com). C: Leonardo Da Vinci's Vitruvian man c1490, original image credit: Luc Viatour (www.Lucnix.be), adapted to include indicator of bilateral plane of symmetry (dashed line). d: Situs solitus organisation of human heart and lungs, drawing from Gray's anatomy of the human body (Gray 1918), adapted to include indicator of bilateral plane of symmetry (dashed line).

Symmetry-breaking event

Telling left from right

How LR asymmetry is established in early development represents a fundamental question in developmental biology and has been an ongoing area of research for over a century (Crampton 1894). Because no macromolecular force differentiates left from right, it poses a puzzle of how the initially symmetrical embryo can orientate its LR axis consistently with respect to the dorsal-ventral (DV) and anterior-posterior (AP) axes. This problem was considered by the eighteenth century philosopher Immanuel Kant, who early on acknowledged that for left and right to be distinct, there must be an immovable reference point of absolute space (Harper 1991). A more modern example of

this same question has been posed as 'the Ozma problem'. Summarised, it asks whether there is any way to communicate the meaning of 'left' to an extra-terrestrial via radio with no common asymmetric reference (Gardner 1990).

There is general agreement that there are three steps in the generation of asymmetry in the developing embryo. Firstly, the radially symmetrical embryo undergoes a symmetry breaking event, in which the LR axis is generated relative to the already established dorsal-ventral (DV) and anterior-posterior (AP) axes. Secondly these asymmetries are translated into differential bilateral gene expression. Finally, the cascade of asymmetrical gene expression determines asymmetric organ positioning and morphology (Vandenberg, Lemire et al. 2013).

There are well documented examples of an asymmetrical cascade of gene expression in the developing embryo resulting in the situs of organ development. The earliest observed asymmetrical signalling pathway in development is the *nodal pathway*. Nodal is a transforming growth factor beta (TGF-β) ligand, first described in the mouse embryo (Zhou, Sasaki et al. 1993). Expression of nodal on one side of the embryo initiates further asymmetrical gene expression of downstream targets, which ultimately determine the lateral positioning of organs (a more in depth description on the regulation and downstream targets of nodal can be found in Shen 2007 and, Grande, Martin-Duran et al. 2014, however this is beyond the scope of this introduction). It has further been shown that expression of nodal on both sides of the embryo results in randomised LR asymmetry (Levin, Johnson et al. 1995, Nonaka, Tanaka et al. 1998). Sided nodal expression as an initiator of an asymmetric gene cascade has been described in a number of species (recently reviewed in Grande, Martin-Duran et al. 2014), however the cause of this initial asymmetrical gene expression remains debated.

A number of mechanisms by which an embryo can become polarised have been described, however a large amount of uncertainty remains regarding the level of conservation of mechanisms across species and how LR asymmetry is, in most cases, consistently established on one of two possible sides (Aw and Levin 2008).

Mechanisms of symmetry-breaking

Ciliary Flow

The most frequently quoted textbook mechanism for the establishment of LR asymmetry is the ciliary flow model (Tabin and Vogan 2003, Aw and Levin 2008). In this model it is proposed that the movement of inherently chiral motile cilia, create a directional fluid flow within a pocket of tissue in the fluid-filled developing embryo prior to gastrulation (Nonaka, Tanaka et al. 1998). The specific details vary according to the model, yet the most popular model assumes the directional flow results

in an asymmetric distribution of particles (Vogan and Tabin 1999), which is then detected by mechano-sensory cilia and ultimately results in asymmetric gene expression (McGrath, Somlo et al. 2003, Tabin and Vogan 2003). In the original mouse model system this directional flow originates at the node (Okada, Nonaka et al. 1999). Ciliary flow has also been exhibited originating from the similar gastrocoel roof plate in *Xenopus* and the Kupffer's vesicle in zebrafish amongst others (Essner, Vogan et al. 2002, Okada, Takeda et al. 2005).

The ciliary flow model provided a convincing explanation for the LR reversals exhibited in *iv* (inverted viscera) mutant mice (Okada, Nonaka et al. 1999) and correlates with the associations of situs inversus and ciliopathies, such as Kartagener's syndrome (Afzelius 1976, Burn 1991, Badano, Mitsuma et al. 2006). It also importantly provides a mechanism in which the LR axis could be established *de novo* due to the inherently chiral motion of the motile cilia (Vogan and Tabin 1999).

However, there are many instances in which the ciliary flow model cannot explain the causal mechanism for LR asymmetry. For example occurrences of situs inversus in the presence of functional cilia and vice versa (Burn 1991) and additionally body asymmetry is observed in organisms prior to the development of cilia or that lack cilia entirely (Manner 2001, Speder, Petzoldt et al. 2007, Okumura, Utsuno et al. 2008).

F molecule

The ability to establish LR asymmetry in organisms without cilia, indicates the existence of an intracellular mechanism to establish LR asymmetry (Levin and Palmer 2007).

Brown and Wolpert (1990) proposed the existence of an inherently chiral molecule, the F molecule, which once oriented with respect to the AP and DV axes would provide a reference point able to distinguish the left side of the embryo from the right within the cell (Figure 2a). This initial asymmetry is then 'converted' into downstream asymmetric pathways and ultimately chirality of the body plan (Brown and Wolpert 1990). This theoretical molecule does not require any 'decision making' to determine left from right, it is intrinsically distinguishable via its structure and provides a molecular method of concordance.

Indeed, the motile cilia responsible for the ciliary flow model represent a potential F-molecule, in that the direction of movement is inherently chiral due to the chiral structure of the cilia (Figure 2b) and not susceptible to change though a reference point, subsequently the ciliary flow model can be considered to encompass the F molecule theory (Levin and Palmer 2007).

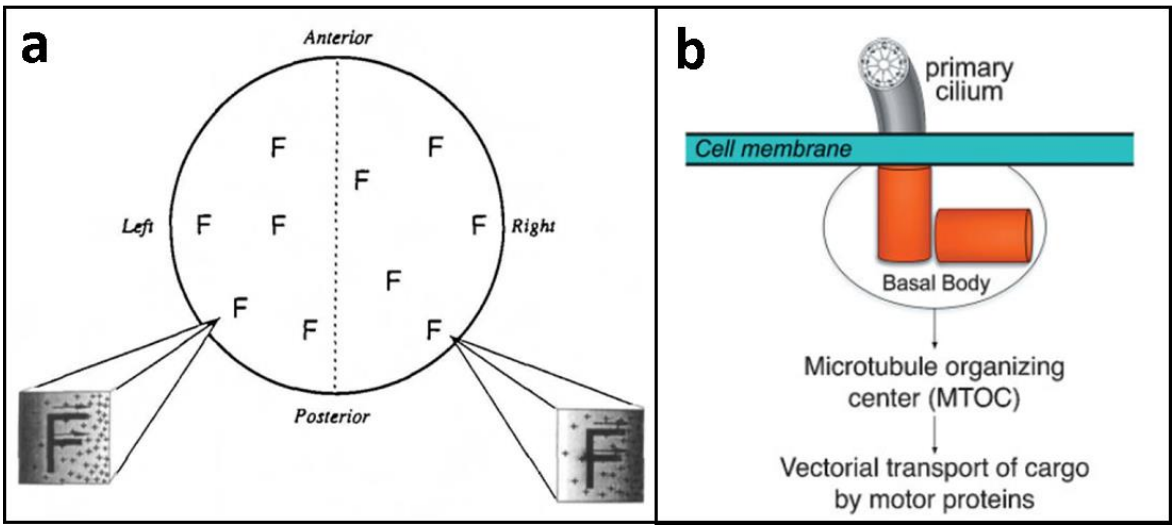


Figure 2 a: Theoretical representation of how a chiral 'F molecule' would enable detection of gradients and subsequently distinguish between left and right within a single cell. Image reproduced from (Brown and Wolpert 1990). b: Composition of cilia, revealing chiral basal body and cytoskeletal functions. Image reproduced from (Levin and Palmer 2007).

Cytoskeletal processes

Cytoskeletal components have been associated with the mechanism of establishing LR asymmetry presented in a number of models. Due to the chiral nature of the cytoskeleton, it provides the possibility for very early determination of LR sides, perhaps as early as the first cell cleavage (Vandenberg, Lemire et al. 2013).

The chromatid segregation model postulates that the chiral cytoskeleton generates an asymmetric distribution of the chromatids during the first cell cleavage, resulting in differentially imprinted chromatids, which are therefore able to differentiate between the two cells/ left and right halves of the embryo (Klar 1994, Klar 2008).

The microtubule organising centre (MTOC), includes the centrosome, and comprises the site of microtubule nucleation and plays a key role in cell division through mitotic spindle organisation (Karsenti and Vernos 2001, Bornens 2012). However, there are some notable exceptions and some organisms lack centrosomes completely (Calarcogillam, Siebert et al. 1983, Mahoney, Goshima et al. 2006, Azimzadeh, Wong et al. 2012). The MTOC is considered to hold functional and structural asymmetries considered essential to maintain cell polarity and asymmetry (Bornens 2012). The other major function of the MTOC is it organises and forms the components of cilia (Figure 2b) (Levin and Palmer 2007).

Cytoskeletal components actin and tubulin have been implicated in the early establishment of LR asymmetry in a number of organisms (Baum 2006, Lobikin, Wang et al. 2012). Although the precise

mechanisms are unclear, the inactivation of actin polymerisation resulted in a loss of asymmetry in early developmental stages of the pond snail *Lymnaea stagnalis* (Shibazaki, Shimizu et al. 2004). Actin molecules have also been shown to undergo spontaneous symmetry breaking in-vitro and self-organisation, facilitating a mechanism of symmetry-breaking within a single cell (Abu Shah and Keren 2014, Mogilner and Fogelson 2015).

Cytoskeletal dynamics have also been incorporated into mechanisms of symmetry breaking based on gap junctions and ion flux gradients (Levin 2003, Oviedo and Levin 2007), which are described in more detail next. It has been proposed that the cytoskeleton actively directs the asymmetric distribution of proteins, including K+ channels and H+ pumps, via motor proteins (Levin and Palmer 2007). It has also been shown that actin inhibition results in a failure of the mechanisms described in ion flux models (Adams, Robinson et al. 2006, Ayerscough 1998, De Brabander, Geuens, et al. 1986).

Gap junctions and Ion flux models

Potassium (K+) channels and Hydrogen (H+) pumps produce consistent biases in the transmembrane voltage and pH, which are able to drive the asymmetric distribution of small molecules through gap junctions from one side to the other. The hydrogen potassium (H+/K+) ATPase transporter has been identified as obligatory for correct LR patterning in early chick and *Xenopus* embryos. Perturbation of the endogenous H+/K+ -ATPase resulted in randomisation of LR asymmetry in both species (Levin, Thorlin et al. 2002). Furthermore, this channel results in the asymmetric localisation of maternal ATPase within 2 hours of fertilisation in the *Xenopus* embryo, indicating that LR asymmetry determination via this mechanism occurs very early in development (Levin, Thorlin et al. 2002). It has since been shown that ion flux is involved in LR asymmetry determination in a number of vertebrates and non-vertebrates (summarised in Adams, Robinson et al. 2006).

Ion transporter proteins and gap junctions also represent a promising model as they allow for subcellular asymmetries to spread, providing a method for how the symmetry breaking event is then amplified across the organism (Levin and Palmer 2007). Finally, an increasing number of studies have documented the function of maternal serotonin in the establishment of LR asymmetry in *Xenopus* and chick embryos (Vandenberg, Lemire et al. 2013). Maternal serotonin, asymmetrically localised by gap junctions, has been implicated in the epigenetic repression of asymmetric gene expression of nodal and consequent randomised LR asymmetry (Carneiro, Donnet et al. 2011).

Conservation between organisms

Nodal signalling

Nodal signalling is involved in a number of different functions and activates various downstream targets in different organisms, however the asymmetric expression of Nodal and it's downstream target 'Pitx' have been shown to be conserved across vertebrates in the establishment of LR asymmetry and regulating gastrulation in embryogenesis (reviewed in Tian and Meng 2006). The apparent absence of Nodal expression in ecdysozoans and platyhelminthes led to the previous assumption that this pathway was specific to the vertebrate development. More recent observations have identified orthologs of Nodal in a number of non-vertebrate deuterostomes (Morokuma, Ueno et al. 2002, Yu, Holland et al. 2002, Duboc, Rottinger et al. 2005) and non-deuterostome groups including Mollusca (Grande and Patel 2009), brachiopoda, chaetognatha (Grande, Martin-Duran et al. 2014) and Cnidaria (Watanabe, Schmidt et al. 2014). It therefore seems likely that Nodal signalling appeared very early in the evolution of the Bilataria.

Although the presence of Nodal is conserved across much of the bilateria, variations in downstream targets, expression domains and characterised functions obscure inferences regarding the ancestral role of the Nodal pathway in the establishment of LR asymmetry (reviewed in Tian and Meng 2006, Grande, Martin-Duran et al. 2014). For example, expression of Nodal in deuterostomes first appears symmetrically expressed on both sides of the embryo and then is restricted to the left side (Levin 1998, Nonaka, Tanaka et al. 1998, Morokuma, Ueno et al. 2002, Yu, Holland et al. 2002) with one notable reversal observed in the sea urchin (Duboc, Rottinger et al. 2005). In the non- deuterostome snail and brachiopod, however, Nodal expression is observed initially asymmetrical on the right side of the embryo (Grande and Patel 2009, Grande, Martin-Duran et al. 2014). This reveals two key differences in the Nodal pathway. Firstly, in the deuterostomes, nodal is restricted from initially symmetrical expression to sided expression, indicating a regulating factor to localise the symmetric nodal expression, whereas in the non-deuterostomes, an upstream factor is directing initially asymmetric Nodal expression (Grande and Patel 2009). Secondly the side of Nodal expression is reversed. The right-sided nodal expression in the snail and brachiopod, in addition to proposals that the ancestral state of snail chirality is dextral (Ponder and Lindberg 1997), suggest that the ancestral pattern of Nodal expression is on the right side. However, it is important to assess more non-deuterostome groups to establish ancestral relationships.

Another major difference observed in the Nodal signalling pathway in is timing. For example, in vertebrate embryos, nodal expression occurs prior to gastrulation (Grande, Martin-Duran et al.

2014), whereas nodal signalling in snails is only observed at a much later developmental stage, notably after LR asymmetry has been established (Grande and Patel 2009). Therefore, although nodal signalling initially arises asymmetrically, it does not represent a symmetry breaking event in the snail.

Timing

The timing of observed symmetry breaking can provide support or outright disprove proposed mechanisms of establishment of LR asymmetry in an organism. As described earlier asymmetries have been observed prior to the formation of cilia and therefore the cilia model cannot explain symmetry breaking in these organisms (Vandenberg and Levin 2010). Furthermore, phenomena such as gynandromorphy, in which an organism is both male and female due to failure for chromosomes to separate properly in first cell division (Barranco, Cabrero et al. 1995), and the dermatological 'CHILD' syndrome, which presents LR bilateral segregation of pigmentation in humans (Happle, Mittag et al. 1995) both reveal striking asymmetrical external morphology apparent across the midline. The chromosomal segregation associations with these disorders indicate the formation of the midline may occur as early as the first cell division (a selection of examples and images are presented in Aw and Levin 2008).

It is important to acknowledge that the earliest observed asymmetry in development is not directly indicative of when asymmetry is established. The initiation of asymmetry is indicated to be as early as the 32 cell stage in *Xenopus* (Vandenberg and Levin 2010) and mouse embryos have been observed to not be LR equivalent by the 8-cell stage (Gardner 2010), far earlier than the appearance of the furrow or node equivalent. Furthermore, certain embryo injection experiments in *Xenopus*, have been shown to only be effective if administered prior to first cell cleavage (Lobikin, Wang et al. 2012). Consequently, timing of LR asymmetry manipulation experiments must be considered very carefully for appropriate interpretation of the outcome. A further important consideration is that of maternal RNAs which will not be effected by experimental inhibitors of gene expression, as they have already been transcribed.

The intracellular models allow for a very early establishment of asymmetry. The plausibility of intracellular symmetry breaking is supported by the establishment of consistent LR asymmetries exhibited at the cellular level (Heacock and Agranoff 1977, Hagmann 1993, Xu, Van Keymeulen et al. 2007, Wan, Ronaldson et al. 2011, Chen, Hsu et al. 2012). The common feature of these systems is the presence of a cytoskeleton. It has also been demonstrated that cytoskeletal components self-organise into chiral structures (Mogilner and Fogelson 2015). What remains to be recognized is the

mechanism by which these intracellular chiral components are translated into larger scale asymmetries.

A unifying model?

It is important to note that the models described here are not mutually exclusive and evidence which supports one model does not necessarily refute another. For example, a number of studies have contested the role of serotonin, ion flux and gap junctions in early establishment of LR asymmetry due to their roles identified within ciliary processes later in development (Beyer, Danilchik et al. 2012, Walentek, Beyer et al. 2012). However, processes that are involved in later LR patterning does not exclude their possibility of performing roles in the early establishment in LR patterning, and it is still possible that the ciliary mechanisms are a by-product or downstream amplifier of an earlier establishment of LR asymmetry by these processes (reviewed in Vandenberg and Levin 2013)

It is also important to make the appropriate conclusions from LR manipulation experiments. Many of the studies in support of the cilia model, refer to the involvement of 'ciliary' proteins of which also perform non-ciliary functions, and many do not specify the localisation of the effect i.e. at the node (Vandenberg and Levin 2013). In a quantitative analysis of the literature approximately half of all studies implicating cilia in the establishment of LR asymmetry provide no measure of morphology or function of the effected cilia and therefore offer unclear definitions of normal/abnormal cilia (Vandenberg 2012). It is likely therefore that many of these studies are affecting cytoskeletal components outside the roles of cilia motility.

Cytoskeletal dynamics perform fundamental roles within all of the intracellular models described, and the MTOC provides the chiral component within the cilia model. Therefore, the cytoskeleton represents a common process in how all organisms establish LR asymmetry. Thus it has been proposed that the establishment of asymmetry is deeply conserved and the cytoskeleton provides the ancestral origin of asymmetry, with the MTOC playing the role of the F molecule proposed by Brown & Wolpert (1990) (Vandenberg and Levin 2013)

It has alternatively been proposed that the developing embryo makes a 'choice' to stochastically utilise one of several available pathways of establishing LR asymmetry (Vandenberg and Levin 2013). Resulting LR interference studies would thus only affect those individuals that had undertaken the pathway being manipulated. This model provides an explanation for the low penetrance of disruption of LR patterning and lack of 100% reversals observed in experimental manipulations.

Each model is supported to a varying extent in different species. Vandenberg *et al.* (2013) provide an insightful quantitative summary of literature cited implicating mechanisms involved in the establishment of LR asymmetry in different species. For example, the cilia model is highly supported in mouse studies yet not so strongly in other vertebrates and not at all in invertebrates. Similarly, the full details of gap junction models have only been characterised in *Xenopus*, although processes have been implicated in a number of other species. Finally, cytoskeletal dynamics are by far the most cited in non-vertebrate model systems and single cells.

If there are, as it appears, different mechanisms across species it is important to identify the best model system for insight into the human condition. The mouse model of nodal flow has shown a number of differences compared to other amniotes, for example the chick embryo has a node but does not have motile cilia, similarly to the pig (Gros, Feistel et al. 2009). Additionally, in syndromes such as CHILD mentioned previously, the mouse does not show pigmentation divided across the midline, yet demonstrates a mosaic pattern (Konig, Happle et al. 2000). Thus the mouse may not represent the best model for medical inferences. It is crucial therefore, to look at multiple model systems to gain insight into which mechanisms are conserved, and thus likely to represent the ancestral state, and establish the level of integration of these processes within derived mechanisms.

LR asymmetry in snails

Spiralian development

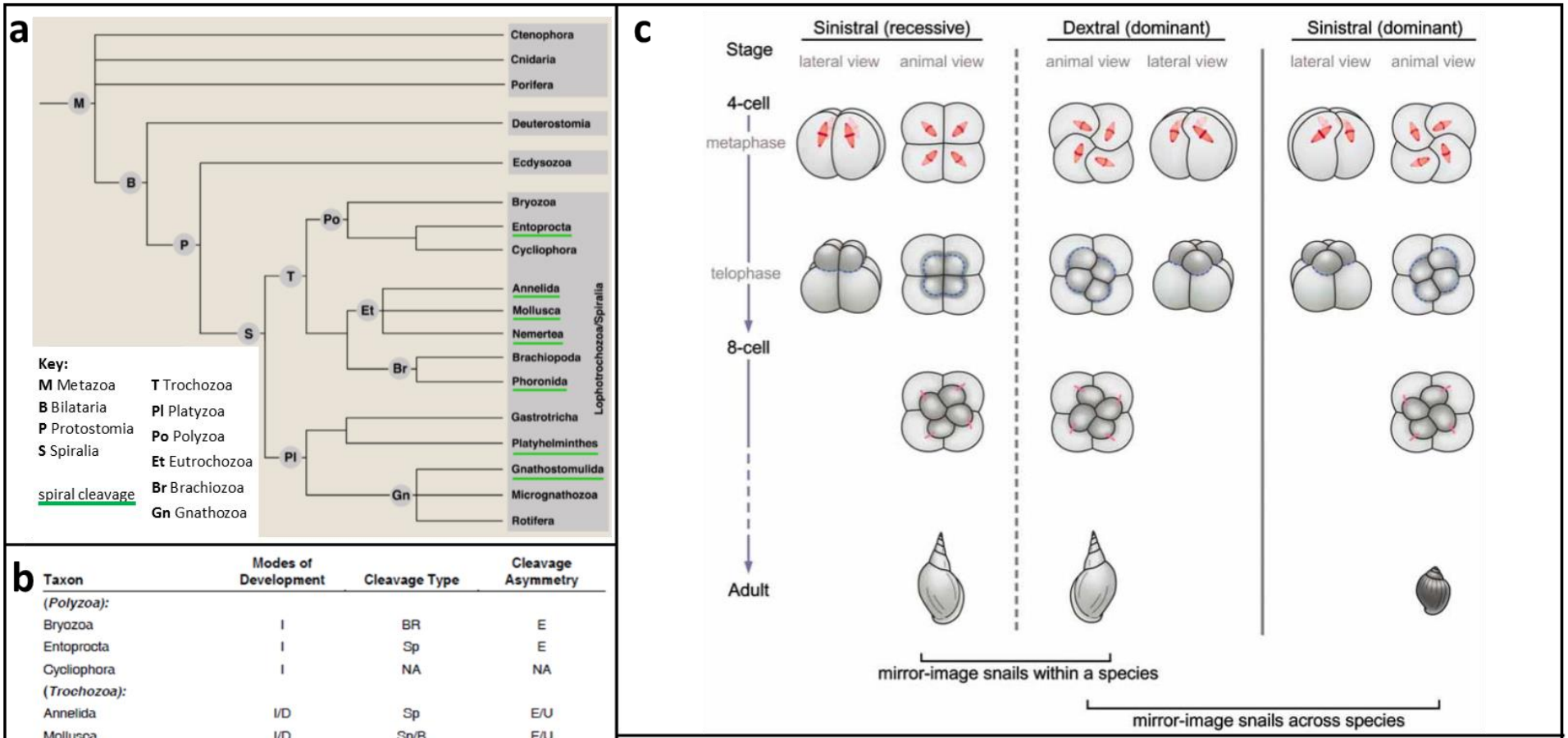
The four most frequently used model organisms in the study of LR asymmetry in development, chick, mouse, *Xenopus*, and zebrafish, are all vertebrates. Invertebrates however can provide useful insight into ancestral traits and a potential universal system of establishing asymmetry. A number of invertebrates are known to exhibit LR asymmetries to varying degrees including, sea urchins, molluscs, *Amphioxus* (an ancestor of vertebrates), *Drosophila* and nematodes (reviewed in Levin 2005, Okumura, Utsuno et al. 2008).

The Spiralia (often used synonymously with Lophotrocozoa) represent one of three major clades within the bilataria and comprise nearly half of the extant metazoans, yet receive far less attention when compared to the other two clades, deuterostomes and ecdysozoans (Henry, 2014, Figure 3a). Within the spiralia there are a number of debated clades, and due to the variability of features in development and the adult body plan, an agreement of the phylogeny continues to be a challenge. A more detailed discussion of the current phylogeny of spiralia is well presented in Henry, 2014.

However, many spiralians are especially useful when studying LR asymmetry due to the method of spiral cleavage in early development.

The spiral cleavage pattern reveals chirality from the third cell cleavage in which the four micromeres do not emerge directly above the four macromeres but are rotationally displaced. This twist occurs either clockwise (dextral) or anticlockwise (sinistral) with respect to the macromeres (Figure 3c) and alternates direction through subsequent cleavage cycles (van den Biggelaar 1991). This is the earliest observed whole-body LR asymmetry in development (Brown and Wolpert 1990). This method of spiral cleavage however, is not universal amongst the spiralia, or even within phyla. Figures 3a and 3b highlight representative spiralian groups and their method of cell cleavage.

All spiral cleavers exhibit this early developmental asymmetry, although many larvae resulting from spiral cleavage are bilaterally symmetrical. Snails, within the phylum Mollusca, provide a valuable study organism because the direction of initial spiral cleavage is continued into their visceral asymmetry and, in most cases, is conserved in the direction of external shell coiling (Crampton 1894, Robertson 1993). As a result, reversals in chirality are easily observed and monitored in snail populations



b

Taxon	Modes of Development	Cleavage Type	Cleavage Asymmetry
(Polyzoa):			
Bryozoa	I	BR	E
Entoprocta	I	Sp	E
Cycliophora	I	NA	NA
(Trochozoa):			
Annelida	I/D	Sp	E/U
Mollusca	I/D	Sp/B	E/U
Nemertea	I/D	Sp	E
Brachiopoda	I/D	R	E
Phoronida	I	BR	E
(Platyzoa):			
Gastrotricha	D	MR	E
Platyhelmintes	I/D	Sp/Id	E
Gnathostomulida	D	Sp	E
Micrognathozoa	NA	NA	NA
Rotifera	D	MR	U
(Other Enigmatic groups):			
Acanthocephala (Rotifera?)	I	MR	U
Myzostomida (Annelida?)	I	Sp	U
Mesozoa (Rhombozoa)	I	Sp	E
Mesozoa (Orthonectida)	I	NA	NA

a: Phylogenetic tree highlighting evolutionary relationships of the spiralia. Phyla which contain species exhibiting spiral cleavage are underlined in green. (Adapted from Henry, 2014).

b: Table highlighting the developmental mechanisms within spiralian taxa including; 'Modes of development', either containing larval stages (indirect: I) or not (direct: D); 'Cleavage Type', whether undergoing spiral (Sp), radial (R), modified radial (MR), bilateral (B), biradial (BR) or idiosyncratic (Id) cleavage patterns, and finally, 'Cleavage Asymmetry' describes whether the first few cleavage divisions are symmetrical (equal: E) or asymmetrical (unequal: U). NA = data not available. (Adapted from Henry, 2014).

c: The third spiral cleavage and consequent adult coiling direction as exhibited in the chiral morphs of the pond snail *Lymnaea stagnalis*, showing the sinistral recessive (left) and dextral dominant (middle). Also shown is the spiral cleavage of another aquatic snail *Physa acuta*, showing the invariable sinistral form (right). Metaphase spindles are shown in red. (Image reproduced from Shibazaki, Shimizu et al. 2004).

Figure 3 Spiralian phylogenetic relationships and early developmental mechanisms. Full details within image.

True enantiomorphs

Snails are rare in that they present true enantiomorphs with multiple chiral reversals having occurred within a number of phylogenetically independent families (Schilthuizen and Davison 2005). Estimates of the frequency of sinistral morphs in snails vary, and regional variation can be quite substantial (Okumura, Utsuno et al. 2008), but it is generally assumed to be less than 10% of species (Asami 1993, VanBatenburg and Gittenberger 1996, Schilthuizen and Davison 2005). The reason for the flexibility of chiral reversals in gastropods is unknown. However, a contributing factor to the number of documented cases may be the ease of observation from shell-coiling direction that far more sinistral species have been reported in snails than other organisms, including slugs, which are derived from snails (Reise, Benke et al. 2002). Conversely it may be that the conspicuous nature of the external shell drives selection not apparent in the slug. It is also possible that there has been an overestimate of reversals in snails, as a result of primary asymmetry wrongly being assumed from the observed secondary asymmetry of shell-coiling (Robertson 1993, McMillen and Goriely 2002), yet the number of chiral reversals reliably documented remains dramatically higher than in other species classes.

Perhaps surprisingly given the frequency of chiral reversals within gastropoda, most species exhibit only one chiral morph. There are many interesting facets of the ecological implications of chiral reversal including barriers to inter-chiral mating due to the incompatible lateral positioning of sexual organs or behavioural biases resulting in the reduced probability of reproduction in the rarer morph (Johnson 1982, Schilthuizen, Scott et al. 2005, Schilthuizen, Craze et al. 2007, Davison, Frend et al. 2009, Koene and Cosijn 2012). This led to the proposal of a possible single-gene speciation event occurring from the emergence of immediately reproductively isolated reversed chiral morphs (Gittenberger 1988, Ueshima and Asami 2003). However, this theory remains contentious. Because of the maternal inheritance of chirality exhibited in the experimentally observed cases of chirality inheritance (described in further detail in section '*Lymnaea* as a model system'), it is believed that gene flow would prevail between opposite chiral morphs (Davison, Chiba et al. 2005). Alternative frequency dependent selection mechanisms have been proposed for dimorphic populations in the presence of chiral predators. For example, sinistral morphs have been shown to survive predation by the snail-eating snake, *Pareas iwasakii*, which bears asymmetric mouth parts apparently adapted to eating dextral coiling snail (Hoso, Kameda et al. 2010).

In light of the potential barriers to mating between chiral morphs, it is notable that there are still limited examples of chiral dimorphism within externally fertilising species which do not suffer the same behavioural or physical barriers, and significantly fewer recorded examples of sinistral snails in

the sea (Hendricks 2009). Therefore, although there is undoubtedly selection on chiral dimorphism, it is unlikely to be the only factor limiting the prevalence of sinistral species, as such there may instead be a selective constraint on the propagation of sinistrality (Vermeij 1975) although there is limited evidence to support this (Davison, Barton et al. 2009).

Species which maintain chiral dimorphism are therefore essential to the study of the inheritance of chirality. Pulmonate snail species from phylogenetically independent families (Wade, Mordan et al. 2006) in which interchiral mating is possible, have shown inheritance of directional LR asymmetry through a maternal effect gene. Interestingly the genetic dominance of the sinistral or dextral form varies between species. *Lymnaea spp.* show dextral coiling is dominant to sinistral (Sturtevant 1923, Boycott, Diver et al. 1930) whereas *Partula spp.* show sinistral coiling to be dominant to dextral (Murray and Clarke 1966).

Expression of nodal has also been documented in a number of snail species (Grande and Patel 2009, Kuroda, Endo et al. 2009). Originally identified in the dextral, *Lottia gigantea* and the sinistral *Biomphalaria glabrata* each exhibited opposite patterns of lateral nodal expression (Grande and Patel 2009). Dextral snails expressed nodal and the downstream target Pitx on the right side whereas the sinistral snails expressed them on the left. This is of great significance due to the apparent absence of nodal in *Drosophila* and nematodes. It has been suggested that their absence of nodal expression may reflect the more derived modes of mesendoderm and LR specification in these systems (Schier 2009).

Lymnaea as a model system

It was observed that primary LR asymmetry in snails (indicated by the direction of shell/whole-body coiling) is determined by a single heritable unit over a century ago (Crampton 1894). Boycott and Diver (1923) documented the inheritance of chirality in the pond snail, *Lymnaea peregra* via multiple laboratory mating crosses between snails of opposite shell-coiling. They described in detail, an unknown pattern of inheritence with five possible outcomes for the resulting ratio of chiral phenotypes within the offspring/broods, including the phenomenon of mixed broods, in which a single clutch of eggs contains offspring of both chiral morphs.

It was shortly thereafter proposed by Sturtevant (1923) to follow a much simpler Mendelian mode of inheritance with a maternal effect, in which the classic Mendelian dominant/recessive mode of phenotype expression is delayed a generation. Therefore, the genotype of mother is expressed through the phenotype of their offspring. This is the presumed mechanism of inheritance of chirality in all chirally dimorphic pulmonate snail species (Asami, Gittenberger et al. 2008). In *Lymnaea spp.*

the dextral form is dominant to the sinistral form. Therefore, the offspring from a heterozygote mother will bear a dextral-coiling body plan. It is important to note that the occurence of mixed broods in Lymnaea spp. is more often documented with dextral coiling offspring occuring in what should be a sinistral brood. Sinistral offspring in a dextral brood was only documented once in the original Boycott and Diver observations (Boycott and Diver 1923).

The sinistral population of *L. peregra* has since been lost, and now the sister species: *L. stagnalis* is used (Morrill 1982, Meshcheryakov 1990). It is believed that the chiral determinant in *L. peregra* and *L. stagnalis* functions in a common manner, albeit with a few minor differences (Kuroda 2014). *L. stagnalis* provides a valuable study system for LR asymmetry for a number of reasons. In addition to the maintained chiral dimorphism, it produces relatively large transparent eggs, approximately 1 mm in diameter, with the yellow embryo visible with the naked eye. It is a hermaphrodite and can reproduce either through self-fertilisation or with a sexual partner (preference), facilitating rapid and effective inbreeding and backcrossing to reduce genetic variability (Hosoiri, Harada et al. 2003).

The direction of spiral rotation of micromeres at the third cell cleavage correlates almost perfectly with eventual organ situs and shell coiling (one exception has been observed (Kuroda 2014)) and as such provides an informative signal of chirality early in development. However, it has recently become apparent that the chiral morphs of *L. stagnalis* are not true enantiomorphs in these early cleavage steps. In the dextral embryo, the third cell cleavage results in the four micromeres emerging at a 45° angle on top on the sister macromeres, whereas in the sinistral embryo, this rotation does not occur immediately and the emergent micromeres sit directly on top of the macromeres with no rotation (Figure 3c) (Shibazaki, Shimizu et al. 2004).

The cytoskeletal components during this phase have been observed, revealing two key steps involving mitotic spindles and microfilaments. In the dextral embryo, firstly 'spindle inclination', reveals a helical orientation of the mitotic spindles with respect to the animal-vegetal axis, which corresponds to the following dextral cleavage of the 4-8 cell embryo. Secondly, filamentous actin, which appears concentrated at each cell boundary and cleavage furrow, precedes a 'spiral deformation' of the blastomeres, which results in the helical emerging micromeres (Figure 3c). The sinistral *L. stagnalis* embryos however, do not undergo either spindle inclination or spiral deformation, yet exhibit radial symmetry throughout the third cell cleavage, and only exhibit sinistral rotation after the emergence of the micromeres (Figure 3c) (Shibazaki, Shimizu et al. 2004).

Although there is no apparent reduction in fitness of the adult sinistral *L. stagnalis*, they are vulnerable in development and suffer a reduced hatch rate (Davison, Barton et al. 2009, Utsuno,

Asami et al. 2011). Developing sinistral *Lymnaea* are also susceptible to reversion to dextrality through unspecified physical manipulation. This was previously believed to be due to the transplanting of the cytoplasmic fluid which surrounds the embryo (ooplasm) from the dominant dextral form (Freeman and Lundelius, 1982). However, this is not the case in *L. stagnalis* and a dextral reversion can occur just through physical disruption (Kuroda 2014).

Due to the observed pathologies (of unknown mechanism) and differences in cleavage pattern in the sinistral developing embryo (Shibazaki, Shimizu et al. 2004), *L. stagnalis* does not provide a perfect comparison of chiral reversal. Still in light of the ease of laboratory rearing and interchiral breeding, *L. stagnalis* provides an invaluable model to study LR asymmetry in development and holds a strong background of previous studies.

Although the gene causing reversal of chirality has not yet been isolated, the mechanism of establishing asymmetry has been highlighted to involve the cytoskeleton. Inhibition of actin polymerisation in genetically dextral embryos resulted in a failure of the emerging micromeres to rotate (spindle deformation and subsequent spindle inclination), mimicking the sinistral wild-type form (described above). Treatment with the same agents on the sinistral embryos had no discernible effect (Shibazaki, Shimizu et al. 2004). Conversely, inhibition of microtubule polymerisation actually enhanced the spindle deformation in the dextral embryos and did not inhibit spindle inclination. Again no change was seen in the sinistral embryos (Shibazaki, Shimizu et al. 2004). This study indicates that the actin cytoskeleton is essential for correct dextral spiral cleavage, whereas microtubule actions occur as a result of the already formed blastomeres. Additionally, it implies that the sinistral form is lacking a functional step in spiral cleavage (spindle inclination).

Further supporting the role of the cytoskeletal structural components in determining chirality was a key study, which physically reversed the orientation of dextral and sinistral *L. stagnalis* embryos by micromanipulation, resulting in 78% successfully LR reversed organisms (Kuroda, Endo et al. 2009). The reversed organisms that reach adulthood produced offspring in accordance with their original coiling direction, indicating that there had been no genetic change in the individuals. This experiment also revealed factors regarding the timing of asymmetry determination. If manipulated prior to the third cell cleavage the embryos would correct themselves to their genetically disposed orientation. Therefore, the cell contacts between macro- and micromeres appear to be the determining step in LR asymmetric patterning (Kuroda, Endo et al. 2009).

Asymmetrical gene expression of Nodal and the downstream Pitx, have been observed in *L. stagnalis*. First detected at cell stages 33-49 and therefore present much earlier in development than

observed in the vertebrates. The physically reversed embryos described above, exhibited reversed asymmetric expression of nodal (Kuroda, Endo et al. 2009). This indicates that the reversal had successfully redirected cell-cell communication and ultimately that the nodal pathway occurs downstream of the symmetry determining step in *L. stagnalis*, which is governed by cytoskeletal dynamics, which are very strongly associated with inheritance of the maternal effect gene.

Whereas the mechanism of establishing LR asymmetry in *L. stagnalis* remains unknown, the method of its inheritance has been known for nearly a century (Sturtevant 1923). Therefore, the Davison research group has focussed on genetic mapping of the 'chirality locus' to identify the causal gene 'D'. There is not yet a fully annotated genome available for *L. stagnalis*, however linkage mapping, via high-throughput sequencing of genetic crosses between lab-reared monomorphic chiral populations, identified three anonymous markers which are tightly linked to the chiral phenotype (Liu, Davey et al. 2013).

This has enabled genotyping of a large number of individual snails according to the chirality locus, as homozygous dominant '*DD*', heterozygous '*Dd*' or homozygous recessive '*dd*'. Throughout this thesis individual samples and populations will be referred to by this nomenclature of chiral genotype. It is important to note that snails are scored by their chiral genotype, which corresponds to the coiling direction of their offspring, not necessarily their own shell coiling direction.

The chirality locus has been localised further through pachytene FISH (**f**luorescent ***in situ*** **h**ybridisation) and fibre-FISH mapping, which enable visualisation of specific genetic transcripts hybridised to whole chromosomes (for a more in-depth description of these techniques please see; Weier, Wang et al. 1995, Garimberti and Tosi 2010). The chromosomal FISH mapping revealed that the three chirality-linked markers all occur on the same chromosome, providing strong evidence that the chirality locus lies within this region (Liu, Davey et al. 2013). Further sequencing via a method called BAC (bacterial artificial clone) walking, which enables genetic sequencing of unknown regions when initiated from an area of known sequence, have provided additional genetic sequence information within this region (for more information regarding this technique please see Kubat 2007). Combined with sequence comparisons of other mollusc species, the distance between the two most tightly linked markers, assumed to contain the chirality gene, has been estimated between 0.4 and 0.6 megabases (Liu, Davey et al. 2013).

Gene expression analysis

RNA: a versatile molecular tool

Sequence data from genomic DNA provides information regarding both protein coding genes and non-coding regions and accordingly holds a significant amount of genetic information not transcribed into RNA. However, RNA expression data offers a wealth of functional information that DNA cannot.

One of the main functional classes of RNAs used in gene function studies are messenger RNAs (mRNAs). The code for each gene is stored in DNA present in almost every cell in an organism. This sequence is transcribed into mRNA when activated. The mRNA sequence then specifies the amino acid sequence to generate the particular protein (Meneely 2009). Thus mRNA represents an intermediate molecule between DNA and the eventual protein and can provide information on which genes are being transcribed, or 'switched on', in the cell.

One of the defining properties of mRNA when it was first discovered in the mid-20th century was the transient nature of the molecule. In yeast the half-life of mRNA varies from 1-100 minutes, whereas in mammals it varies from less than 20 minutes to up to 50 hours. mRNA decays at varying rates, partly this can be attributed to tertiary structure of the mRNA, however there are a number of documented mechanisms affecting the stability or active degradation of mRNAs, revealing another method in which gene expression is regulated within the cell (Elliot and Ladomery 2011, pp. 307-321).

RNA analysis can provide much more information than simply which genes are being transcribed and quantitatively regulated. This was made apparent through the genome sequencing project, which identified substantially fewer protein coding genes than the number of actual proteins known to be present in humans (the proteome). This discrepancy has been explained through post-transcriptional RNA processing. Transcriptional regulation determines which genes are switched on or off in a cell, whereas the complexity of the proteome is in a large part due to post-transcriptional RNA processing, such as alternative splicing (Elliot and Ladomery 2011, pp. 158-192). Alternative splicing produces different mRNA sequences from the same gene, as such these isoforms would not be observable from DNA sequencing.

It has now become apparent that the functional aspects of RNA are not limited to the protein-coding mRNAs. Non-coding RNAs (ncRNAs) can be divided into long and short. Long ncRNAs are considered to be >200 nucleotides (nt) in length, arbitrarily due to RNA extraction protocols which omit short RNAs (Mercer, Dinger et al. 2009). The field of long ncRNA research is relatively recent. Although

many predicted functions are uncharacterised, long ncRNAs are predicted to have widespread functionality including chromatin modification and modulating protein binding interactions, pre and post-transcriptional regulation (Mercer, Dinger et al. 2009, Wilusz, Sunwoo et al. 2009). Small ncRNAs have been better characterised and are largely involved in the regulation of gene expression and particularly gene silencing, although many classes likely remain to be identified (Mattick and Makunin 2006). There is a growing body of evidence highlighting that micro RNAs (19-25 nt in length) and recently classified Piwi-interacting RNAs (24-30 nt in length) play key roles in the regulation of animal development (Stefani and Slack 2008).

A number of other features of RNA provide insight into the workings of the cell and gene expression, such as catalytic enzymatic functions, recognition site and protein interactions, complex secondary and tertiary structures resulting in varied reactivity and structural components such as the ribosome (Elliot and Ladomery 2011). Although these facets are beyond the scope of this project, it is important to recognise the vast potential of RNA analysis.

Patterns of gene expression

To infer biological meaning from RNA data, patterns of gene expression must be identified. A common assessment of gene expression simply compares the quantitative level of gene expression between two samples. This directly implicates genes, or specific isoforms, which are being affected by the variable considered. Due to the transient nature of RNA, quantitative comparisons can be performed between samples or within samples over time to gain both temporal and spatial comparisons of gene expression. Quantitative patterns of gene expression can reveal differences between organisms indistinguishable through genomic approaches (as demonstrated by Wolf, Bayer et al. 2010).

The locality of gene expression can reveal insightful functional clues regarding the role of the RNA transcript. Therefore, it is essential to compare specific tissues within quantitative experiments. Additionally, techniques such as *in situ* hybridisation, allow for the visualisation of gene expression within tissues or in the case of whole-mount *in situ* hybridisation even whole organisms (due to size limitations this is usually limited to developmental stages) (Hemmati-brivanlou, Frank et al. 1990).

It is important to remember than specific gene expression is almost always in combination with that of other transcripts. Observing expression patterns of multiple target transcripts can reveal gene networks and clusters. Similarities in gene expression patterns can indicate similar functions of the genes involved and expression pathways, thus enabling functional interpretations of undescribed

genes (Chua, Robinson et al. 2004, Janky, van Helden et al. 2009). This however, requires a certain level of functional annotation available, which is not the case for the majority of species.

Proving causality

It is essential when exploring genetic relationships linked with an observed phenotype to clarify whether the gene causes the phenotype or is simply associated with phenotype. This is especially apparent with the advent of next generation sequencing and the growth of large sequence datasets, where the need to justify cause as opposed to correlation is of increasing concern.

This is not a new problem and represents a natural caveat of any experiment. How does one prove the order of cause and effect? Further still, proving that a factor precedes an effect does not denote that it caused it. Entire books have been written on the mathematical and philosophical properties of causality (Pearl 2000) however this thesis is not one of them. This problem can be minimised by performing highly controlled randomised experiments. In genetics however, experimental conditions (genotypes) cannot be randomly allocated and therefore studies are outside the statistical requisites of true experimentation and leave room for doubt (Rubin 1991, Rosenbaum 1995). Therefore, the role of the experimenter is to minimise residual doubt.

This process has been coherently reviewed in (Page, George et al. 2003), in which they summarise possible origins of an association between a gene/polymorphism and a phenotype. Firstly, it may be that the gene does in fact cause the observed phenotype, a true causal relationship. Alternatively, statistically significant associations may represent false positives identified due to chance. The gene may be associated with the trait due to disequilibrium together with the true causative gene. Finally, the association may have been identified through systematic bias within the experiment. In order to support the first hypothesis that the gene is a true causative polymorphism, the other specious possibilities must be eliminated.

Reducing the occurrence of false positives by chance can be achieved by increasing the probability (p) value threshold for statistical significance. Such is the case for p value corrections to account for false discovery rate (FDR) due to multiple comparisons (Benjamini and Hochberg 1995, Benjamini, Drai et al. 2001). However, simply increasing the p value will not correct for the alternative sources of erroneous associations.

Identifying a polymorphism in linkage disequilibrium provides a primary indicator of its association with the observed trait. However, in order to ascertain whether it represents the causative polymorphism, efforts must be taken to classify (and subsequently eliminate) all other

polymorphisms within the region of DNA, which may also be in disequilibrium (Demuth and Wade 2006).

Systematic bias can arise in two forms, experimental bias or biological bias. All experimental studies include a level of error. Replication, although an assumed part of any scientific experiment should be considered for both technical or 'operational replication', in which the same methodology is employed in order to achieve the same result, and 'constructive replication' in which the same result is achieved via a different methodology (outlined and exemplified in Lykken 1968). Each of these forms of replication highlights different sources of experimental error. It is subsequently essential to appropriately calculate and incorporate error into the resulting analyses and inferences. For example, if the data does not fit the statistical model employed, the resulting p value will be irrelevant, no matter how 'significant' the value.

Knowing the biological system being examined is critical in order to design an experiment free of systematic bias. For example, the study populations, if not able to be controlled must have their evolutionary or population history known to enable incorporation of error due to admixture or other confounding genetic effects, such as environmental effects on gene function (Page, George et al. 2003).

Although many have proffered guidelines for the minimum requirements to infer causation (Koch 1882 cited in: Page et al. 2003, Hill 1965, Glazier, Nadeau et al. 2002, MacArthur, Manolio et al. 2014, amongst others), a universal definition is not appropriate as the capabilities for experiments are not equal across study systems. For example, genomic studies of wild subpopulations on an ecological genetics grant cannot gain the same level of proof as a controlled laboratory population with access to an annotated reference genome. Therefore, with the aim to prove causality of an associated candidate gene via the inability to refute it, the advised methods of proving causality applicable to the Davison research group's chirality gene study in *L. stagnalis* are outlined below (Glazier, Nadeau et al. 2002, Page, George et al. 2003, Weigel and Nordborg 2005).

The initial step in identifying causal genes is firstly identifying linkage of the genetic polymorphism with the observed phenotype, enabling accurate genotyping of samples. The associated polymorphism must subsequently be located to a specific region of the genome by fine-scale mapping and further sequence analysis to describe the genes present within the region. As described earlier (section '*Lymnaea* as a model system'), the Davison research group have already finely mapped the region of the genome tightly linked with the chirality phenotype (Liu, Davey et al. 2013). Through further sequencing analysis a number of genes have been identified that lie within

the region, which could represent the cause of inherited chirality. A frame-shift mutation has been identified in the gene coding for a diaphanous formin in the sinistral genotype (Davison et al., *awaiting publication*). This gene represents the primary candidate as the 'chirality gene' in *L. stagnalis* and is described in further detail in Chapter 3.

The presence of a mechanistic link between the candidate gene and the observed phenotype will provide 'biological plausibility' to support the causal relationship of the gene. As *L. stagnalis* lacks a fully annotated reference genome this may not be considered essential, however those genes with functions considered likely to be involved in axis specification and structural developmental will need to be eliminated as candidates with stronger evidence than those that do not.

Support can also be gained for a causal gene by recognition of the same genetic polymorphism resulting in the altered phenotype in another population or species. Unfortunately, there is only one sinistral population of *L. stagnalis* currently available. Therefore, all findings of genetic linkage between chiral morphs are unable to be supported through independent populations. New sinistral individuals of *Lymnaea spp.* have recently been recovered in Kauai, Hawaii and efforts are underway to establish a new laboratory line (Dr Angus Davison, University of Nottingham, Dr. Kenneth Hayes, Howard University, Dr Norine Yeung, Bishop Museum, Hawaii, *pers. comm.*). Yet presently the primary candidate diaphanous formin has not been found to be associated with chirality in other chirally dimorphic snails, *Euhadra* and *Partula* (Davison et al., *awaiting publication*).

Further functional tests will be necessary to support the causative role of the candidate gene. Ideally gene knock-down experiments would provide the proof that the specific candidate gene is directing chirality determination in *L. stagnalis*. Due to the maternal effect of the chirality gene, the mRNA transcripts and likely the gene product are already present and as such, expression interference methods will not be effective in the developing embryo. However, the development of genome editing technologies such as CRISPR-cas (clustered regularly interspaced short palindromic repeats - cas), enable permanent gene modification which is subsequently passed to the next generation (Cong, Ran et al. 2013, Friedland, Tzur et al. 2013). Accordingly, this method of gene knock-down could be effective in disrupting the chirality phenotype. The CRISPR-cas approach needs only a specified target gene sequence similar to PCR primer design (Sander and Joung 2014) and so can be readily applied to *L. stagnalis*. However, the method of delivery to the target genome remains a challenge. As such gene knock-down is a growing possibility for this system, although currently beyond the current capabilities of the Davison research group.

In the absence of gene knock-down experiments, functional properties can be inferred by alternative means. For example, drug treatments, which interfere with protein function specific to the candidate gene, can mimic the effects of gene knockdown (Davison et al., *awaiting publication*). Additionally, gene expression patterns can highlight tissue (spatially) and temporally specific functionality of candidate genes between chiral genotypes.

Project Aims

The aim of this project was to elucidate gene expression patterns associated with chiral variants of *L. stagnalis*. With the intention of providing support for the causal relationship of the primary candidate diaphanous formin and likely downstream effects, whilst also identifying possible mechanistic explanations for the negative pleiotropic effects observed in sinistral *L. stagnalis* embryos (Davison, Barton et al. 2009), this was performed using two different scales of expression analysis. Gene-specific quantitative expression differences were assessed for a number of candidate genes using quantitative real-time PCR (qPCR). Additionally, a transcriptomic sweep was performed using a novel next generation RNA sequencing method to explore the extent of gene expression differences between the three chiral genotypes.

Due to the inherent importance of locality of expression in the determination of LR asymmetry, *in situ* hybridisation methods were also undertaken to explore the localised expression of a number of key candidate genes in *L. stagnalis* embryos. This was performed thanks to collaboration with Dr Daniel Jackson at the Georg August Unversität Göttingen, Germany. However due to limited time to develop the gene-specific assays, results were inconclusive and subsequently not presented in depth within this thesis. A summary of the project is presented in the supplementary information (SI).

Chapter 2:
Validating endogenous control genes for use in quantitative real-time PCR in *Lymnaea stagnalis*

Introduction

The use and misuse of quantitative real-time PCR

Quantitative real-time PCR (qPCR) has become the principal technique for detection and quantification of gene expression and ultimately determining whether varying conditions have an effect on the expression of a specific gene. qPCR offers a flexible method, facilitating a dynamic range of input quantities and accurate within a two-fold range (AppliedBiosystems 2014). A brief description of the principles of qPCR follows. For a historical review of the technique and its development please see Van Guilder et al. (2008).

qPCR employs the conventional polymerase chain reaction (PCR) to amplify a specific DNA sequence within a sample via a pair of complementary sequence primers. qPCR however uses complementary DNA (cDNA) reverse-transcribed from RNA as a template. The level of gene expression is assumed to be reflected in the amount of gene-specific RNA transcripts and subsequently cDNA present within each sample. The addition of a fluorescent dye, which binds to double-stranded DNA, allows the qPCR machine to record a quantitative measure of the amount of PCR product within each reaction, inferred from the intensity of the fluorescent signal after each thermodynamic cycle. Using this information, the qPCR system can generate a real-time amplification curve for each reaction (AppliedBiosystems 2014).

Once the fluorescent signal in a sample has exceeded that of the background fluorescence, it signposts the exponential phase of the reaction and the start of the amplification curve. During the exponential phase of the qPCR, the fluorescent signal will be directly proportional to the amount of template. A cycle threshold (Cq) value can then be calculated from the intersection of the amplification curve and the 'threshold line' (Figure 4). This value corresponds to the number of cycles required to exceed the threshold. Therefore, a high Cq value indicates a low starting quantity of the specific transcript. The threshold line will be the same for all samples compared within the same run (AppliedBiosystems 2011). The Cq value is also frequently referred to as the C_T, C_P, or

occasionally TOP. Here 'Cq' is the term used, as advised by the MIQE guidelines and the Real-time PCR Data Mark-up Language, RDML (www.rdml.org , Bustin, Benes et al. 2009).

There are two different approaches to qPCR; absolute and relative quantification. The analyses throughout this project were intended to determine whether there are differences in gene expression between the chiral genotypes of *Lymnaea stagnalis*. As such absolute quantification was not required, simply to detect whether one genotypic group has a significantly different quantity of transcript relative to the others. Therefore, only relative quantification was performed. In order to calculate the relative quantities of a transcript, one sample within the analysis functions as a reference or calibrator sample. All Cq values are then converted to a fold-change expression level relative to the calibrator sample. However, ensuring that samples are accurately measured relative to one another can prove to be a not insignificant challenge.

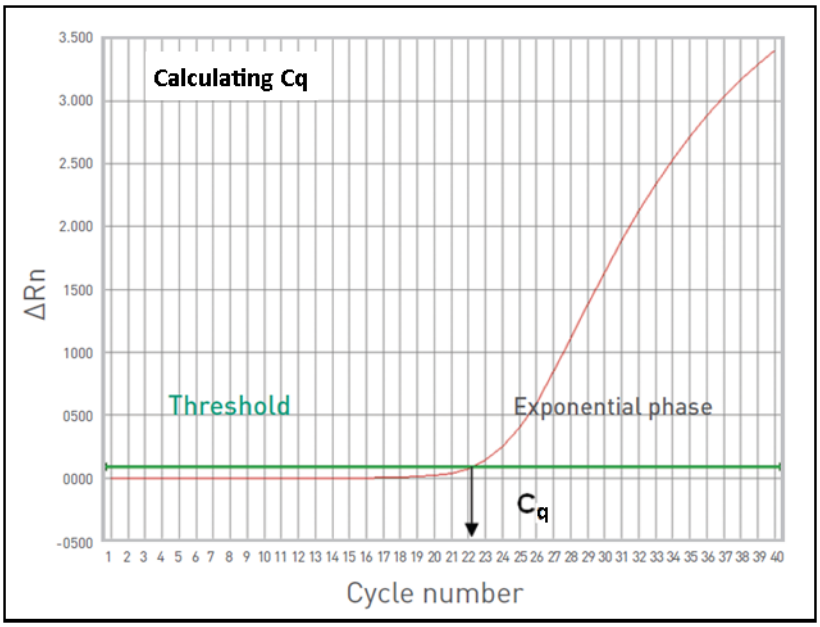


Figure 4 Graphical representation of the calculation of cycle threshold (Cq) values from qPCR data. The fluorescence of the reporter dye, normalised to the passive reference dye and background fluorescence, (ΔRn) is shown according to reaction 'cycle number'. Image adapted from (AppliedBiosystems 2011).

Recently it has become apparent that the ease of access and flexibility of the qPCR technique has led to potentially misleading results, some of which in medical research may have harmful consequences (Baker 2011, Bustin, Benes et al. 2013). The MIQE guidelines published in 2009 (Bustin, Benes et al. 2009) have provided a well-regarded checklist for the **M**inimum **I**nformation for publication of **Q**uantitative real-time PCR **E**xperiments (**MIQE**). The guidelines contain 85 check points for quality control assessment. To complete the entire check checklist however, may not be feasible for all experimental designs due to limited sample, resources and time. A summary of the most important guidelines are described in the following section in order to highlight and explain the potential pitfalls common to previous qPCR methodologies.

Priorities for experimental design

Sample quality

The RNA sample provides the initial template for the qPCR. Thus to ensure biologically meaningful results, samples must be of high quality. Firstly, the presence of chemical impurities (often occurring from carryover in extraction methods, for example, phenol) can inhibit both reverse transcription reactions and qPCR. Contaminating carryover DNA will also significantly compromise the accuracy of the qPCR due to the potential amplification of non-specific products. Therefore, appropriate measures must be taken to ensure a clean and specific method of RNA extraction, including the removal of DNA from samples prior to reverse transcription (AppliedBiosystems 2008).

Secondly the RNA must be structurally intact to provide a reliable template for reverse transcription. RNA integrity can be better maintained by appropriate handling and storage of tissues and RNA samples. 'RNase' is a term which refers to the numerous enzymes which accelerate the degradation of RNA molecules. Due to the general prevalence of RNases, minimal handling and immediate reverse transcription of RNA is recommended to minimise sample degradation (Taylor, Wakem et al. 2010). Furthermore, RNA sampling time/conditions should be controlled between samples. The temporal nature of RNA provides a wealth of biological information, however to make accurate inferences between experimental groups, confounding factors, such as age, diet, diurnal cycle, which may influence expression of the target gene must be minimised.

The vast majority of total RNA consists of ribosomal and other non-coding RNAs, whilst messenger RNAs (mRNAs) are predicted to only represent 1-3% (Palmer and Prediger 2015). mRNAs are more likely to reveal functional variation due to their essential role in protein coding. Therefore, it is common to enrich for mRNA by treating the total RNA sample prior to cDNA synthesis, using Oligo

dT binding. Oligo dTs bind to the distinctive polyadenylated (poly(A)) tail present at the 3' end of mRNAs and as a result can select for mRNA molecules (Aviv and Leder 1972, DeFranscesco 1998).

This mechanism has also been applied to enrich for mRNAs during cDNA synthesis. Reverse transcription requires a short primer sequence complimentary to the RNA strand to initiate the reaction, which follows to generate a copy of the template sequence in the 5'-3' direction. By using oligo dT primers, which complement the poly-A tail present at the 3' end of mRNA, reverse transcription is initiated only on mRNAs. However, with increasing levels of research indicating the functional importance of non-coding RNAs (Mattick and Makunin 2006, Stefani and Slack 2008, Wilusz, Sunwoo et al. 2009), over-enrichment may be obscuring overall expression patterns. Moreover, mRNA enrichment of this manner does not only filter out non-coding RNAs; due to reverse transcription being initiated at the 3' end, long genes may struggle to gain even coverage and suffer a significant underrepresentation of the 5' region. Therefore, a combination of Oligo dT and random hexamer primers are recommended for a well-balanced cDNA synthesis (Taylor, Wakem et al. 2010).

Amplification efficiency

Theoretically during the exponential phase of a qPCR reaction, every copy of the target transcript is doubled after each thermodynamic cycle. However, this assumption is not always met. To reliably quantify relative expression levels requires the Cq values of the target gene to be normalised to those of a stable endogenous control gene (described in greater detail in section 'Endogenous control genes'). To be compared accurately, all primer pairs must perform with the same amplification efficiency, or alternatively, the comparative Cq calculation must incorporate differences in amplification efficiency. Therefore, average primer efficiencies must be known for each primer pair used in the qPCR experiment.

Inter-run calibration

There are two sources of variation in a qPCR experiment; technical and biological. Each qPCR reaction will be slightly different from the last by some level of variation. Therefore, technical replicates must be performed in addition to biological replicates. Three technical replicates per sample is preferred (AppliedBiosystems 2014). If the total number of samples to be assessed in the experiment exceeds one 96 well plate, which is highly likely considering the advised number of controls undertaken, plate-plate technical variation will also be introduced.

If all samples to be compared within the same target gene occur on the same plate, plate-plate noise will not affect the relative comparisons within that gene. However, if samples are to be compared over multiple plates, plate-plate error must be corrected for by the inclusion of the selected

calibrator sample on all experimental plates to be compared. This can become costly in both samples and reagents and it is therefore worth considering plate design for sample maximisation as opposed to gene maximisation (Hellemans, Mortier et al. 2007).

Endogenous control genes

To verify that quantitative differences observed from the qPCR experiment are a true reflection of the relative gene expression between the individuals assessed and not a consequence of technical differences in sample quantity or quality, samples must be standardised across the experiment. There are ways to control for starting quantity across samples, for example, using an equal number of cells in RNA extraction or an equal starting quantity of RNA in cDNA synthesis. However, these methods are not precise enough. A robust method for standardising starting concentrations between samples is to 'normalise' the relative quantities of each target gene of interest (GOI) to those of an internal reference gene, hereafter referred to as an 'endogenous control gene'.

Previously commonly referred to as 'housekeeping genes', perhaps exacerbating the misconception that any 'cell maintenance' gene can be used as a stable calibrator. It is now becoming widely accepted that there are no universal endogenous control genes, and each gene intended for use as an endogenous control must be validated as consistently expressed across all experimental conditions. It is also recommended to use a minimum of three endogenous control genes (Vandesompele, De Preter et al. 2002).

Candidates for stable endogenous control genes in *L. stagnalis*

Of the ten published studies using relative qPCR in *L. stagnalis* (Web of Science, November 2015), only one has described any method of validation of the endogenous control genes used (Bouetard, Besnard et al. 2013). The research areas of these studies for the most part, involve the central nervous system and as such are not likely to be using chiral variants. Consequently, to date, no stable control genes have been described for use across chiral variants. With one notable exception (Bulloch, Diep et al. 2005), the majority of experiments default to the use of ribosomal RNA (rRNA), actins or tubulins as generalised endogenous controls (van Kesteren, Carter et al. 2006, van Nierop, Bertrand et al. 2006, Ribeiro, Schofield et al. 2010, Bavan, Straub et al. 2012, Bouetard, Besnard et al. 2013, Carter, Rand et al. 2015). Others have not used endogenous control genes at all (Hatakeyama, Sadamoto et al. 2004, Wagatsuma, Sadamoto et al. 2005, Azami, Wagatsuma et al. 2006).

rRNAs, such as 18S and 28S rRNA, although commonly employed, are generally not considered to provide suitable endogenous control genes due to a number of reasons. Firstly, if the sample has not

been enriched for mRNA, the over-abundance of rRNAs relative to the target mRNA sequence can lead to problems in accurate normalisation. If the sample has in fact been selected for mRNAs, the rRNAs will not be present in the sample, due to their lack of a poly-A tail (although exceptions have been observed (Slomovic, Laufer et al. 2006)). Additionally, rRNAs are transcribed through an independent pathway from mRNAs and therefore not being regulated in the same manner, and therefore may not relative to the mRNAs being quantified (Radonic, Thulke et al. 2004). In many cases the use of actin and tubulin could provide an appropriate endogenous control, however, due to the focus of this study on cytoskeletal processes, it would have been contradictory to propose genes from within the functional groups expected to vary between conditions as stably expressed control genes.

In the absence of verified stable control genes for use in qPCR studies in *L. stagnalis*, applicable to the analysis of chirality, it was essential to design and evaluate new primers to provide suitable control genes for the intended differential expression analysis of candidate chirality genes.

This experiment aimed to identify differential expression between maternal transcripts and developmental processes. Therefore, endogenous control genes that had been verified as stable in reproductive tissue and throughout developmental stages in a wide range of other species including; frog (Sindelka, Ferjentsik et al. 2006), plant (Pellino, Sharbel et al. 2011), pig (Kuijk, du Puy et al. 2007), and mouse (Jeong, Choi et al. 2005) were considered to be suitable candidates for control genes in this system. Additionally, those genes found to be most stable across a variety of tissues in the original geNorm study (human) (Vandesompele, De Preter et al. 2002), were included as candidates to better accommodate the inclusion of somatic foot tissue.

Endogenous control genes must be validated as stable in all tissues that are to be used in the qPCR experiments. However, the same gene does not have to be used across all experiments. For example, one set of control genes may be used to standardise an analysis of foot tissue samples, whereas another gene may be more appropriate for use in single cell embryo tissue.

Measures of gene expression stability

Three different freely available algorithms, namely; geNorm, (Vandesompele, De Preter et al. 2002), NormFinder (Andersen, Jensen et al. 2004) and BestKeeper (Pfaffl, Tichopad et al. 2004) were employed and their capabilities evaluated. Each of the three methods for calculating gene expression stability includes a unique aspect and therefore using multiple methods will not only support inferences through repeated analysis but will provide additional information.

Methods

Sample Preparation

Three separate tissues were assessed in this study from laboratory reared populations of *L. stagnalis* each with an inbreeding coefficient of more than 98% whilst maintaining chiral dimorphism.

Single-cell embryos of individual self-fertilised mothers were decapsulated and stored in RNAlater® solution (Ambion) at 4°C. Once a sufficient number of embryos had been collected per individual snail (>100), total RNA was extracted using the RNeasy micro kit (Qiagen), yielding approximately 0.5 ng total RNA per embryo (Table 1). All protocols were carried out under the product guidelines.

The ovotestis (hermaphrodite gonad) and foot tissue samples were dissected from individual adult snails and snap frozen using a dry ice/ethanol slurry and total RNA was immediately extracted using TRI Reagent® solution (Applied Biosystems) (Table 2, Table 3).

Complementary DNA (cDNA) was then synthesised from a maximum of 500 ng total RNA, using the first strand synthesis procedure within the home-brew protocol provided by Dr. Susan Bassham (University of Oregon) (Box 1).

The RNeasy micro kit included a DNase treatment in the protocol. The TRI Reagent extracted samples however, required additional steps to remove residual DNA carryover. The total RNA from the ten foot samples was re-extracted using the RNeasy micro kit as a trial to remove carryover DNA using the DNase I provided within the kit. Due to the failure of this process to sufficiently remove carryover DNA, this was not performed on the ovotestis samples. Consequently, no DNase treatment was performed on the ovotestis samples used in this experiment.

The genotypes of the snails used in the ovotestis tissue analyses were inferred from additional DNA extractions and subsequent PCRs using genetic markers previously established for this population. The generalised PCR protocol is described in Box 2. The specific PCR used is referred to in Table 3 and the primer sequences are described in the SI (S2). PCRs were performed using 1 µl of a 1:10 dilution of the DNA sample. The homozygote single cell egg samples were collected from mothers descended from homozygous dominant '*DD*' or recessive '*dd*' lines originating from the same heterozygote (*Dd*) virgin and therefore of known genotype. The foot tissue samples were also extracted from the same homozygote populations and therefore were of known genotype.

cDNA synthesis protocol

Combine:

2 µl	Random Primer Mix (NEB)	
0.8 µl	10 mM dNTP mix	
X µl	(up to 0.5 µg) RNA	
X µl	(if necessary) RNase free water to bring total to 13 µl	**Total: 13 µl**

Heat to 65°C for 5 minutes, then ice

Collect contents at bottom of tube by brief centrifugation.

Add:

4 µl	5x First-Strand Buffer (Invitrogen)	
1 µl	0.1 mM DTT (Invitrogen)	
1 µl	RNase inhibitor (RNAseOUT™ 40u/µl, Invitrogen)	
1 µl	Superscript III reverse transcriptase (200 u/µl, Invitrogen)	**Total: 20 µl**

Mix by gentle aspiration

25°C for 10 min.

[Reaction can be scaled up to accommodate more starting RNA]

Synthesis: Incubate at 50°C for 50 minutes.

Inactivation: 85°C for 5 minutes. Chill on ice, collect contents to bottom by short spin.

Box 1 In-house laboratory protocol for the synthesis of single-stranded cDNA from total RNA. The volume of template varied between reactions and as such is represented by 'x'. The volume of H_2O was adjusted to the input volume of template to attain a final reaction volume of 20 µl thus is also represented by 'x'.

The total RNA samples were often too small to allow for a thorough sample quality assessment. However, every sample was quantified via a spectrophotometer (NanoDrop2000, Thermo Fisher Scientific). Additionally, a small aliquot of a number of total RNA samples were visualised via electrophoresis on agarose gel to provide a visual overview of the sample quality. This was performed at least once for each RNA extraction method (S3).

Non-quantitative PCRs were also performed on all cDNA samples according to the protocol described in Box 2. PCR amplification of the target gene sequences functioned as a positive control for both the primer pair and the cDNA sample prior to commencing the qPCR reactions. Another PCR employing primers specific to an intronic sequence region was performed on all samples to test for the presence of carryover genomic DNA (S5). A consistent genomic DNA sample of an individual '*DD*' *L. stagnalis* was used as a positive control and PCR grade water as a negative control in all reactions. The PCR products were visualised via gel electrophoresis using ethidium bromide as a fluorescent marker.

All RNA samples were stored at -80°C and all cDNA samples were stored at -20°C. Aliquots were made of the experimental working concentration dilutions of cDNA to reduce freeze-thaw cycles, whereas serial dilutions were performed independently for each standard curve experiment. All cDNA samples were moderately vortexed before use and prior to each serial dilution step.

Non-quantitative PCR reaction setup and cycle parameters

Per reaction:

2 µl	10x PCR Buffer (ThermoFisher Scientific)
1.2 µl	$MgCl_2$ solution (ThermoFisher Scientific)
0.8 µl	8 µM dNTP mix
2 µl	10 µM forward & reverse primer mix
0.1 µl	AmpliTaq Gold® DNA polymerase (ThermoFisher Scientific)
x µl	Template (DNA/cDNA)
x µl	H_2O (PCR grade)

[Total reaction volume = 20 µl]

Thermocycling parameters:

1. 98°C 10 mins
2. 98°C 30 secs
3. 58°C 30 secs
4. 72°C 60 secs
5. Cycle from step 2, 34 more times
6. 72°C 5 mins

-END-

Box 2 Generalised non-quantitative PCR protocol. The volume of template varied between reactions and as such is represented by 'x'. The volume of H_2O was adjusted to the input volume of template to attain a final reaction volume of 20 µl thus is also represented by 'x'.

Table 1 Details of RNA extraction and cDNA synthesis for the single cell embryo samples used in the endogenous control gene stability assessment. Table includes: sample identifier (ID) and genotype (Geno) of the mother snail; Spectrophotometry data of the Total RNA sample including sample concentration (ng/µl) and 260/280 & 260/230 absorbance ratios; volume (µl RNA) and quantity (ng RNA) of total RNA used for cDNA synthesis.

ID	Tissue	Geno	Extraction Date	RNA extraction method	DNase Treatment	Total RNA			cDNA synthesis	
						ng/µl	260/280	260/230	µl RNA	ng RNA
11289	Embryo	*DD*	25/11/2014	RNeasy micro kit	DNase I	13.80	1.80	0.22	10.0	138.0
11292	Embryo	*DD*	07/11/2014	RNeasy micro kit	DNase I	9.90	2.40	0.73	9.3	92.1
11293	Embryo	*DD*	20/11/2014	RNeasy micro kit	DNase I	15.10	2.20	0.41	10.0	151.0
11295	Embryo	*DD*	25/11/2014	RNeasy micro kit	DNase I	10.20	1.74	1.28	10.0	102.0
11297	Embryo	*DD*	07/11/2014	RNeasy micro kit	DNase I	11.60	1.83	0.87	9.4	108.5
11298	Embryo	*DD*	28/10/2014	RNeasy micro kit	DNase I	21.00	1.82	1.05	9.0	189.0
11282	Embryo	*dd*	03/11/2014	RNeasy micro kit	DNase I	13.70	2.44	0.60	8.8	120.6
11283	Embryo	*dd*	05/11/2014	RNeasy micro kit	DNase I	9.80	1.92	0.44	10.0	98.0
11284	Embryo	*dd*	27/10/2014	RNeasy micro kit	DNase I	12.90	1.73	1.12	9.3	120.0
11287	Embryo	*dd*	05/11/2014	RNeasy micro kit	DNase I	8.00	2.02	1.51	9.4	75.2
11301	Embryo	*dd*	20/11/2014	RNeasy micro kit	DNase I	12.10	2.23	0.65	10.0	121.0
11303	Embryo	*dd*	25/11/2014	RNeasy micro kit	DNase I	13.30	2.30	1.60	10.0	133.0

Table 2 Details of RNA extraction and cDNA synthesis for the foot tissue samples used in the endogenous control gene stability assessment. Table includes: sample identifier (ID) and genotype (Geno) of the individual snail; Spectrophotometry data of the total RNA sample including sample concentration (ng/µl) and 260/280 & 260/230 absorbance ratios; volume (µl RNA) and quantity (ng RNA) of total RNA used for cDNA synthesis.

ID	Tissue	Geno	Extraction Date	RNA extraction method	DNase Treatment	Total RNA			cDNA synthesis	
						ng/µl	260/280	260/230	µl RNA	ng RNA
11347	Foot	*DD*	11/03/2015	TRIreagent, RNeasy micro kit	DNase I	72.02	1.95	1.40	6.9	496.9
11350	Foot	*DD*	12/03/2015	TRIreagent, RNeasy micro kit	DNase I	49.54	1.96	1.51	10.1	500.4
11351	Foot	*DD*	12/03/2015	TRIreagent, RNeasy micro kit	DNase I	85.33	2.19	2.26	5.9	503.4
11352	Foot	*DD*	12/03/2015	TRIreagent, RNeasy micro kit	DNase I	67.62	2.10	1.95	7.4	500.4
11357	Foot	*DD*	13/03/2015	TRIreagent, RNeasy micro kit	DNase I	62.15	2.23	1.25	8.0	497.2
11348	Foot	*dd*	12/03/2015	TRIreagent, RNeasy micro kit	DNase I	74.69	2.03	2.12	6.7	500.4
11349	Foot	*dd*	12/03/2015	TRIreagent, RNeasy micro kit	DNase I	70.75	2.05	2.03	7.1	502.3
11353	Foot	*dd*	13/03/2015	TRIreagent, RNeasy micro kit	DNase I	69.2	2.41	1.36	7.2	498.2
11354	Foot	*dd*	13/03/2015	TRIreagent, RNeasy micro kit	DNase I	78.2	2.07	1.72	6.4	500.5
11356	Foot	*dd*	13/03/2015	TRIreagent, RNeasy micro kit	DNase I	76.98	2.09	1.98	6.5	500.4

Table 3 Details of RNA extraction and cDNA synthesis for the ovotestis tissue samples used in the endogenous control gene stability assessment. Table includes: sample identifier (ID) and genotype (Geno) of the individual snail, [2] PCR 1315-507 used to identify genotype; Spectrophotometry data of the total RNA sample including sample concentration (ng/µl) and 260/280 & 260/230 absorbance ratios; volume (µl RNA) and quantity (ng RNA) of total RNA used for cDNA synthesis. † sample removed from analysis.

ID	Tissue	Geno	Extraction Date	RNA extraction method	DNase Treatment	Total RNA			cDNA synthesis	
						ng/µl	260/280	260/230	µl RNA	ng RNA
10627†	Ovotestis	*DD*[2]	25/09/2013	TRIreagent	n/a	60.1	1.8	1.5	16.6	498.9
10633	Ovotestis	*DD*[2]	12/09/2013	TRIreagent	n/a	57.8	1.9	0.9	17.3	499.9
10636	Ovotestis	*DD*[2]	25/09/2013	TRIreagent	n/a	82.4	1.8	1.8	12.1	498.6
10638	Ovotestis	*DD*[2]	13/09/2013	TRIreagent	n/a	67.5	1.8	1.9	14.8	499.2
10622	Ovotestis	*Dd*[2]	25/09/2013	TRIreagent	n/a	62.1	1.8	1.6	16.1	499.7
10629	Ovotestis	*Dd*[2]	13/09/2013	TRIreagent	n/a	59.5	1.9	1.7	16.8	499.7
10631†	Ovotestis	*Dd*[2]	11/09/2013	TRIreagent	n/a	68.6	1.9	1.1	14.6	500.4
10639	Ovotestis	*Dd*[2]	11/09/2013	TRIreagent	n/a	208.9	1.9	1.8	4.7	490.9
10626	Ovotestis	*dd*[2]	25/09/2013	TRIreagent	n/a	74.2	1.9	1.0	13.4	497.2
10630	Ovotestis	*dd*[2]	12/09/2013	TRIreagent	n/a	64.1	1.8	1.0	15.6	499.7
10640†	Ovotestis	*dd*[2]	11/09/2013	TRIreagent	n/a	73.1	1.8	1.3	13.7	500.9
10642	Ovotestis	*dd*[2]	13/09/2013	TRIreagent	n/a	53.2	1.9	1.7	18.7	497.4

Primer design

Targets

Candidate endogenous control genes were selected based on a number of previously published qPCR studies which indicated good potential normalising controls for the reproductive tissue and developmental stages. Of the studies which demonstrated good validation of their endogenous controls, common control genes included; elongation factors, various ubiquitin genes, Actin 2, glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*), and various histone proteins (Vandesompele, De Preter et al. 2002, Jeong, Choi et al. 2005, Sindelka, Ferjentsik et al. 2006, Kuijk, du Puy et al. 2007, Pellino, Sharbel et al. 2011). Finally, those identified as the most stably expressed in the original geNorm paper, including GAPDH, Hypoxanthine phosphoribosyl-transferase 1 (HPRTI), Ubiquitin C (UBC), Ribosomal protein L13a (RPL13A), succinate dehydrogenase (SDHA) and Tyrosine 3-monooxygenase/ tryptophan 5-monooxygenase activation protein, zeta polypeptide (YWHAZ), represented wider-ranging candidate genes likely to be stably expressed in multiple tissues (Vandesompele, De Preter et al. 2002).

Although actin variants frequently appear as stable normalising control genes, none were included in this analysis because cytoskeletal processes are likely to be involved in the processes investigated.

Using transcriptome resources of 1-2 cell stage *L. stagnalis* embryos (Liu, Davey et al. 2014), six genes relating to those highlighted above were selected for analysis. Gene functions were predicted from sequence similarity to published human housekeeping genes. These included; short-chain specific acyl-CoA dehydrogenase (*Lacads*), as a substitution for SDHA; elongation factor 1-alpha (*Lef1a*); histone protein, H2A (*Lhis2a*); 60S ribosomal protein L14 (*Lrpl14*); Ubiquitin-conjugating enzyme E2 (*Lube2*); and 14-3-3 protein zeta (*Lywhaz*). This provided species specific sequence information and the confidence that the transcripts were present in the one cell stage embryo and ovotestis. Additionally, the level of expression was predicted to be neither extremely high nor low inferred from human housekeeping gene expression data (data presented in the SI, S4 (Eisenberg and Levanon 2013, Liu, Davey et al. 2014)). All gene abbreviations were given the prefix 'L' to denote the *L. stagnalis* specific gene sequence (Table 5).

All primer pairs were designed from the aforementioned 1-2 cell stage transcriptome sequence data, using freely available software Primer 3 (Untergasser, Cutcutache et al. 2012). Primer pairs were designed to have a Tm range within 2°C of each other and amplicon product size between 110-130bp. GC clamps where included where possible, and primer pairs were selected with the lowest available 'Th' scores. The increased strength of bonding between G and C bases help to promote

specific binding at the 3' end of the primer and 'Th' scores provide a measure of the likelihood of the primer binding to a region other than that specified based on thermodynamic secondary structure alignments. Both of these features were included in attempt to increase primer specificity. The melting temperature of all primer pairs ranged from 57.45°C-59.75°C (Table 5). Due to the limited variability in melting temperature, the same cycling parameters were used for all qPCR experiments.

All primers pairs were also designed to be exon-spanning. The position of introns was inferred by performing a local blast of the transcriptome sequence against the current *Lymnaea* genome assembly (version 10), generated by the Davison research group. Pairwise alignments of the two sequences were then generated using NCBI blast online (www.ncbi.nlm.nih.gov). Consequently, in the event of genomic contamination, the primers were either unable to span the length of the intron and did not amplify a genomic product, or amplified a product at a substantially different size, detectable during the melt curve stage of the qPCR (described in section 'Quantitative real-time PCR').

Primer specificity

All primer pairs were first tested in a conventional non-quantitative PCR alongside a genomic control sample (described in section 'Sample Preparation') and the products visualised on an agarose gel to verify the expected size of products. Additionally, the sequence specificity of the amplicons of all six primer pairs was verified through Sanger sequencing of conventional PCR products generated from pooled heterozygote single cell embryo cDNA samples.

Primer amplification efficiency

To calculate the primer efficiencies, standard curve qPCR experiments were performed on each primer pair used in this experiment. Five standardised concentrations were used with an additional negative control. Because absolute concentrations of the cDNA samples were unknown, standard concentrations were produced by using serial dilutions of the original cDNA sample. Five step serial dilutions were performed using a dilution factor of 1:5, 1 part cDNA/previous cDNA dilution, plus 4 parts PCR grade water. Primer efficiencies for all six endogenous control gene primer pairs were established using the same reference sample, created from pooling all 12 ovotestis cDNA samples.

The starting concentration of the serial dilutions alternated between a 1:3 or 1:6 dilution of the original concentration. Therefore, some primer efficiency trials provide results for the amplification efficiency of cDNA sample quantities ranging from 33.33% - 0.05% of the original concentration, whereas others assessed 16.67% - 0.03% (full details of the range of concentrations assessed are presented in Table 4). The standard curve qPCR was carried out by the same method as that for the

comparative qPCR described in section 'Quantitative real-time PCR'. Accordingly, 3 µl of cDNA was used in each reaction (Box 3).

Average primer efficiencies for each primer pair were calculated via the arithmetic mean of a minimum of two successful standard curve experiments. A standard curve experiment was considered successful if it produced a R^2 value of >0.98. Values from the lowest concentration dilutions were omitted if they dramatically reduced the amplification efficiency or R^2 value of an experiment. The average primer efficiencies are quoted as the amplification efficiency within the concentration range they were successfully calculated from. This range indicates the limits of acceptable working concentration/dilution factor for an experimental qPCR assessment.

Table 4 Details of the five-step serial dilutions used for standard curve qPCR experiments to assess amplification efficiency using a starting concentration of 1:3 full or 1:6. Concentrations are represented as both a percentage of the full concentration cDNA (% full conc.) and dilution ratio (ratio).

Serial dilutions	Starting Conc.	1:5	1:25	1:125	1:625
	1	2	3	4	5
% full conc., / ratio	33.33 / 1:3	6.67 / 1:15	1.33 / 1:75	0.27 / 1:375	0.05 / 1:1875
	16.67 / 1:6	3.33 / 1:30	0.67 / 1:150	0.13 / 1:750	0.03 / 1:3750

Quantitative real-time PCR

It cannot be assumed that the pattern of gene expression will be equal across different tissues. Therefore, separate analyses were performed for each of the three tissues included in this study. All tissue samples were extracted from separate individual snails.

The embryo analysis was performed on cDNA from pools of single-cell embryos collected from 12 individual *L. stagnalis*. These comprised six *DD*, and six *dd* individuals (Table 1). The foot tissue analysis was performed on cDNA of ten adult *L. stagnalis* individuals. These comprised five *DD*, and five *dd* individuals (Table 2). The ovotestis tissue analysis was performed on cDNA of nine adult *L. stagnalis* individuals. These comprised of three dextral homozygous (*DD*), three dextral heterozygous (*Dd*) and three sinistral homozygous individuals (*dd*) (Table 3).

Cq values were obtained from qPCR experiments using the AB 7500 fast system (Applied Biosystems) and Primer Design's fast SYBR® green master mix. 3 µl of cDNA were used in each well. All samples were used at a 1:30 dilution of the original cDNA concentration (alternately described as 3.33% of the full concentration). Because 3 µl were used instead of the more commonly used 1 µl, this could be considered to represent a 1:10 dilution. Mastermixes were prepared for each target gene experiment following the reaction setup described in Box 3. Also presented in Box 3 are the thermocycling parameters used in each qPCR experiment.

Table 5 Primer sequence information for amplification of endogenous control gene targets including: primer name and associated protein with accession number (Acc. No.) of its most closely related human gene; gene abbreviation (Abv.) used throughout this analysis; Primer sequence in the 5' to 3' direction; Primer length (P.L) & amplicon length in nucleotides (A.L); primer melting temperature (Tm) and the difference between melting temperature within each primer pair (Tm diff); the estimate of mispriming to any sequence (Any th) and specifically mispriming at the 3' end (3' th); and the predicted intron size between the two primers. * primer lies on an exon boundary. †full intron information unknown due to the transcriptomic sequence crossing two genomic contigs, the minimum intron size is presented.

Primer Name	Associated Protein	Acc. No.	Abv.	Sequence 5'-3'	P.L	A.L	Tm	Tm diff	%GC	Any th	3' th	Predicted Intron
ACA_11210_F1	Short-chain specific acyl-CoA dehydrogenase	NM_014049	*Lacads*	TGCACTCTCTAAACGAACTTCC	22	117	58.35	0.42	45.45	0	0	866
ACA_11210_R1				TCCCTTGATTGTGCTGTTGAC	21		58.77		47.62	0	0	
EF1_8940_F1	Elongation factor 1-alpha	NM_006620	*Lef1a*	CGTCACAACCAGCATATCCC	20	113	58.7	0.77	55	0	0	663
EF1_8940_R1				AGAGTTCGAGGGCTGCTTAC*	20		59.47		55	0	0	
HiS_8200_F1	Histone H2A	NM_012412	*Lhis2a*	TCAGAGGAGATGAGGAGTTGG	21	123	58.26	0.63	52.38	0	0	785†
HiS_8200_R1				CCCCAAGTTATGCTGCCTTC	20		58.89		55	0	0	
RPL_2341_F2	60S ribosomal protein L14	NM_003973	*Lrpl14*	TAATAAGTCGGTTGCGCGC*	19	114	59	1.55	52.63	27.26	27.26	1255
RPL_2341_R2				GGGAACAGTCTACTTGGGC	19		57.45		57.89	0	0	
UB_3288_F2	Ubiquitin-conjugating enzyme E2	NM_003336	*Lube2*	GCGGATCCTCTTGCAATCTT*	20	131	58.33	0.3	50	0	0	3224
UB_3288_R2				TCTGTGGACTGCATATCACTCT	22		58.63		45.45	0	0	
YWHAZ_562_F1	14-3-3 protein zeta	NM_006761	*Lywhaz*	GGAGGAGCTGAAGTCAATATGC	22	125	58.86	0.78	50	0	0	711
YWHAZ_562_R1				AGTCACCCTGCATTTTGAGG	20		58.08		50	0	0	

All samples were performed in triplicate repeat for each of the six reference genes and negative control wells were included per mastermix in duplicate repeat. This necessitated the use of 192, 174 and 228 wells for the foot, ovotestis and embryo analyses respectively. The whole analysis for each tissue exceeded the capacity for one 96 well plate and consequently was divided across multiple plates. All samples per gene were included on the same plate. No experiment exceeded three plates and when multiple plates were used all master mixes and plates were prepared at the same time to reduce experimental noise (Figure 5).

A temperature melt curve step was included at the end of all qPCR reactions. During this step the temperature of the reaction was incrementally increased whilst continuing to record fluorescence (Box 3). Because the SYBR® green dye fluoresces when associated with double stranded DNA, the signal will decrease as the DNA is melted and becomes single-stranded. A sharp single peak in the melt curve indicates that only one specific PCR product has been amplified

qPCR reaction setup & cycle parameters

Per well:

5µl	SYBR green Mastermix (2x) (Primer Design)
0.5µl	forward & reverse primer mix (4µM)
3µl	cDNA (concentration specified per experiment)
1.5µl	H_2O (PCR grade)

[Total reaction volume = 10µl]

Thermal cycling parameters:

1. 95°C 20s [initial temperature ramp & hold]
2. 95°C 3s
3. 60°C 30s (data collection, Cq) [cycle to step 2; 39 more times]
4. 95°C 15s
5. 60°C 60s
6. slow ramp 1% (data collection; Tm melt curve)
7. 95°C 15s
8. 60°C 15s

-END-

Box 3 Details of qPCR reaction setup per well and the following thermal cycling parameters used for all qPCR experiments described.

Cq values were exported for each well of the experiment using the 7500 software. Average Cq values derived from triplicate repeats of each sample were used in analyses. Only samples with standard deviation (SD) of <0.5 were used in the analysis. This occasionally involved removing perceived outliers (observable from the presence of a substantially different shaped amplification curve) from the dataset, leaving some individual samples in only duplicate repeat. Due to failure to amplify in two of the 3 replicates, one sample was included with no data from technical replicates.

Normalising control software:

Three algorithms were used to assess the same qPCR data, all of which run as macros within Microsoft Excel 2003. The BestKeeper applet used raw Cq values, whereas NormFinder and geNorm required linearised Cq values. All could accommodate corrections for amplification efficiency.

Linearised relative Cq values were calculated for each sample by first subtracting the average Cq value of the nominated calibrator sample from the average Cq of the sample to create a relative, or delta, Cq value (ΔCq). Amplification efficiency corrected ΔCq values were then calculated by multiplying the efficiency by the power of the ΔCq value (Equation 1). The BestKeeper applet ran entirely from raw Cq values and corrected for amplification efficiency via the inbuilt formulas within the applet using the manually input amplification efficiency values.

Equation 1 Formula based on Pfaffl's method (Hellemans, Mortier et al. 2007) to calculate linearised Cq values which incorporate the amplification efficiency of each target. E = amplification efficiency, represented as a value between 1 and 2.

$$E^{\Delta Cq(\text{Sample Cq- Calibrator Cq})}$$

Results

General QC:

Primer specificity

Sanger sequencing confirmed that the amplicons of all six primer pairs were specific to the transcript they were designed from. Additional PCRs on cDNA and genomic DNA analysed via fluorescent gel electrophoresis, showed amplified products to be of the expected size and additionally demonstrated the difference in amplicon size of a product generated from a cDNA or a genomic DNA reaction. There was no visible amplification of multiple products from any of the samples (Figure 6, cDNA product).

a

Em	1	2	3	4	5	6	7	8	9	10	11	12
A	11289	11292	11293	11295	11297	11298	11282	11283	11284	11287	11301	11303
B	11289	11292	11293	11295	11297	11298	11282	11283	11284	11287	11301	11303
C	11289	11292	11293	11295	11297	11298	11282	11283	11284	11287	11301	11303
D	11289	11292	11293	11295	11297	11298	11282	11283	11284	11287	11301	11303
E	11289	11292	11293	11295	11297	11298	11282	11283	11284	11287	11301	11303
F	11289	11292	11293	11295	11297	11298	11282	11283	11284	11287	11301	11303
G	x	x	x	x	x	x	x	x	x	x	x	x
H	H_2O	H_2O	H_2O	H_2O	x	x	x	x	x	x	x	x

b

Fo	1	2	3	4	5	6	7	8	9	10	11	12
A	11347	11348	11349	11350	11351	11352	11353	11354	11356	11357	11347	11348
B	11347	11348	11349	11350	11351	11352	11353	11354	11356	11357	11349	11350
C	11347	11348	11349	11350	11351	11352	11353	11354	11356	11357	11351	11352
D	11347	11348	11349	11350	11351	11352	11353	11354	11356	11357	11353	11354
E	11347	11348	11349	11350	11351	11352	11353	11354	11356	11357	11356	11357
F	11347	11348	11349	11350	11351	11352	11353	11354	11356	11357	H_2O	H_2O
G	11347	11348	11349	11350	11351	11352	11353	11354	11356	11357	H_2O	H_2O
H	11347	11348	11349	11350	11351	11352	11353	11354	11356	11357	H_2O	H_2O

c

Ov	1	2	3	4	5	6	7	8	9	10	11	12
A	10633	10636	10638	10622	10629	10639	10626	10630	10642	H_2O	10633	10622
B	10633	10636	10638	10622	10629	10639	10626	10630	10642	H_2O	10636	10629
C	10633	10636	10638	10622	10629	10639	10626	10630	10642	H_2O	10638	10639
D	10633	10636	10638	10622	10629	10639	10626	10630	10642	H_2O	10626	H_2O
E	10633	10636	10638	10622	10629	10639	10626	10630	10642	H_2O	10630	x
F	10633	10636	10638	10622	10629	10639	10626	10630	10642	H_2O	10642	x
G	10633	10636	10638	10622	10629	10639	10626	10630	10642	H_2O	x	x
H	10633	10636	10638	10622	10629	10639	10626	10630	10642	H_2O	x	x

Figure 5 qPCR endogenous control experimental plate setup. The total number of embryo samples included plus the negative controls (H_2O) enabled the analysis of two control genes per plate (indicted by the different background colour)(a), whereas the fewer samples in the foot experiments (b) and the ovotestis experiments (c) enabled the inclusion of three control genes per plate. Unused wells are indicated by 'x'.

All of the primer pairs used in this experiment were designed to contain introns. As a result, some could not amplify a product from a genomic template. This was the case for *Lef1a*, *Lrpl14* and *Lube2*. It was not always possible to generate the best primers over a long intron. As such, some primer pairs could amplify a genomic product, although the resulting amplicon was of a substantial size difference, as seen in *Lacads*, *Lhis2a* and *Lywhaz* (Figure 6, genomic DNA product).

The melt temperature (Tm) curves of all qPCR reactions showed distinct peaks for experimental samples. A smaller peak at a lower Tm was occasionally visible in the negative controls of some genes, namely *Lhis2a*, *Lrpl14*, *Lube*2 and *Lywhaz* (Figure 7). This is indicative of primer dimer.

Sample quality/Genomic contamination

The representative total RNA samples that were visualised by gel electrophoresis, generally displayed two distinct bands, which are indicative of the abundant size-specific rRNA transcripts (gels are presented in the SI). Therefore, the samples were assumed to be non-degraded (S3). The intronic PCRs showed no amplification of a genomic product in any of the embryo sample, whereas the ovotestis and foot samples all amplified a genomic product (S5).

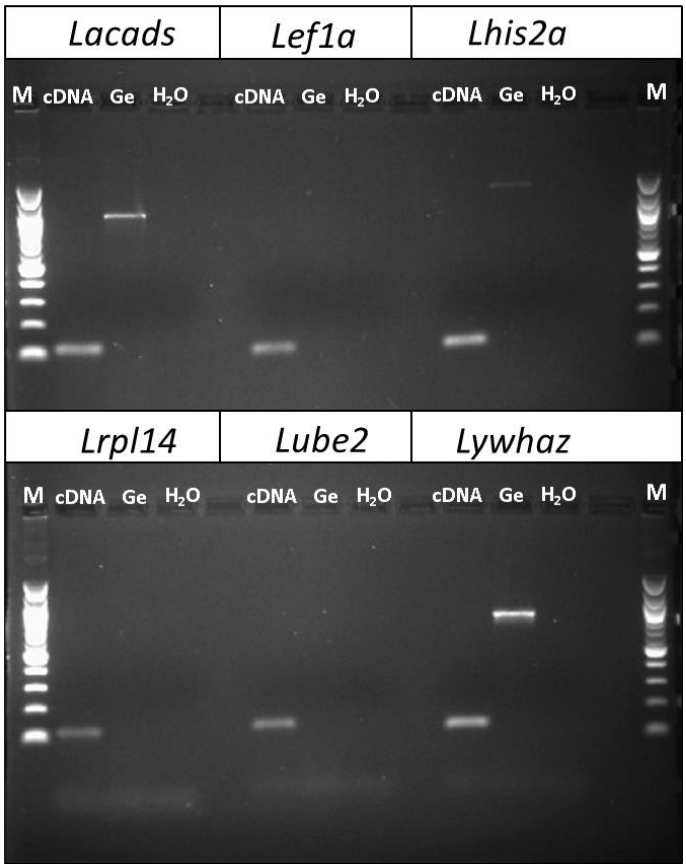


Figure 6 UV visualisation via agarose gel electrophoresis of PCR products of the six endogenous control genes, amplified from three different templates: cDNA, genomic DNA (Ge) and a negative control (H_2O). 100 base pair ladder was included as a size marker (M).

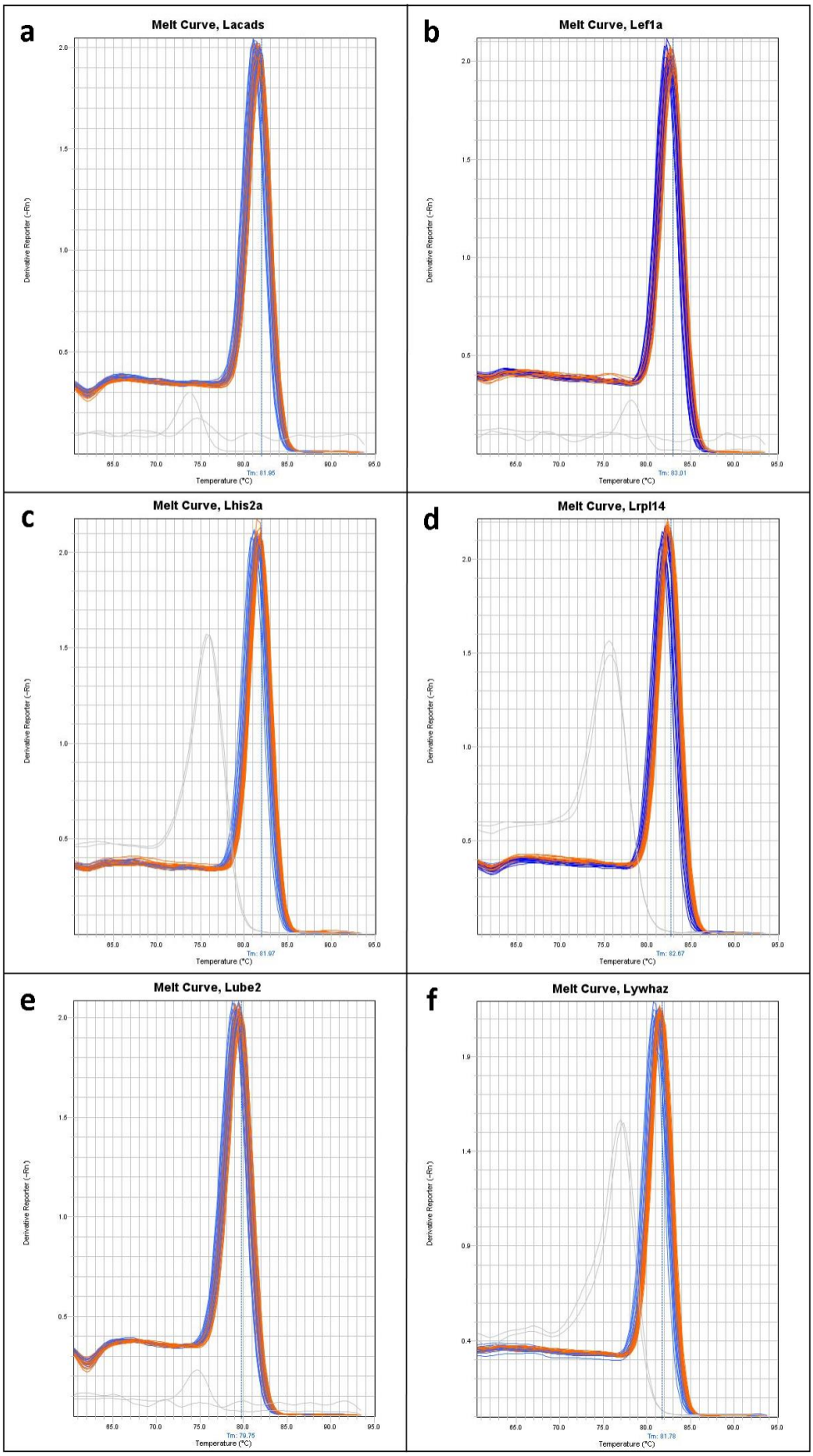


Figure 7 Representative temperature melt (Tm) curves of qPCR amplification of *Lacads* (a); *Lef1a* (b); *Lhis2a* (c); *Lrpl14* (d); *Lube2* (e) and *Lube2* (f). Tm curves were produced from *DD* (blue) and *dd* (red) embryo samples. Negative controls are shown in grey. The melt curves are presented to demonstrate specificity via shape not absolute values.

Primer efficiencies

Amplification efficiencies of each primer pair were estimated by performing standard curve quantification of serial dilutions of a pooled cDNA sample. Primer pairs demonstrated amplification efficiencies between 1.906 and 2.115 with R^2 values exceeding 0.98 and are presented in Table 6. The minimum detectability template concentrations are presented as a percentage of the undiluted full concentration cDNA required for the qPCR reaction to perform within the estimated amplification efficiency. All primers demonstrated acceptable amplification efficiency in dilutions of up to 1:150/0.67% of full concentration (Table 6). The working concentration of a 1:30 dilution used in the qPCR experiments falls well within these limits.

Raw Cq data and linearised Cq values

The raw Cq data is presented in Table 7, Table 8 and Table 9. The linearised Cq data for each tissue is shown in Table 10. The omission of perceived outliers in technical replicates is denoted for each sample by the n value. Descriptive details of the omitted data points are presented in the SI (S6).

Table 6 Amplification efficiency estimates of each primer pair for the six endogenous control genes assessed represented by their gene abbreviation (Abv.). The average efficiency is quoted as the amount each template will increase per qPCR cycle (between 1 and 2). The minimum dilution is presented as a percentage of the undiluted original cDNA concentration required in the qPCR reaction. Additionally, the number of runs included to generate the average amplification efficiency is quoted and the tissue the experiments were performed on: ^{O} = Ovotestis reference sample.

Primer Pair	Abv.	Efficiency (R2 > 0.98)	Minimum Dilution (%)	No. of runs included
ACA_11210_F1R1	*Lacads*	1.912	0.27	2^{O}
EF1_8940_F1R1	*Lef1a*	2.115	0.67	2^{O}
HiS_8200_F1R1	*Lhis2a*	1.943	0.03	2^{O}
RPL_2341_F2R2	*Lrpl14*	1.906	0.03	2^{O}
UB_3288_F2R2	*Lube2*	1.923	0.03	2^{O}
YWHAZ_562_F1R1	*Lywhaz*	1.918	0.03	3°

Table 7 Average Cq values (Cq Mean) and associated standard deviation (SD) calculated from technical replicates (n) of 12 embryo samples for six endogenous control genes. Including sample ID, genotype (Geno) and tissue description. *amplification observed in negative controls

Raw Cq data			*Lacads*			*Lef1a*			*Lhis2a**			*Lrpl14**			*Lube2*			*Lywhaz**		
ID	**Geno**	**Tissue**	**Cq Mean**	**Cq SD**	**n**	**Cq Mean**	**Cq SD**	**n**	**Cq Mean**	**Cq SD**	**n**	**Cq Mean**	**Cq SD**	**n**	**Cq Mean**	**Cq SD**	**n**	**Cq Mean**	**Cq SD**	**n**
11289	*DD*	Embryo	23.177	0.052	3	26.868	0.266	3	22.167	0.032	3	23.116	0.110	3	25.927	0.222	3	24.757	0.071	3
11292	*DD*		22.558	0.183	3	26.268	0.429	3	21.881	0.064	3	22.823	0.060	2	25.326	0.065	3	23.833	0.063	3
11293	*DD*		22.811	0.046	3	26.508	0.322	3	21.989	0.061	3	22.708	0.078	3	25.710	0.134	3	24.356	0.065	3
11295	*DD*		23.891	0.091	3	27.805	0.239	2	22.687	0.045	3	23.802	0.145	3	26.286	0.077	3	24.937	0.092	3
11297	*DD*		22.836	0.106	3	26.682	0.406	3	21.901	0.053	3	22.633	0.077	3	25.584	0.157	3	24.201	0.067	3
11298	*DD*		22.042	0.083	3	25.773	0.021	3	21.177	0.088	3	21.857	0.088	3	24.819	0.130	3	23.413	0.019	3
11282	*dd*		22.572	0.115	3	26.430	0.158	3	21.841	0.119	3	22.910	0.109	3	25.315	0.084	3	23.616	0.103	3
11283	*dd*		23.455	0.062	3	27.424	0.132	3	22.801	0.033	3	23.260	0.100	3	26.304	0.076	3	24.803	0.117	3
11284	*dd*		22.066	0.109	3	26.075	0.276	3	21.528	0.027	3	22.166	0.112	3	25.065	0.089	3	23.782	0.029	3
11287	*dd*		23.403	0.122	3	27.768	0.468	2	22.653	0.164	3	23.838	0.072	3	26.785	0.346	3	24.556	0.167	3
11301	*dd*		23.630	0.251	3	27.355	0.296	3	22.531	0.048	3	23.483	0.042	3	26.393	0.068	3	24.906	0.062	3
11303	*dd*		22.600	0.118	3	27.442	0.130	2	22.373	0.182	3	23.309	0.108	3	26.294	0.065	3	23.988	0.144	3

Table 8 Average Cq values (Cq Mean) and associated standard deviation (SD) calculated from technical replicates (n) of 10 foot samples for six endogenous control genes. Including sample ID, genotype (Geno) and tissue description. *amplification observed in negative controls

Raw Cq data			*Lacads*			*Lef1a*			*Lhis2a**			*Lrpl14**			*Lube2*			*Lywhaz**		
ID	**Geno**	**Tissue**	**Cq Mean**	**Cq SD**	**n**	**Cq Mean**	**Cq SD**	**n**	**Cq Mean**	**Cq SD**	**n**	**Cq Mean**	**Cq SD**	**n**	**Cq Mean**	**Cq SD**	**n**	**Cq Mean**	**Cq SD**	**n**
11347	*DD*	Foot	28.918	0.137	3	35.870	n/a	1	25.899	0.120	3	21.493	0.045	3	25.341	0.123	3	24.361	0.026	3
11350	*DD*		25.876	0.068	3	32.601	0.003	2	21.520	0.032	3	18.721	0.037	3	21.208	0.077	3	20.685	0.031	3
11351	*DD*		27.559	0.078	3	34.042	0.225	2	23.825	0.050	3	19.724	0.066	3	23.440	0.021	3	22.670	0.086	3
11352	*DD*		26.685	0.033	3	33.649	0.204	2	22.327	0.066	3	18.921	0.038	3	21.822	0.058	3	21.318	0.050	3
11357	*DD*		26.187	0.073	3	32.478	0.358	2	21.777	0.105	3	18.975	0.010	3	21.374	0.175	3	20.930	0.019	3
11348	*dd*		27.463	0.084	3	33.350	0.070	2	23.649	0.135	3	19.446	0.056	3	22.626	0.115	3	22.173	0.064	3
11349	*dd*		27.470	0.204	3	33.394	0.495	3	23.436	0.107	3	19.408	0.046	3	22.577	0.073	3	22.151	0.024	3
11353	*dd*		27.246	0.166	3	33.218	0.418	3	23.788	0.008	3	19.563	0.078	3	22.647	0.058	3	22.150	0.076	3
11354	*dd*		26.498	0.035	3	32.317	0.266	3	22.591	0.020	3	19.267	0.087	3	21.428	0.088	3	21.277	0.115	3
11356	*dd*		27.035	0.089	3	32.778	0.137	3	23.559	0.109	3	19.876	0.035	3	22.955	0.059	3	22.236	0.023	3

Table 9 Average Cq values (Cq Mean) and associated standard deviation (SD) calculated from technical replicates (n) of 9 ovotestis samples for six endogenous control genes. Including sample ID, genotype (Geno) and tissue description. *amplification observed in negative controls

Raw Cq data			*Lacads*			*Lef1a*			*Lhis2a**			*Lrpl14**			*Lube2*			*Lywhaz**		
ID	**Geno**	**Tissue**	**Cq Mean**	**Cq SD**	**n**	**Cq Mean**	**Cq SD**	**n**	**Cq Mean**	**Cq SD**	**n**	**Cq Mean**	**Cq SD**	**n**	**Cq Mean**	**Cq SD**	**n**	**Cq Mean**	**Cq SD**	**n**
10633	*DD*	Ovotestis	23.422	0.072	3	30.233	0.400	3	21.929	0.282	3	18.914	0.060	3	22.924	0.220	3	20.008	0.055	3
10636	*DD*		23.613	0.043	3	30.285	0.175	3	21.762	0.103	3	19.229	0.075	3	23.139	0.246	3	19.947	0.015	3
10638	*DD*		24.060	0.029	3	31.212	0.256	3	22.477	0.024	3	19.550	0.073	3	23.945	0.168	3	20.971	0.030	3
10622	*Dd*		22.699	0.060	3	30.869	0.341	3	21.822	0.014	3	19.094	0.194	3	22.578	0.100	3	19.391	0.031	3
10629	*Dd*		23.421	0.042	3	31.568	0.301	3	22.451	0.016	3	19.252	0.058	3	22.893	0.060	3	20.174	0.027	3
10639	*Dd*		24.107	0.039	3	30.927	0.254	3	21.763	0.011	3	19.415	0.009	3	23.367	0.061	3	19.673	0.024	3
10626	*dd*		22.598	0.034	3	30.756	0.144	3	21.729	0.143	3	18.899	0.070	3	22.582	0.298	3	19.386	0.020	3
10630	*dd*		24.505	0.030	3	31.694	0.457	3	22.591	0.066	3	19.381	0.045	3	23.333	0.024	3	20.412	0.046	3
10642	*dd*		23.953	0.046	3	30.971	0.024	2	22.358	0.125	3	19.128	0.038	3	23.110	0.062	3	19.854	0.051	3

Table 10 Linearised Cq values for each tissue analysis. Including sample ID, genotype (Geno) and tissue description.

Linearised Cq values								
ID	**Geno**	**Tissue**	***Lacads***	***Lef1a***	***Lhis2a***	***Lrpl14***	***Lube2***	***Lywhaz***
11289	*DD*	Embryo	0.629	0.496	0.664	0.627	0.571	0.890
11292	*DD*		0.421	0.316	0.552	0.520	0.385	0.487
11293	*DD*		0.497	0.378	0.592	0.482	0.495	0.685
11295	*DD*		1.000	1.000	0.929	0.977	0.722	1.000
11297	*DD*		0.505	0.431	0.560	0.460	0.456	0.619
11298	*DD*		0.302	0.218	0.351	0.279	0.277	0.371
11282	*dd*		0.425	0.357	0.538	0.550	0.382	0.423
11283	*dd*		0.754	0.751	1.000	0.689	0.730	0.917
11284	*dd*		0.306	0.274	0.440	0.340	0.325	0.471
11287	*dd*		0.729	0.973	0.909	1.000	1.000	0.780
11301	*dd*		0.845	0.714	0.840	0.795	0.774	0.980
11303	*dd*		0.433	0.762	0.759	0.711	0.726	0.539
11347	*DD*	Foot	1.000	1.000	1.000	1.000	1.000	1.000
11350	*DD*		0.139	0.086	0.055	0.167	0.067	0.091
11351	*DD*		0.415	0.254	0.252	0.319	0.289	0.332
11352	*DD*		0.235	0.189	0.093	0.190	0.100	0.138
11357	*DD*		0.170	0.079	0.065	0.197	0.075	0.107
11348	*dd*		0.390	0.151	0.224	0.267	0.169	0.241
11349	*dd*		0.391	0.157	0.195	0.261	0.164	0.237
11353	*dd*		0.338	0.137	0.246	0.288	0.172	0.237
11354	*dd*		0.208	0.070	0.111	0.238	0.077	0.134
11356	*dd*		0.295	0.099	0.211	0.352	0.210	0.251
10633	*DD*	Ovotestis	0.495	0.335	0.644	0.664	0.513	0.534
10636	*DD*		0.561	0.348	0.577	0.813	0.590	0.513
10638	*DD*		0.749	0.697	0.927	1.000	1.000	1.000
10622	*Dd*		0.310	0.539	0.600	0.745	0.409	0.357
10629	*Dd*		0.495	0.910	0.912	0.825	0.503	0.595
10639	*Dd*		0.772	0.563	0.577	0.917	0.686	0.430
10626	*dd*		0.291	0.495	0.564	0.657	0.410	0.356
10630	*dd*		1.000	1.000	1.000	0.897	0.670	0.695
10642	*dd*		0.699	0.582	0.857	0.762	0.579	0.483

geNorm Analysis

geNorm provides a stability value 'M value' for each gene; a lower M value indicates a more stable gene. A graphical output of the most stable genes is then produced, culminating in the most stable pair of genes. Due to the algorithms within geNorm, it can only output the most stable pair of genes, not a single best gene. A second graph is produced, indicating the optimal number of genes to include in the experiment to provide the most stable normalisation. An advised cut-off value of less than 0.15 indicates that that combination of genes will provide a reliable normalisation factor (PrimerDesign 2014). The results are presented in Table 11 and Figure 8 and described below per tissue.

Embryo

geNorm placed *Lhis2a* and *Lube2* as the most stable pair of genes, with a combined stability score of 0.196. The inclusion of any number of the genes provided a V score of <0.15, although the lowest V score was achieved with the inclusion of the five genes; *Lhis2a*, *Lube2*, *Lrpl14*, *Lacads* and *Lywhaz.*

Table 11 geNorm results per tissue, including the number of samples included in analysis (n). Endogenous control genes (Target) are ranked in order of decreasing stability 1-6, based upon their stability score. Calculated V scores and individual M scores are also provided.

geNorm Results					
Tissue	**Ranking**	**Target**	**Stability score**	**V score**	**M score**
Embryo, n=12	1/2	*Lhis2a/Lube2*	0.196	0.061	0.246/0.259
	3	*Lrpl14*	0.204	0.064	0.267
	4	*Lacads*	0.242	0.052	0.282
	5	*Lywhaz*	0.262	0.049	0.324
	6	*Lef1a*	0.285	n/a	0.330
Foot, n=10	1/2	*Lywhaz/Lube2*	0.217	0.092	0.325/0.401
	3	*Lhis2a*	0.269	0.091	0.461
	4	*Lacads*	0.327	0.082	0.407
	5	*Lrpl14*	0.376	0.088	0.489
	6	*Lef1a*	0.444	n/a	0.579
Ovotestis, n=9	1/2	*Lrpl14/Lube2*	0.250	0.097	0.367/0.363
	3	*Lywhaz*	0.292	0.070	0.384
	4	*Lhis2a*	0.309	0.079	0.360
	5	*Lacads*	0.360	0.077	0.473
	6	*Lef1a*	0.409	n/a	0.507

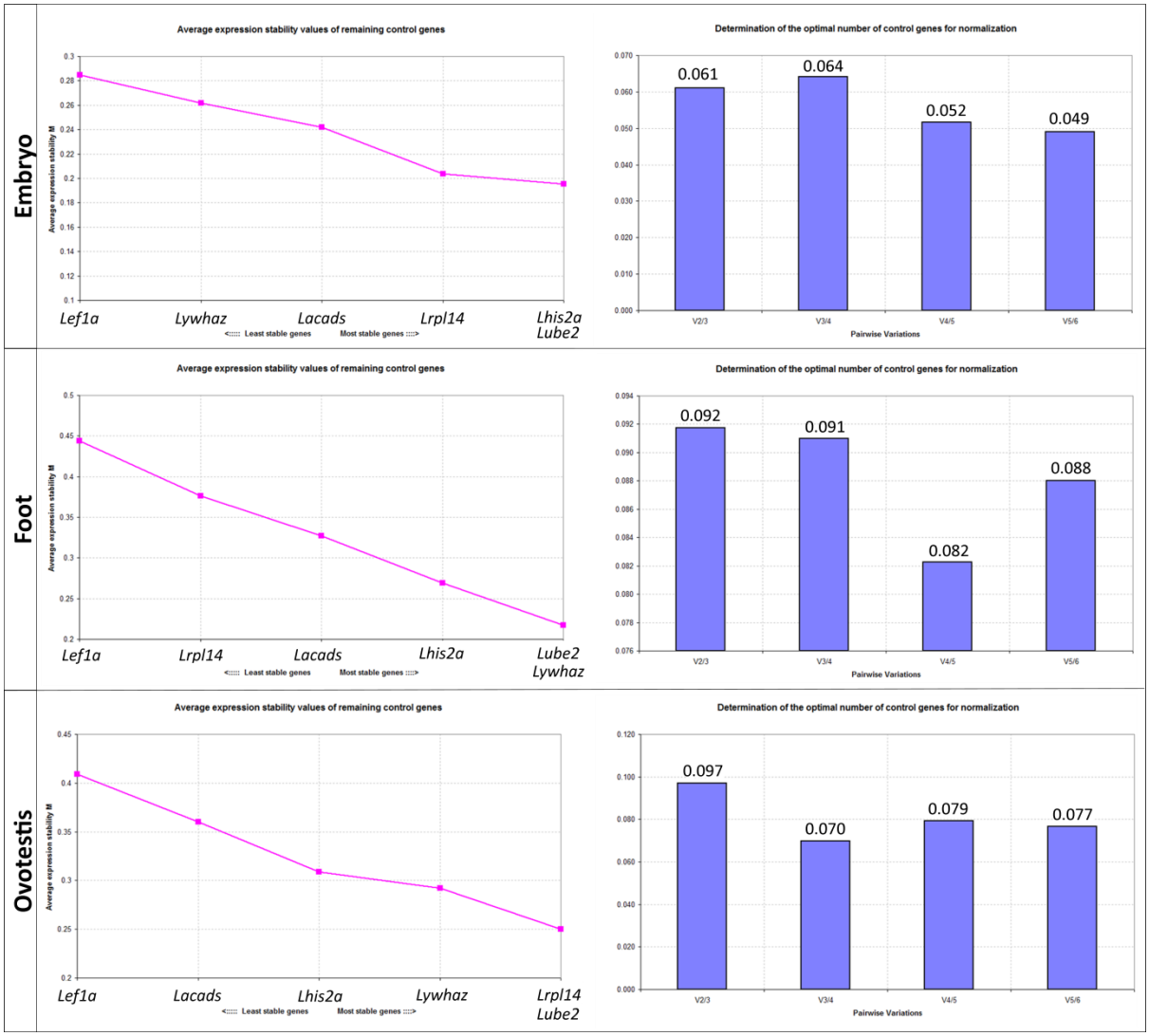


Figure 8 Graphical output of the geNorm analysis. Each tissue analysis generates one graph displaying the average expression stability values of remaining control genes (left) and one graph showing the optimum number of control genes for normalisation (right).

Foot

geNorm placed *Lywhaz* and *Lube2* as the most stable pair of genes with a combined stability score of 0.217. The inclusion of any number of the genes provided a V score of <0.15, although the lowest V score was achieved with the inclusion of the four genes; *Lywhaz*, *Lube2*, *Lhis2a* and *Lacads*.

Ovotestis

geNorm placed *Lrpl14* and *Lube2* as the most stable pair of control genes with a combined score of 0.250. *Lhis2a* bore the lowest M score of all the target genes, at 0.360, yet it was placed fourth in the combined stability score. The inclusion of any number of the genes provided a V score of <0.15, although the lowest V score was achieved with the inclusion of the three genes; *Lrpl14*, *Lube2* and *Lywhaz*.

Of all the tissues, the embryo analyses yielded the lowest V scores, followed by ovotestis and then foot. *Lef1a* was consistently found to be the least stable gene in all tissues.

NormFinder

NormFinder outputs an estimation of both the most stable pair of genes and the single most stable gene. It also has the capacity to incorporate a grouping factor, in this instance genotype. A 'stability score' is calculated for each gene and a combined stability score is output for the best pair of genes. A low stability score indicates better expression stability. The results are presented in Table 12 and described below per tissue.

Embryo:

Lhis2a was identified as the single most stable gene in the embryo analysis with a stability value of 0.058. However, it was not included in the best combined pair, which comprised *Lacads* and *Lube2*, with a score of 0.047.

Foot:

Lywhaz was identified as the most stable gene in the foot tissue, with a stability score of 0.074. When combined with *Lube2*, this was reduced to 0.066.

Ovotestis:

Lhis2a was identified as the single most stable gene in the ovotestis tissue, generating an individual stability score of 0.124. However, the best pair of genes was calculated to be *Lef1a* and *Lywhaz*, with a combined stability value of 0.083, despite the fact that *Lef1a* presented the worst single gene stability value: 0.243.

Across all tissues, the single cell embryo analysis yielded the lowest stability values, followed by foot tissue and ovotestis. With the exception of the embryo analysis, in which it ranked second least stable, *Lef1a* was found to be the least stable gene in all tissues. In all analyses, the stability value of the best combined pair of genes was lower than that of any individual gene stability score.

Table 12 NormFinder results per tissue, including the number of samples included in analysis (n). Endogenous control genes (Target) are ranked in order of decreasing stability 1-6, based upon their individual stability score. The best combined pair of genes is also presented with its associated stability score.

NormFinder Results					
Tissue	**Ranking**	**Target**	**Stability score**	**Best Combined Pair**	**Stability Value**
Embryo, n=12	1	*Lhis2a*	0.058	*Lacads/Lube2*	0.047
	2	*Lrpl14*	0.076		
	3	*Lube2*	0.086		
	4	*Lacads*	0.104		
	5	*Lef1a*	0.122		
	6	*Lywhaz*	0.124		
Foot, n=10	1	*Lywhaz*	0.074	*Lube2/Lywhaz*	0.066
	2	*Lube2*	0.133		
	3	*Lacads*	0.151		
	4	*Lrpl14*	0.176		
	5	*Lhis2a*	0.215		
	6	*Lef1a*	0.298		
Ovotestis, n=9	1	*Lhis2a*	0.124	*Lef1a/Lywhaz*	0.083
	2	*Lrpl14*	0.147		
	3	*Lube2*	0.153		
	4	*Lywhaz*	0.171		
	5	*Lacads*	0.206		
	6	*Lef1a*	0.243		

BestKeeper

BestKeeper firstly calculated the SD values associated with the geometric mean Cq of all samples within each gene. Those genes with a SD of <1 are considered stable and their Cq data is included in the generation of the 'BestKeeper index' (BK index). All of the endogenous control genes assessed exhibited a SD less than 1 and therefore all contributed to the generation of the BK index. A regression model was then fitted to estimate the correlation (r value) of the Cq data of each gene to the BK index. Consequently, the BestKeeper applet provides two measures of gene stability. A low SD and a high r value indicate a more stable control gene. The results of the BestKeeper analysis are presented in Table 13 and described below per tissue.

Embryo

The gene ranked as most stable in the embryo tissue according to SD was *Lhis2a* (0.408), whereas the least stable gene was *Lef1a* (0.577). Every gene in the single cell embryo analysis resulted in a highly significant positive correlation with the BK index (p = 0.001). *Lhis2a* demonstrated the highest correlation with the BK index, with an r value of 0.979, and *Lywhaz* the lowest with an r value of 0.900.

Foot

The gene ranked as most stable in the foot tissue according to SD was *Lrpl14* (0.500), whereas the least stable gene was *Lhis2a* (0.947). Each gene resulted in a highly significant positive correlation with the BK index (p = 0.001). *Lywhaz* had the strongest correlation bearing an r value of 0.998, and *Lef1a* the lowest correlation, with an r value of 0.907.

Ovotestis

The gene ranked as most stable in the ovotestis tissue according to SD was *Lrpl14* (0.176), whereas the least stable gene was *Lacads* (0.500). With the exception of *Lef1a*, every gene resulted in a significant correlation with the BK index. *Lywhaz* had the strongest correlation, with an r value of 0.894. *Lef1a* did show a positive correlation with the BK index generating an r value of 0.655; however this correlation only resulted in a p value of 0.056 and is therefore not considered statistically significant.

Table 13 BestKeeper results per tissue, including the number of samples included in analysis (n). Endogenous control genes (Target) are ranked in order of decreasing stability 1-6, based upon either their correlation with the BestKeeper index (r) or the standard deviation (SD) associated with the average Cq per gene (Mean Cq). Included also are the associated probability values (p) of the correlation.

BestKeeper Results							
Tissue	**Ranking, r**	**Ranking, SD**	**Target**	**r**	**p**	**Mean Cq**	**SD**
Embryo, n=12	1	1	*Lhis2a*	0.979	0.001	22.122	0.408
	2	6	*Lef1a*	0.969	0.001	26.859	0.577
	3	5	*Lube2*	0.962	0.001	25.811	0.514
	4	3	*Lrpl14*	0.957	0.001	22.985	0.476
	5	4	*Lacads*	0.949	0.001	22.913	0.493
	6	2	*Lywhaz*	0.900	0.001	24.257	0.457
Foot, n=10	1	4	*Lywhaz*	0.998	0.001	21.973	0.754
	2	5	*Lube2*	0.993	0.001	22.512	0.867
	3	6	*Lhis2a*	0.984	0.001	23.207	0.947
	4	2	*Lacads*	0.981	0.001	27.082	0.638
	5	1	*Lrpl14*	0.964	0.001	19.526	0.500
	6	3	*Lef1a*	0.907	0.001	33.356	0.695
Ovotestis, n=9	1	4	*Lywhaz*	0.894	0.001	19.974	0.366
	2	2	*Lube2*	0.877	0.002	23.093	0.313
	3	6	*Lacads*	0.876	0.002	23.590	0.500
	4	1	*Lrpl14*	0.853	0.003	19.206	0.176
	5	3	*Lhis2a*	0.831	0.005	22.095	0.330
	6	5	*Lef1a*	0.655	0.056	30.942	0.369

Table 14 Ranking summary of endogenous control gene (Target) stability decreasing from 1-6 as estimated through geNorm (GN), NormFinder (NF) & BestKeeper according to correlation with the BestKeeper index (BK, r) and the standard deviation (BK, SD). Genes included in the 'best-combined pair' within NormFinder are indicated with *.

Target	**Embryo**				**Foot**				**Ovotestis**			
	GN	**NF**	**BK, r**	**BK, SD**	**GN**	**NF**	**BK, r**	**BK, SD**	**GN**	**NF**	**BK, r**	**BK, SD**
Lacads	4	4*	5	4	4	3	4	2	5	5	3	6
Lef1a	6	5	2	6	6	6	6	3	6	6*	6	5
Lhis2a	1/2	1	1	1	3	5	3	6	4	1	5	3
Lrpl14	3	2	4	3	5	4	5	1	1/2	2	4	1
Lube2	1/2	3*	3	5	1/2	2*	2	5	1/2	3	2	2
Lywhaz	5	6	6	2	1/2	1*	1	4	3	4*	1	4

Discussion

Indication of best genes to use as endogenous controls

Embryo

geNorm listed *Lhis2a* and *Lube2* and the two most stable genes in the embryo tissue. Interestingly the addition of a third gene decreased stability, yet the optimal stability value is gained from including all of the genes, except *Lef1a* (Table 11). This is not necessarily because the inclusion of *Lef1a* will significantly reduce the stability, but a consequence of how geNorm calculates the V score by removing the least stable gene, followed by the next least stable and so on. Just as you can only ever have an estimation of the two most stable genes, you can only have a V score of all genes minus one. All genes exhibited good M values. Such that the highest M value observed in the embryo analysis (*Lef1a*, M=0.330), was still lower than the lowest M value in the ovotestis analysis (*Lhis2a*, M=0.360) and comparable to that in the foot analysis (*Lywhaz*, M=0.325).

NormFinder ranked *Lhis2a* as the most stable gene in the single cell embryo, with a stability score of 0.058. However, the combined best pair was comprised of *Lube2* and *Lacads*, which individually ranked 3rd and 4th most stable respectively (Table 12). This discrepancy reflects the variation in stability assessment when utilising the full capabilities of NormFinder, including the group identifiers to incorporate inter and intra group effects, compared to only using the individual stability values.

BestKeeper found highly significant (p=0.001) positive correlations with all genes and the BK index and as such any choice of gene is considered acceptable for use as a reference. *Lhis2a* was ranked as most stable using both the correlation with the BestKeeper index and the SD (Table 13). The BestKeeper rankings display larger discrepancies between the stability rankings based upon r value or SD than comparisons across software. It is important to acknowledge both of the measures of gene stability. Nevertheless, if using more than one endogenous control gene (which is strongly recommended), and given that all of the endogenous control genes assessed have an acceptable range of SD to be included in the BestKeeper index, the correlation of patterns of gene expression, would likely provide a more informative measure of expression variation, as opposed to the level of perhaps negligible SD within a single gene.

All three algorithms ranked *Lhis2a* as the single most stable single gene. However, there is less of a consensus for the rankings of the remaining endogenous controls (Table 14). Generally, *Lhis2a, Lrpl14* and *Lube2* remain in the top three most stable genes across software.

Foot

The results of the geNorm analysis show that even the minimum inclusion of the two most stable genes, *Lywhaz* and *Lube2* provide an acceptable endogenous control measure (V score <0.15). Indeed, the addition of 5/6 of the genes would still provide a stable endogenous control measure (Table 11). This indicates that all of the genes could provide stable endogenous control genes in the foot tissue, with the possible exclusion of *Lef1a* because, for reasons described previously, it was not included in the calculation of the V score.

The optimum number of genes to include according to geNorm is four, however due to finite amount of sample and the cost associated with running qPCR plates, the proposed increase in stability must be counterbalanced with realistic laboratory practice. Three is often quoted as the minimum number of controls to use (Vandesompele, De Preter et al. 2002) although here, the addition of a third control gene provides little increase in stability, V score from 0.92 to 0.91.

NormFinder and BestKeeper (r value) also found *Lywhaz* and *Lube2* to be the most stable pair of genes in the analysis (Table 14). It is of note again that the BestKeeper rankings according to SD are substantially different to those for r value, but for the reasons described previously, the r values are assumed to be more informative in this instance.

Due to the agreement of the different algorithms it is assumed that *Lywhaz* and *Lube2* should be used as endogenous normalising controls for analyses in foot tissue. It will depend on the individual experiment whether or not the small increase in stability gained from the addition of a third control gene, as calculated by geNorm, will be worth the additional time and resources in accommodating another endogenous control gene into the experimental design.

Ovotestis

The results of the geNorm analysis in the ovotestis show that even the minimum inclusion of the two most stable genes, *Lrpl14* and *Lube2,* provide an acceptable endogenous control measure with a V score of 0.097. The inclusion of a third gene substantially improves this score to 0.070, and also indicates the most stable combination of genes (Table 11).

It is interesting however, that the gene with the lowest M value, *Lhis2a*, was placed fourth in the combination of genes to use together. This is inferred as a reflection of how stable *Lhis2a* is when calculated based on its own variability, compared to how it correlates with other genes.

The NormFinder results of the ovotestis tissue were also a little conflicting with regards to the individual gene stability score and that of the best combined pair. *Lhis2a* was clearly ranked as the

most stable gene, with a stability score of 0.124 and *Lef1a* as the least stable with a score of 0.243, almost double that of *Lhis2a*. However, the top combined genes were revealed to be *Lef1a* and *Lywhaz*, which were ranked the 6th and 4th most stable genes, respectively.

These discrepancies are believed to reflect the difference in calculation of expression variability when measured within a single gene and when utilising the full capabilities of the NormFinder software. Although useful insights can be gained by incorporating the inter and intra-group effects, it seems improbable that *Lef1a* truly represents one of the most stable genes in the ovotestis. This is largely due the high average Cq values (Table 9), which are likely to result more variable data (AppliedBiosystems 2014) and additionally due to its poor performance within the other software analyses.

BestKeeper ranked *Lywhaz*, followed by *Lube2* as the most stable genes based on r value (Table 13). The top two genes according to BestKeeper's calculation of SD were *Lrpl14* and *Lube2*, the same top two genes as calculated by geNorm. *Lhis2a* however was ranked in 5th place according to r value and 3rd place according to SD, contrary to its individual top ranking in geNorm and NormFinder (Table 14).

It appears that *Lhis2a*, if used alone may provide a more stable reference gene than any other, however when using, as recommended, more than one endogenous control gene, *Lhis2a* fails to be the most stable choice.

Comparisons across tissue analyses

Lube2 indicates overall most stable endogenous control

All of the six genes presented acceptable endogenous controls for use in all of the tissues assessed with the potential exception of *Lef1a*. Across all of the tissues, *Lef1a* and *Lacads* were generally ranked the least stable. *Lube2* represented one of the two most stable genes in the geNorm analysis of all three tissues (Table 14). Therefore, if the same endogenous control genes must be used for all tissue analyses, *Lube2* should be included.

Lef1a represents the least stable endogenous control

Lef1a was consistently ranked least stable in all analyses of foot and ovotestis tissue, and often in the embryo. It is important to note that it is still considered acceptable for use within an experiment as a stable endogenous control, however would not be selected as the gene of choice.

The reason for the poor performance of *Lef1a* may be due to its low level of expression rather than any true expression variability of this gene. The average Cq values for *Lef1a* in the foot and ovotestis analyses are all above 30 (Table 8, Table 9). This is above the Cq for optimal qPCR analysis and may result in greater variability during the qPCR (AppliedBiosystems 2008). Additionally, *Lef1a* exhibited the amplification efficiency furthest from 100%/2 and the only gene to show efficiency >100% which is indicative of inhibition (Table 6) which may lead to skewed results (AppliedBiosystems 2014). Furthermore, it is listed in the BestKeeper requirements that genes with values over 30 are not suitable for analysis in the current version of BestKeeper and that the issues will be addressed in another version awaiting release (Pfaffl, Tichopad et al. 2004). Of the analyses described here, the only instance where a gene has not been considered stable for use as an endogenous control gene, was that of *Lef1a* in the BestKeeper analysis of the ovotestis tissue, however with an average Cq of over 30, this gene data was technically not suitable for analysis within BestKeeper and therefore should be inferred with caution. The same applies for the foot analysis within BestKeeper in which the *Lef1a* Cq values also exceeded 30 (Table 13).

The lower expression level of *Lef1a* was indicated in the amplification efficiency experiments, which showed a higher minimum working concentration for *Lef1a* compared to the other genes (Table 6). *Lef1a* may provide a reliable endogenous control gene when using an increased cDNA concentration.

The ovotestis tissue was found to be the most variable

Compared to the other tissues assessed, the ovotestis analysis was generally shown to be more variable. There are many reasons why some tissues may be more variable than others. The only systematic difference between the populations of *Lymnaea within* each tissue analysis is their chiral genotype. It may be that there are greater chirality-associated variations in expression of the endogenous control genes in the ovotestis, than compared to the embryo and foot tissue. However, the sampling method of the ovotestis tissue holds greater potential for inconsistency than that of the embryo and foot tissue, which may better account for the increased variability.

Extraction of the individual snails' ovotestis was not rigorously temporally controlled. RNA extractions were performed in the morning and each snail was observed to be sexually mature (having lain at least one clutch of eggs in its lifespan). Ideally the ovotestis would have been extracted shortly before egg laying, resulting in an ovotestis sample containing both sperm and eggs. However, this is close to impossible, as the snails lay eggs at varied intervals, which are largely unpredictable and because the ovotestis is inside the snail, it is impossible to view to condition of the tissue prior to dissection.

In addition to the temporal variation in the ovotestis tissue samples, there is the inevitable contamination of the 'liver' tissue, (this organ represents an overall digestive gland however is commonly referred to as the liver). Due to the internal organisation of *L. stagnalis*, the ovotestis is closely associated with the liver (Figure 9). When dissecting, the ovotestis can be pulled up out of the liver and is generally free of the dark brown liver tissue, which is clearly distinguishable from the cream coloured ovotestis. However, liver tissue carryover is inevitable, albeit reduced and each individual ovotestis sample will have a varying amount of liver contamination.

The foot tissue was also not temporally controlled beyond morning extractions and that the individual snails were of a similar age. The increased overall consistency of the endogenous control genes in the foot tissue compared to that of the ovotestis is perhaps an indication of the general stability of the foot tissue expression compared to the ovotestis tissue, which contains potentially more variable factors i.e. time since egg-laying. Additionally, the foot tissue does not have the same issues with carryover of non-specific tissues as does the ovotestis, although care must be taken not to include mouth parts or brain tissue.

It is clear however that the embryo samples exhibit much less variation than both of the other tissues. The embryo samples represent a very clean and temporally controlled sample, with RNA storage occurring almost immediately after decapsulation of the egg within hours of it being laid. Additionally, no new transcripts are being generated within the single cell embryo, which would further reduce variability.

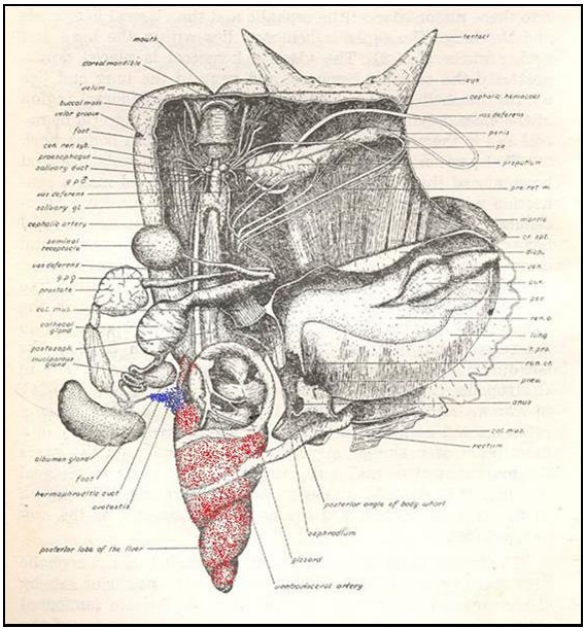


Figure 9 Internal organisation of *Lymnaea stagnalis*. The liver is highlighted in red and the ovotestis highlighted in blue. Adapted from original image from Wisconsin Academy of Science, Arts and Letters.

Choice of endogenous controls for between tissue comparisons

It must be noted that if making comparisons between tissue analyses, the tissues were firstly extracted by different methods resulting in potential variation in sample quality and especially genomic contamination. The snails used in the ovotestis analysis were also of an earlier lab population than those used in the embryo and foot tissue analysis.

The gene stability assessments here indicate the most suitable endogenous control gene to normalise expression patterns between genotype within the same tissue. No controlled stability assessments have been compared across the different tissues. However due to the general agreement of the stability of *Lube2* and *Lhis2a* in all tissues, these likely represent stable genes across all three tissues (Table 14).

Quality controls

Sample quality

It must be acknowledged that the quality of each total RNA sample was not been exhaustively assessed prior to cDNA synthesis. When representative samples were visualised via gel electrophoresis, the total RNA showed distinct bands. These are assumed to represent specific-sized abundant rRNA and the general lack of smearing indicated that the transcripts had maintained their full length and ultimately that the samples were of good quality (presented in SI, S3). Nevertheless, not all samples were visualised on a gel, as such there may have been differences in sample quality that were not detected in the Nanodrop quantification, for example degradation of long transcripts. However, no sample included in the analyses indicated a tendency for error, due to each instance of omitted replicates having occurred in different individuals.

Additionally, there was no mRNA enrichment performed on any of the samples. This may have reduced the quality of the samples through interference of dominating rRNA transcripts during qPCR. Yet this is not considered to be of concern due to the success of the amplification efficiency tests.

Of greater concern, may be the genomic DNA present in the foot and ovotestis tissue samples. Due to the lack of multiple Tm peaks in any of the experimental samples, there is no evidence that any confounding genomic transcripts have been amplified. If there has been amplification of genomic transcripts it was at a level undetectable by the qPCR instrument and therefore considered negligible. Carryover genomic DNA if not able to amplify, may still negatively impact the qPCR reaction through interference. The level of interference would vary with the extent of genomic

carryover. However, again due to the success of the amplification efficiency tests, which were performed using the ovotestis sample, which demonstrated the highest level of genomic carryover, this is not considered to be of concern. The difference in performance of the DNase treatments is still of technical interest however, and so a discussion of the protocols follows in the SI (S5.4).

Primer quality

The presence of Tm peaks in the negative controls of genes; *Lhis2a*, *Lrpl14*, *Lube2*, and *Lywhaz* may lead to some concern over the specificity of the primers used. Due to these peaks being of significantly lower Tm and generally lower signal intensity, the peaks are assumed to represent primer dimer created from the primer pairs binding to each other as opposed to the target sequence (Figure 7).

It is also possible that the multiple Tm peaks were created from a single amplicon. As described in the introduction, the Tm curve is generated through the reduction in the fluorescent signal as the double-stranded amplicon melts and the SYBR® green dye dissociates. Therefore if a region of the amplicon melted slower than another (GC rich regions for example) then it is possible that two peaks could be created (Downey 2014).The peaks observed in this experiment however, do not occur in the experimental samples, which supports the assumption that they represent primer dimer, which is more likely to occur in the absence of a target sequence.

It is generally advised to redesign any primer pair which generates primer dimer because the primer dimers will still be formed within the experimental samples (AppliedBiosystems 2014). However, the lack of multiple Tm peaks in the experimental samples indicates that this too represents a negligible limitation on the accuracy of the qPCR and should not create a systematic bias between the genotypes compared.

Experimental design

Number of samples

The minimum recommended number of samples to include in a geNorm experiment is ten, these should also represent all experimental conditions (PrimerDesign 2014). As the establishment of endogenous controls is ideally the first step in a relative qPCR experiment it is usually performed on the preliminary samples and minimum experimental setup.

The embryo experiment was initially planned to compare only the two homozygous chiral genotypes and included six samples of each. Heterozygote samples were later added to the qPCR experiments. Therefore, the endogenous controls have not strictly been assessed in all genotypes within the

embryo. Due to the limited variability observed in endogenous controls within the embryo tissue, it is unlikely that the heterozygotes would behave differently.

Only the homozygous genotypes were selected for comparison in the foot tissue, which included five samples of each. All ten of the foot samples used in the qPCR experiments were used in the endogenous control assessment.

Initially twelve ovotestis samples were prepared for use in the endogenous control gene assessment, four *DD*, four *Dd* and four *dd*. However, after an initial qPCR run it was quickly apparent that one of the samples, namely 10631; a *Dd* sample, frequently generated high SD and a failure to amplify. This sample was removed from analyses. However, due to the comparative nature of the stability assessment, it is recommended to include equal numbers of each experimental condition/genotype (PrimerDesign 2014) and so another *DD* and *dd* sample were also removed from the experimental setup. Therefore, the ovotestis endogenous control assessment, with nine samples, falls just under the advised minimum input, although due to the overall good levels of stability in the endogenous control genes this is not believed to significantly compromise the findings.

This experiment was originally designed to compare expression differences between chiral genotypes, although it became apparent that tissue comparisons would also be possible and potentially very informative. However, an additional gene expression stability experiment was not factored into the experimental design and as such there is no stability information for endogenous control genes across tissues.

Number of genes

The number of genes to include in the search to find multiple stable endogenous control genes largely depends on resources. The more genes tested, the higher the chance of finding the most suitable endogenous controls. However, qPCR experiments are costly and especially when the RNA sample is limited, as was the case with the embryo samples, it is not advisable to test more genes than necessary. The company Primer Design produce kits for establishing endogenous controls with the geNorm software (PrimerDesign 2014). The kits are provided with either six or twelve candidate endogenous controls to test. It was therefore assumed that six candidate endogenous control genes would be a sufficient starting point to identify stable control genes.

Choice of reporter dye

SYBR® green was chosen as the reporter dye for the experiments due to its flexibility of use, which can be considered a benefit or a limitation. Alternatives such as TaqMan® require specific

fluorescent primers to be made for each target gene. This method greatly increases specificity of the quantitative data since the fluorescent signal will only be omitted from the specific target sequence (as opposed to any double stranded product, as is the case with SYBR® green). However, each TaqMan® probe is considerably more expensive than the standard oligo which can be used with SYBR® green. The use of SYBR® green additionally allows for the Tm curve to be produced, providing a valuable indication of the specificity of every qPCR reaction, which is not possible with the TaqMan® probe.

Capabilities of software

Of the three methods chosen here, geNorm and BestKeeper both function through pairwise comparisons, whereas NormFinder is model based, calculating variability resulting from the experimental grouping factors. Each method provides a unique aspect of analysis.

The geNorm applet calculates gene stability by performing pairwise comparisons between all genes included and provides an 'M value' based upon the geometric mean of the SD of each pairwise comparison per gene, therefore a low M value indicates lower variability/greater stability. The software then follows to progressively omit the most variable gene pairs until the most stable pair of genes remains. As a result of this method, any genes exhibiting a similar expression pattern will be considered more stable. Therefore, it is of great importance to ensure that the any potential control genes are not co-regulated. None of the genes included here are believed to be co-regulated based on their largely unrelated functions, however this does not completely eliminate the possibility of co-regulation.

The step-wise process of the stability assessment of the endogenous control genes performed in geNorm additionally allows for the calculation of the optimum number of genes to include, whether this results from the most stable pair of genes or the inclusion of more. However, as a result of the step wise omission of the least stable gene, there is only ever a calculation for all genes minus one, therefore the most stable combination may be generated from the inclusion of all genes, yet this is not calculated.

For geNorm, or indeed any pairwise comparison approach, to work it must have a stable gene included in the analysis. NormFinder however, estimates gene stability using a 'model-based approach' which evaluates the level of intra- and inter-group variation within each gene. This is believed to provide a more robust estimation as the genes are classified by the level of systematic error as opposed to their similarity of expression pattern to the other genes included in the analysis.

As such NormFinder is less sensitive to the risk of co-regulated genes and accommodates the grouping factors of the experimental analysis.

BestKeeper provides two measures of gene stability. The first simply provides a measure of the SD of the average Cq values per gene. Any gene with an SD of less than 1 is considered stable and is included in the calculation of the BestKeeper index. A Pearson correlation coefficient (r) is then performed providing probability (p) values to evaluate the relatedness of each gene's expression pattern with the BestKeeper index. Genes can then be ranked according to the strength of their correlation and associated p-values. This provides a measure of how similar one gene's expression pattern is compared to that of the other genes included and thus works in a similar way to geNorm. Therefore, it is again of great importance to ensure no genes are co-regulated.

A number of previous studies citing the use of BestKeeper have simply ranked to genes using the SD score (Hibbeler, Scharsack et al. 2008, Bouhaddioui, Provost et al. 2014). This method omits the pairwise comparison element of the BestKeeper applets function, and simply classifies genes based upon their independent variability. This does provide a useful measure of how the gene performs and is not biased by the relationship of the other genes included, however this should be combined with the inferences based upon the similarities of gene expression, especially if the study intends to use multiple control genes (Taki, Abdel-Rahman et al. 2014).

Some more comprehensive analyses of endogenous control genes have included some approaches that were not used here (Jacob, Guertler et al. 2013, Taki, Abdel-Rahman et al. 2014), namely the delta C_T method (Silver, Best et al. 2006) and RefFinder (Xie, Xiao et al. 2012). The delta C_T method also performs pairwise comparisons and as such was not believed to add a substantial amount of additional information to warrant the additional analysis. RefFinder provides the useful capability of combining the results of multiple methods, and calculating a geometric mean of the multiple rankings (Xie, Xiao et al. 2012). RefFinder was not employed here, partly due to difficulty in accessing the web-based program at the cited address, yet largely due to doubt surrounding the appropriateness of combining alternative methods to create a summative ranking. For example, if three pairwise comparison approaches have been performed and only one model based method, the rankings will be biased according to the methodology.

Of the approaches discussed here, geNorm is the most cited at 4,352 times. Studies citing the use of NormFinder are less than half of this number at 1,734 and BestKeeper 1,215 times. Thus geNorm is the most widely accepted as the method of choice for verifying the stability of endogenous control genes. The latest geNorm software is also provided in kits by the qPCR company Primer Design,

providing increased accessibility of this method. The paper introducing the delta/comparative C_T method for verifying stability of endogenous controls has only been cited 215 times, although because no actual software package is required for this method, it may not be cited in the same manner. RefFinder is the most recent of the methods discussed here (September 2012), which may account for its relatively low 37 citations (All citation counts are quoted from Web of Science™ correct as of the 24th September 2015).

All three methods used here have provided a unique aspect of the data analysis. geNorm provides a measure of the optimum number of genes to include in the analysis and an advised cut-off value (V, <0.15) for an acceptable endogenous control gene combination. BestKeeper outputs a quotable measure of SD for each gene and a statistical measure of the relatedness of gene expression. Finally, NormFinder provides valuable information on the experimental design; calculating variation created both within and between experimental groups and importantly provides an alternative to pairwise comparison methods.

Conclusion

It has been established that any of the six genes would provide acceptable endogenous controls to standardise gene expression between chiral genotypes within any of the three different tissues, perhaps with the exception of *Lef1a*. Once published, these primers will enable other researchers of *L. stagnalis* to quickly verify endogenous controls suitable for use in qPCR experiments assessing ovotestis, foot and embryo tissue within and between chiral variants, which was lacking previously. Additionally, the apparent unsuitability of *Lef1a* is of interest as it is a common choice for endogenous control genes.

The software is largely in agreement that *Lef1a* and *Lacads* are the least stable genes across all tissues. BestKeeper provides no information on the stability of pairings or trios, yet provides a convenient measure of SD for each gene and valuable support to the inferences of the other methods. NormFinder is informative in its independent rankings and ability to incorporate experimental group into the analysis, yet does not provide 'best combined trio'. Due to some of the surprising inclusions of lower ranked genes in the best combined pair, it would be difficult to estimate which, if any, additional genes would increase stability.

If the experiment could only employ one endogenous control gene, which is not advised, however in some circumstances, such as very small starting material, becomes necessary; *Lhis2a* represents the most stable choice of gene for the embryo and ovotestis tissue based on its individual SD. However,

when using more than one endogenous control, other genes provide more stable alternatives. *Lywhaz* largely represents the single most stable gene in the foot tissue. Furthermore, if the experiment could only employ one endogenous control gene across the three tissues, *Lube2* represents a common top ranked gene.

Due to frequent recommendations to use a minimum of three endogenous control genes if possible, the results from the geNorm analysis appear to be the most informative as this software provides information on the stability of more than two genes.

Chapter 3: Quantitative gene expression analysis in *Lymnaea stagnalis*

Introduction

Quantitative gene expression analysis

One of the fundamental aims of this project was to identify expression variation between chiral variants of *L. stagnalis*. A total of thirteen candidate genes, potentially associated with chirality determination in *L. stagnalis* were selected for qPCR analysis. As introduced in the previous chapter, qPCR represents a gold standard technique in quantitative expression analysis, yet common misuse can often lead to misinterpretation of data. The experimental priorities already outlined were employed here to ensure correct experimental practice. Any significant differences in the pattern of gene expression were hoped to elucidate functional processes associated with chiral dimorphism.

Selecting genes of interest

Functional targets

The original project aim was to further analyse a selection of candidate genes identified as differentially expressed (DE) from the eRAD dataset described in Chapter 4. Any conclusions from the high throughput bioinformatic analysis would ideally be supported in an *in situ* experiment, assessing more individuals under better controlled settings. Preliminary DE analysis of the eRAD dataset identified a number of loci as significantly DE between chiral genotypes. A selection of loci representing genes bearing functions likely associated with laterality determination or cytoskeletal processes were selected for further analysis via qPCR. The eRAD analysis which identified the DE loci was obtained through unsuitable parameters and as such is not presented in Chapter 4. However, the target genes selected still hold functional associations with laterality determination and so are still of interest to assess via qPCR. The original eRAD data provided the sequence information required to assess the specific gene targets in *L. stagnalis*.

Actin-related proteins

Cytoskeletal processes have been highlighted in the majority of models of symmetry-breaking and especially those in early development (as introduced in Chapter 1). Cytoskeletal actins have been specifically implicated in LR axis specification in *L. stagnalis* (Shibazaki, Shimizu et al. 2004) and

therefore represent likely targets for expression variation or regulation in chiral variants. Additionally, the primary candidate for the chirality gene identified in the Davison research group is a diaphanous related formin, a Rho GTPase protein which is known to regulate actin assembly (Li and Higgs 2003, Kovar 2006).

Myosins

Similarly, motor proteins, such as myosins, have been highlighted with potential functions in intracellular symmetry breaking, by controlling asymmetric distribution of polarity determinants, (molecular cargo transport) (Vandenberg, Lemire et al. 2013). It is also expected that the myosins will interact with other cytoskeletal processes, especially actin dynamics as they largely represent actin-dependent motor proteins which are often involved in forming actin filaments (Sellers 2000) and have been previously linked with the establishment of LR asymmetry (Baum 2006, Hozumi, Maeda et al. 2006, Speder, Adam et al. 2006).

Additional candidates

A collagen-related target gene was also selected for further analysis representing an alternative extracellular structural component. Additionally, a staufen-related gene was assessed. Staufen is a gene which has been associated with regulation of gene expression and asymmetric mRNA localisation in *Drosophila* embryos (Matsuzaki, Ohshiro et al. 1998, Houchmandzadeh, Wieschaus et al. 2002, Martin and Ephrussi 2009). Finally, a largely uncharacterised gene unc93a was included, which has shown potential phenotypes relating to egg laying (de la Cruz, Levin et al. 2003) and ovarian membranes (Liu, Dodds et al. 2002) in addition to muscle function (Hoebe and Beutler 2008). It is hoped any patterns of gene expression identified in unc93a may contribute to the functional characterisation of this gene.

Proximal targets

As described in Chapter 1, the region of the chromosome which contains the single heritable unit that determines chirality in *L. stagnalis* has been identified by the Davison research group through continued genetic analysis of chiral variants. This involved mating crosses between chiral lines and creating linkage maps to identify genetic markers tightly linked to the chirality phenotype. Due to the lack of a reference genome for *L. stagnalis*, a method called BAC walking was employed to obtain sequence information within this region and identify the genes present (Liu, Davey et al. 2013). This has provided a selection of candidate genes in close proximity to, and in linkage with the chirality locus.

The primary candidate as the causal gene for establishing LR asymmetry in *L. stagnalis* is a diaphanous formin related gene, hereafter referred to as '*Ldia2'*. A single base deletion has been identified in the sinistral copy of the gene, present in the very early coding region (Davison *et al. awaiting publication*). This deletion creates a coding frameshift and will likely result in a number of downstream consequences (Streisin.G, Okada et al. 1966). It is unknown whether this will affect the quantitative levels of the transcript. There is nothing currently known to be inhibiting the transcription of the mRNA and so consequences of the frameshift may only be observable at the protein level. Yet regulatory processes such as nonsense mediated decay may result in quantitative differences between the chiral genotypes (Neu-Yilik, Gehring et al. 2004, Conti and Izaurralde 2005).

There is another diaphanous formin gene, *Ldia1*, in close proximity to the primary candidate gene with a small number of genetic sequence differences, maintaining approximately 90% conserved sequence with *Ldia2* (Davison et al, *awaiting publication*). These two highly conserved genes are likely to have resulted from a previous gene duplication in *L. stagnalis*. The gene lacking the frameshift in the coding region is indicative of being the ancestral form prior to the duplication event, due to its greater sequence similarity to the single gene copy present in the closely related snail species *Biomphalaria* and *Physa* and is therefore referred to as '*Ldia1*'. It will be essential to identify whether there are also expression differences in this gene to ascertain whether it is the frameshift in *Ldia2* or generally an associated function of the diaphanous formins causing any observed pattern of gene expression.

There are a number of other genes within this region of the chromosome which represent alternative candidates for the causal gene of LR determination in *L. stagnalis* and therefore must also be compared alongside the main candidate *Ldia2*. Two additional genes were chosen for qPCR analysis based on both their close proximity to the chirality locus and associated functions.

The first, within the 'fat' group of the cadherins, which have functions in cell adhesions, is a fat1 like gene, hereafter referred to as '*Lfat1*'. It has been suggested that the more divergent cadherins, such as those in the fat group, have a range of more diverse cell functions (Suzuki 2000, Tanoue and Takeichi 2005, Halbleib and Nelson 2006). The fat group has been linked to the actin dynamics in the cytoskeleton, and specifically f-actin (Tanoue and Takeichi 2004), therefore it will be important to investigate gene expression patterns in addition to the primary candidate *Ldia2*.

The second, is a gene involved in maintaining integrity of polarised cellular extensions in morphogenesis, described in *Drosophila*, known as 'furry' here referred to as '*Lfry*' (Cong, Geng et al.

2001). The cellular extensions are composed of cytoskeletal components, and therefore may have important interactions with cytoskeletal dynamics in development.

Selecting tissues for comparison

One of the major benefits of RNA analysis, as introduced in Chapter 1, is the ability to gain insight into the dynamics of gene regulation, revealing which genes are being 'switched on' or overexpressed relative to another individual or tissue. Therefore, it is important to select, not only representative individuals but also appropriate tissues to be used within the gene expression comparison.

Chirality associated differences represent a significant factor during *L. stagnalis* development, exemplified by the low hatch rate observed in sinistral embryos (Davison, Barton et al. 2009), whereas the functional differences between genotypes later in life are apparently negligible, with the exception of some behavioural traits (Davison, Frend et al. 2009). As such it is predicted that chirality-associated differences will be most prevalent during development and reproduction. Furthermore, the establishment of chirality is known to arise from a maternal effect and occur before the third cell cleavage. To identify expression differences associated with the causal gene and not later downstream processes, the ideal tissue would be unfertilised eggs. Accordingly, this experiment examined the ovotestes of sexually mature *L. stagnalis* to provide a representative sample of gametic expression patterns and potentially very early stage zygote.

Single cell embryos pooled from individual snails of known genotype were also included in the experiment. The zygote is believed not to start expressing its own transcripts until the 24 cell stage (Morrill 1982) although zygotic nuclear transcription has been observed from the 8 cell stage (Liu, Davey et al. 2014). Therefore, when assessing the one cell stage embryo tissue, it can be assumed that only the maternal transcripts will be present.

In addition to these functionally related tissues, the foot tissue was included in analyses to provide a somatic control tissue comparison.

Predicted outcomes

Genotype associated patterns of gene expression provide insight into the functional consequences of genetic variation. Furthermore, allele specific expression patterns can reveal regulatory mechanisms effecting only one allele, such as x chromosome silencing and epigenetic gene imprinting, although are not limited to such occurrences (Lo, Wang et al. 2003, Serre, Gurd et al. 2008, Yang, Graze et al. 2011). Due to the high level of genetic similarity of the Davison laboratory

population of *L. stagnalis* (>98%) whilst maintaining chiral dimorphism, it is expected most differences in gene expression will be associated with chirality. Additionally, the inclusion of the heterozygote genotype is hoped to reveal quantitative expression patterns regarding genetic dominance at the chirality locus.

In light of previous studies describing that sinistral developing *L. stagnalis* lack a functional step during spiral cleavage (Shibazaki, Shimizu et al. 2004) and suffer a reduced hatch rate (Davison, Barton et al. 2009), sinistrals may be exhibiting loss of function through reduced or interrupted gene expression in a number of genes.

Specifically, the primary candidate gene, *Ldia2*, has a deletion in the sinistral copy resulting in a frameshift very near the start of the coding sequence (Davison et al, *awaiting publication*). If not observable through quantification of the missense transcript, the resulting protein level changes will likely result in downstream consequences in other genes. For example, diaphanous formin is directly involved in actin polymerisation and self-assembly. Consequently, it is likely that the expression of the arp2/3 complex genes will be effected.

The gene-duplication of the diaphanous formin related gene may function to 'rescue' the faulty *Ldia2*. This would provide an explanation for why the sinistral *L. stagnalis* do not exhibit a complete loss of function, and may even assume overexpression of the *Ldia1* gene to compensate. The foot tissue was included as a somatic tissue control. Due to the limited observations of effects of chirality in adult *L. stagnalis* (Davison, Barton et al. 2009), chirality-associated differences in expression were not expected. However, if the frameshift in the sinistral *Ldia2* gene, results in a transcript monitoring response, such as non-sense mediated mRNA decay, this would be expected to occur in all tissues.

In addition to the genotypic comparisons, differences in gene expression between tissues may provide functional inferences. For example, the transcripts already present in the one-cell embryo direct development until the onset of zygotic transcription, and potentially after (Baroux, Autran et al. 2008, Liu, Davey et al. 2014). Therefore, any transcripts relatively overexpressed in the one-cell embryo compared to the foot will likely have increased functional significance in early development.

The experiments performed here will not provide comprehensive answers to these questions; however quantitative patterns in gene expression can elucidate potential regulatory processes and highlight functional importance of these 13 candidate genes in chiral variants.

Methods

Sample Preparation

Three separate tissues from laboratory reared populations of *L. stagnalis* were assessed in this study. The samples included offspring from multiple specific mating crosses. Each population was generated from the main laboratory population of the Davison research group, although the level of inbreeding is present to different extents. All samples have an inbreeding coefficient of more than 98% similarity whilst maintaining chiral dimorphism.

Embryo

The same embryonic cDNA samples used in the validation of endogenous control genes experiment (Chapter 2) were used for the differential expression analyses described here with the addition of the heterozygote, *Dd*, samples described below.

Five single-cell embryo samples collected from individual *Dd* mothers were added to the embryo analysis. Having observed self-fertilised, anticlockwise shell-coiling mothers to produce clockwise-coiling offspring, the genotype of the mother was known to be *Dd*. Egg collection, RNA extraction and cDNA synthesis protocols followed those described in Chapter 2. Due to the later extraction date, the *Dd* samples are not representative of same single genetic cross as the *DD* and *dd* embryo samples. A pooled *Dd* sample was also generated by pooling single cell embryos from multiple *Dd* mothers prior to RNA extraction. This sample was run as an additional reference sample in the embryo and foot experiments. In total, the embryo dataset comprised of six *DD*, five *Dd* and six *dd* individual samples with an additional reference sample extracted from multiple individuals, referred to as '1 cell pool' (Table 15).

Due to the limited quantity of the embryonic RNA samples, only one round of cDNA synthesis could be performed resulting in a maximum of 12 µl full concentration cDNA. Additionally, the cDNA was synthesised from less than the standardised 500 ng total RNA (Table 15).

Foot

The foot samples used in this experiment are the same as those described in Chapter 3. There are no *Dd* representative genotypes in the foot tissue (Table 16).

Ovotestis

In addition to the nine ovotestis samples used in Chapter 3, three ovotestis samples that were not utilised in the endogenous control analyses, although were generated at the same time and therefore of the same genetic cross, were included (sample ID: 10627, 10631 & 10640). Another

fifteen ovotestis samples were included (sample ID: 8515-9014): these cDNA samples were synthesised from total RNA extracted from the individual snails included in eRAD library 3 & library 4 of the same ID, prior to mRNA enrichment (described in Chapter 4). A final ten samples were added to the ovotestis analysis (sample ID: 11347-11357). These cDNA samples were synthesised from total RNA extracted from the ovotestis of the same individual snails as the foot tissue samples. In summary the ovotestis datasets contained fourteen *DD*, nine *Dd* and fourteen *dd* individual samples. The samples span three different genetic crosses and varying sample storage duration (Table 17).

DNase treatment

As described in Chapter 3, (Methods, Sample Preparation), the embryo samples were extracted using the RNeasy micro kit (Qiagen), which includes a DNase treatment step, DNase I. Total RNA of the ten foot samples was subsequently re-extracted using the RNeasy micro kit and therefore treated with DNase I.

No DNase treatment was performed on the earlier two rounds of ovotestis RNA extractions (sample ID: 8515-9014; 10627-10642). Two alternative DNase treatments were tested on ten of the ovotestis total RNA samples (11347-11357). Firstly, Ambion's DNA Free™ method was used and cDNA was then synthesised from 500 ng of the treated total RNA as per the protocol described in Chapter 2. Having failed to prevent intronic amplification from the cDNA generated (see SI, S5 for further details), another DNase treatment, Primer Design's Precision DNase, was applied to the same RNA sample. cDNA was then synthesised from 500 ng of the treated RNA. Both DNase treatments were applied in accordance with the protocols provided.

As in the previous chapter, at least one standard non-quantitative PCR was performed on all cDNA samples. PCR amplification of the gene of interest (GOI) functioned as a positive control for both the primer pair and the cDNA sample prior to commencing the more expensive qPCR reactions. Another PCR, utilising primers specific to intronic regions, was performed on all samples to test for the presence of contaminating carryover genomic DNA. A consistent genomic DNA sample of an individual *DD L. stagnalis* was used as a positive control and PCR grade water as a negative control in all reactions. The PCR products were visualised via gel electrophoresis using ethidium bromide as a fluorescent marker.

All RNA samples were stored at -80°C and all cDNA samples were stored at -20°C. Aliquots were made of the experimental working concentration dilutions of cDNA to reduce freeze-thaw cycles, whereas serial dilutions were performed independently for each standard curve experiment. All cDNA samples were moderately vortexed before use and prior to each serial dilution step.

Table 15 Details of RNA extraction and cDNA synthesis for the single cell embryo samples used in the qPCR experiments. Table includes: sample identifier (ID) and genotype (Geno) of the mother snail; Spectrophotometry data of the Total RNA sample including sample concentration (ng/µl) and 260/280 & 260/230 absorbance ratios; volume (µl RNA) and quantity (ng RNA) of total RNA used for cDNA synthesis. The individual used as the calibrator sample in the genotype analysis is indicated by 'C'.

ID	Tissue	Geno	Extraction Date	Extraction method	DNase Treatment	Total RNA			cDNA synthesis	
						ng/µl	260/280	260/230	µl RNA	ng RNA
11289	Embryo	*DD*	25/11/2014	RNeasy micro kit	DNase I	13.80	1.80	0.22	10.0	138.0
11292	Embryo	*DD*	07/11/2014	RNeasy micro kit	DNase I	9.90	2.40	0.73	9.3	92.1
11293	Embryo	*DD*	20/11/2014	RNeasy micro kit	DNase I	15.10	2.20	0.41	10.0	151.0
11295[C]	Embryo	*DD*	25/11/2014	RNeasy micro kit	DNase I	10.20	1.74	1.28	10.0	102.0
11297	Embryo	*DD*	07/11/2014	RNeasy micro kit	DNase I	11.60	1.83	0.87	9.4	108.5
11298	Embryo	*DD*	28/10/2014	RNeasy micro kit	DNase I	21.00	1.82	1.05	9.0	189.0
11358	Embryo	*Dd*	20/04/2015	RNeasy micro kit	DNase I	14.40	1.72	0.60	10.0	144.0
11359	Embryo	*Dd*	20/04/2015	RNeasy micro kit	DNase I	13.90	2.40	0.87	9.0	125.1
11360	Embryo	*Dd*	30/04/2015	RNeasy micro kit	DNase I	23.30	1.84	0.63	10.0	233.0
11361	Embryo	*Dd*	30/04/2015	RNeasy micro kit	DNase I	15.80	2.04	0.80	10.0	158.0
11363	Embryo	*Dd*	11/05/2015	RNeasy micro kit	DNase I	18.60	2.56	1.30	10.0	186.0
1 cell pool	Embryo	*Dd*	11/05/2015	RNeasy micro kit	DNase I	12.00	2.57	0.46	10.0	120.0
11282	Embryo	*dd*	03/11/2014	RNeasy micro kit	DNase I	13.70	2.44	0.60	8.8	120.6
11283	Embryo	*dd*	05/11/2014	RNeasy micro kit	DNase I	9.80	1.92	0.44	10.0	98.0
11284	Embryo	*dd*	27/10/2014	RNeasy micro kit	DNase I	12.90	1.73	1.12	9.3	120.0
11287	Embryo	*dd*	05/11/2014	RNeasy micro kit	DNase I	8.00	2.02	1.51	9.4	75.2
11301	Embryo	*dd*	20/11/2014	RNeasy micro kit	DNase I	12.10	2.23	0.65	10.0	121.0
11303	Embryo	*dd*	25/11/2014	RNeasy micro kit	DNase I	13.30	2.30	1.60	10.0	133.0

Table 16 Details of RNA extraction and cDNA synthesis for the foot tissue samples used in the qPCR experiments. Table includes: sample identifier (ID) and genotype (Geno) of the individual snail; Spectrophotometry data of the total RNA sample including sample concentration (ng/µl) and 260/280 & 260/230 absorbance ratios; volume (µl RNA) and quantity (ng RNA) of total RNA used for cDNA synthesis. The individual used as the calibrator sample in the genotype analysis is indicated by 'C'.

ID	Tissue	Geno	Extraction Date	Extraction method	DNase Treatment	Total RNA			cDNA synthesis	
						ng/µl	260/280	260/230	µl RNA	ng RNA
11347	Foot	*DD*	11/03/2015	TRI Reagent, RNeasy micro kit	DNase I	72.02	1.95	1.4	6.9	496.9
11350[C]	Foot	*DD*	12/03/2015	TRI Reagent, RNeasy micro kit	DNase I	49.54	1.96	1.51	10.1	500.4
11351	Foot	*DD*	12/03/2015	TRI Reagent, RNeasy micro kit	DNase I	85.33	2.19	2.26	5.9	503.4
11352	Foot	*DD*	12/03/2015	TRI Reagent, RNeasy micro kit	DNase I	67.62	2.1	1.95	7.4	500.4
11357	Foot	*DD*	13/03/2015	TRI Reagent, RNeasy micro kit	DNase I	62.15	2.23	1.25	8	497.2
11348	Foot	*dd*	12/03/2015	TRI Reagent, RNeasy micro kit	DNase I	74.69	2.03	2.12	6.7	500.4
11349	Foot	*dd*	12/03/2015	TRI Reagent, RNeasy micro kit	DNase I	70.75	2.05	2.03	7.1	502.3
11353	Foot	*dd*	13/03/2015	TRI Reagent, RNeasy micro kit	DNase I	69.2	2.41	1.36	7.2	498.2
11354	Foot	*dd*	13/03/2015	TRI Reagent, RNeasy micro kit	DNase I	78.2	2.07	1.72	6.4	500.5
11356	Foot	*dd*	13/03/2015	TRI Reagent, RNeasy micro kit	DNase I	76.98	2.09	1.98	6.5	500.4

Table 17 Details of RNA extraction and cDNA synthesis for the ovotestis tissue samples used in the qPCR experiments. Table includes: sample identifier (ID) and genotype (Geno) of the individual snail, PCR used to identify genotype: 1: cb3g FP1 F8R8, 2: 1315-507, 3: n/a (homozygous lines); Spectrophotometry data of the total RNA sample including sample concentration (ng/µl) and 260/280 & 260/230 absorbance ratios; volume (µl RNA) and quantity (ng RNA) of total RNA used for cDNA synthesis. The individual used as the calibrator sample in the genotype analysis is indicated by 'C'.

ID	Tissue	Geno	Extraction Date	Extraction method	DNase Treatment	Total RNA			cDNA synthesis	
						ng/µl	260/280	260/230	µl RNA	ng RNA
8515	Ovotestis	*DD*[1]	25/06/2012	TRI Reagent	n/a	89.4	1.8	1.0	12.6	375.3
8548[C]	Ovotestis	*DD*[1]	19/06/2012	TRI Reagent	n/a	112.6	1.8	1.3	13.5	506.6
8582	Ovotestis	*DD*[1]	18/06/2012	TRI Reagent	n/a	61.4	1.7	1.4	25.5	522.1
8583	Ovotestis	*DD*[1]	26/06/2012	TRI Reagent	n/a	77.8	1.8	0.9	22.1	571.8
9014	Ovotestis	*DD*[1]	29/06/2012	TRI Reagent	n/a	169.7	1.8	1.8	8.7	492.1
8554	Ovotestis	*Dd*[1]	18/06/2012	TRI Reagent	n/a	141.6	1.9	1.4	12.0	566.4
8555	Ovotestis	*Dd*[1]	18/06/2012	TRI Reagent	n/a	143.1	1.9	0.4	10.8	515.0
8559	Ovotestis	*Dd*[1]	18/06/2012	TRI Reagent	n/a	137.0	1.8	1.1	11.1	506.8
8562	Ovotestis	*Dd*[1]	18/06/2012	TRI Reagent	n/a	141.0	1.8	1.3	12.0	563.8
9013	Ovotestis	*Dd*[1]	29/06/2012	TRI Reagent	n/a	149.3	1.8	1.8	10.2	507.8
8806	Ovotestis	*dd*[1]	19/06/2012	TRI Reagent	n/a	149.5	1.9	1.1	10.5	523.1
8808	Ovotestis	*dd*[1]	25/06/2012	TRI Reagent	n/a	110.3	1.8	1.5	12.6	463.1
8996	Ovotestis	*dd*[1]	29/06/2012	TRI Reagent	n/a	106.1	1.7	1.5	15.0	530.5
9005	Ovotestis	*dd*[1]	26/06/2012	TRI Reagent	n/a	110.0	1.7	1.5	14.1	517.0
9007	Ovotestis	*dd*[1]	29/06/2012	TRI Reagent	n/a	107.3	1.8	1.1	15.0	536.7
10627	Ovotestis	*DD*[2]	25/09/2013	TRI Reagent	n/a	60.1	1.8	1.5	16.6	498.9
10633	Ovotestis	*DD*[2]	12/09/2013	TRI Reagent	n/a	57.8	1.9	0.9	17.3	499.9
10636	Ovotestis	*DD*[2]	25/09/2013	TRI Reagent	n/a	82.4	1.8	1.8	12.1	498.6
10638	Ovotestis	*DD*[2]	13/09/2013	TRI Reagent	n/a	67.5	1.8	1.9	14.8	499.2

10622	Ovotestis	*Dd*[2]	25/09/2013	TRI Reagent	n/a	62.1	1.8	1.6	16.1	499.7
10629	Ovotestis	*Dd*[2]	13/09/2013	TRI Reagent	n/a	59.5	1.9	1.7	16.8	499.7
10631	Ovotestis	*Dd*[2]	11/09/2013	TRI Reagent	n/a	68.6	1.9	1.1	14.6	500.4
10639	Ovotestis	*Dd*[2]	11/09/2013	TRI Reagent	n/a	208.9	1.9	1.8	4.7	490.9
10626	Ovotestis	*dd*[2]	25/09/2013	TRI Reagent	n/a	74.2	1.9	1.0	13.4	497.2
10630	Ovotestis	*dd*[2]	12/09/2013	TRI Reagent	n/a	64.1	1.8	1.0	15.6	499.7
10640	Ovotestis	*dd*[2]	11/09/2013	TRI Reagent	n/a	73.1	1.8	1.3	13.7	500.9
10642	Ovotestis	*dd*[2]	13/09/2013	TRI Reagent	n/a	53.2	1.9	1.7	18.7	497.4
11347	Ovotestis	*DD*[3]	11/03/2015	TRI Reagent	DNA-free™& Precision DNase	58.9	1.1	0.4	8.5	500.7
11350	Ovotestis	*DD*[3]	12/03/2015	TRI Reagent	DNA-free™& Precision DNase	95.8	1.5	0.7	5.2	497.9
11351	Ovotestis	*DD*[3]	12/03/2015	TRI Reagent	DNA-free™& Precision DNase	67.7	1.5	0.3	7.4	500.9
11352	Ovotestis	*DD*[3]	12/03/2015	TRI Reagent	DNA-free™& Precision DNase	79.1	1.3	0.3	6.3	498.2
11357	Ovotestis	*DD*[3]	13/03/2015	TRI Reagent	DNA-free™& Precision DNase	61.1	1.4	0.2	8.2	500.8
11348	Ovotestis	*dd*[3]	12/03/2015	TRI Reagent	DNA-free™& Precision DNase	63.7	1.2	0.2	7.8	496.7
11349	Ovotestis	*dd*[3]	12/03/2015	TRI Reagent	DNA-free™& Precision DNase	48.0	1.0	0.3	10.2	489.8
11353	Ovotestis	*dd*[3]	13/03/2015	TRI Reagent	DNA-free™& Precision DNase	68.7	1.3	0.4	7.3	501.7
11354	Ovotestis	*dd*[3]	13/03/2015	TRI Reagent	DNA-free™& Precision DNase	63.3	1.3	0.3	7.9	499.7
11356	Ovotestis	*dd*[3]	13/03/2015	TRI Reagent	DNA-free™& Precision DNase	94.5	1.5	0.7	5.3	500.7

Primer Design

All primers were designed using Primer 3 (Untergasser, Cutcutache et al. 2012). All primer pairs were designed to have a Tm range within 2°C of each other and amplicon product size between 110-130 base pairs (bp). As described in Chapter 2 (Methods; Primer design), to improve amplification specificity GC clamps were included where possible, and primer pairs were selected with the lowest possible 'Th' ('Th'ermodynamic secondary structure alignments) scores (Table 18, Table 19).

Where possible the primer pairs were designed to include an intron. As a result, the pair either did not amplify a product from contaminating genomic material or the product produced was of a larger size than that amplified from transcriptomic cDNA, detectable through gel electrophoresis and the Tm melt curve step. The only primer pairs in this experiment that did not span multiple exons were *Ldia1 3' UTR* and *Ldia2 3' UTR*. This was due to the lack of an intron being present in the 3' UTR region. These were the only primer pairs which amplified the same sized amplicon from both genomic and cDNA templates.

Functional targets

The DE loci selected for further analysis identified through the previous eRAD sequencing analysis, also contained a paired-end (pe) contig sequence, which was assembled from the eRAD dataset (full details of the method are described in Chapter 4, although these specific loci were described from a previous analysis not presented). The contig was matched to a predicted protein in the UniRef90 database via 'Blast' to identify the closest related cluster sequence and associated gene/protein description. The pe-contig was paired with a genomic contig from the latest alignment of *Lymnaea* genomic sequence data (version 10, note that this has since been updated) via a local Blast, to identify the position of introns to enable the design of exon-spanning primer pairs, and additionally to ascertain whether or not there were multiple regions within the *L. stagnalis* genome that the contig sequence may specify.

The name and predicted functions of the nine GOIs selected for further analysis are described below.

Actin-related protein complex 2/3

Two separate genes specific to different subunits of the actin-related protein (Arp) 2/3 complex were identified in the eRAD sequence data. The Arp 2/3 complex is comprised of seven separate subunits and is recognised to regulate the nucleation process of actin filaments and have strong interactions with formins (Welch, DePace et al. 1997, Goley and Welch 2006, Pollard 2007). There have been indications that the subunits of the Arp 2/3 complex may have specialisations and as a result be differentially regulated (Gournier, Goley et al. 2001). Therefore both subunit genes were

included in the experiment. These represent the Arp 2/3 complex subunit 1a, and Arp 2/3 complex subunit 3, referred to here as '*Larp2/3 1a*' and '*Larp2/3 3*' respectively.

Heavy chain myosin

Transcripts relating to two forms of heavy chain myosin (mhc) were identified, one muscle and one non-muscle form, referred to here as '*Lmhc*' and '*Lmhc nm*' respectively. The *Lmhc* transcript was found most closely related in sequence to a form of mhc called 'catchin' which is formed from a splice variant of the mhc, and is specific to the Molluscan catch muscle (Yamada, Yoshio et al. 2000). Because *L. stagnalis* does not contain a catch muscle, it is likely that this transcript is not actually specific to catchin but instead represents a similar mhc gene.

The non-muscle myosin represents an isoform of myosin II, which is associated with actin-binding and cell-cell adhesion (Vicente-Manzanares, Ma et al. 2009). Although *Lmhc nm* has more functional implications with embryonic polarity than *Lmhc* (Guo and Kemphues 1996, Vicente-Manzanares, Ma et al. 2009), both were included within the experiment for comparison.

Unconventional myosins

Unconventional myosins are expected to have a diverse range of functions within the cell (Wu, Jung et al. 2000, Redowicz 2007, Maravillas-Montero and Santos-Argumedo 2012) and have been directly linked to LR asymmetry in *Drosophila* (Hozumi, Maeda et al. 2006, Speder, Adam et al. 2006). Two specific unconventional myosin related genes were included in this experiment: myosin Va and myosin XVIIIa, here referred to as '*Lmyo5a*' and '*Lmyo18a*' respectively.

Myosin V has been characterised as a processive actin-based motor, transporting cargo along actin tracks, however does not form actin filaments (Cheney, Oshea et al. 1993, Mehta, Rock et al. 1999, Sellers and Veigel 2006). Myosin V has also been observed to interact with a number of cytoskeletal elements not just actin (Nagashima, Torii et al. 2002).

Myosin XVIIIa, is a more recently described myosin class (Furusawa, Ikawa et al. 2000) and as such is less studied, yet has been observed to co-localise with microfilaments and may have roles associated with the golgi membrane (Yamashita, Sellers et al. 2000, Dippold, Ng et al. 2009).

Collagen, staufen & unc-93a

The collagen included is specific to the collagen type XI alpha subunit 2 or 1 and is here referred to as 'Lcol11a 2/1'. The group XI collagens are recognised as fibrillar collagens, which self-assemble to form a structural network of striated fibrils which function to resist pulling forces (Keene, Oxford et

al. 1995, Kadler, Holmes et al. 1996). Although collagen fibrils are a major component of the cartilage, their expression is not restricted to cartilaginous tissues (Bernard, Yoshioka et al. 1988).

The transcript specific to the RNA binding protein staufen, identified in the eRAD sequence data was included in the experiment to explore alternatives to direct structural associations, and is here referred to as '*Lstau*'. This is the only gene included, which has specific functions linked to maternal mRNAs (St Johnston, Beuchle et al. 1991).

The unc93a like gene represents somewhat more of a wild card addition to the experiment. The protein is largely uncharacterised in function, however has been linked to egg laying and the ovaries in other species (Liu, Dodds et al. 2002, de la Cruz, Levin et al. 2003).

Proximal targets

The gene sequences for targets identified through regional genomic analyses were obtained by the Davison group. Intronic regions were located by performing a local blast of the genomic gene sequence to the transcriptomic resources for *L. stagnalis* also available in the Davison research group (Liu, Davey et al. 2013, Liu, Davey et al. 2014). Pairwise alignments were then generated of the two sequences using NCBI blast online (http://blast.ncbi.nlm.nih.gov/Blast.cgi).

Due to the highly conserved sequence similarity between the diaphanous formins, *Ldia1* and *Ldia2*, regions of gene specific sequence large enough to design primers were scarce. The untranslated regions (UTRs) held the largest number of sequence variations between the two genes. Consequently, primers were designed for each gene in the 3' UTR; '*Ldia1 3'UTR*' and '*Ldia2 3'UTR*'.

Another pair of primers was designed to target the region of the open reading frame (ORF) in *Ldia2*, which includes the frameshift suspected to cause chiral reversal in *L. stagnalis;* '*Ldia2 ORF'*. This additional target was included in attempt to infer any information about the regulation of the ORF compared to the 3' UTR. However, there were far less sequence differences between *Ldia1* and *Ldia2* in the ORF. The primers were designed to amplify the *Ldia2* candidate gene and contain three bases different from the *Ldia1* sequence, including the last two consecutive bases on the leading edge of the forward primer, increasing the likelihood of amplifying only the *Ldia2 ORF*.

Primer specificity

All primer pairs were first tested via a conventional non-quantitative PCR using a representative cDNA sample and a genomic control sample and water negative control. The products were visualised via fluorescent agarose gel electrophoresis to verify the expected size of products (Box 2).

In addition, the sequence specificity of the amplicons generated from *Ldia1 3' UTR*, *Ldia2 3' UTR* & *Ldia2 ORF* were verified through Sanger sequencing of conventional PCR products generated from a pooled *DD* single cell embryo cDNA template. Due to the presence of variable Tm peaks in the *dd* samples in the *Ldia2 ORF* qPCR experiments, Sanger sequencing of *Ldia2 ORF* was additionally performed on a PCR product generated from cDNA of both a pooled *DD* and a pooled *dd* embryo template. Sanger sequencing was performed in both forward and reverse directions using the original qPCR primers. No cloning was undertaken. The protocols for the Sanger sequencing sample preparation are available in the SI (S7).

Primer amplification efficiency

Primer efficiencies were calculated for each of the fourteen primer pairs used in the qPCR experiments using the same methodology as described in Chapter 2 (Methods, Primer amplification efficiency) although a different reference sample was used. The standard curve experiments of the functional GOIs (*Larp2/3 1a*, *Larp2/3 3*, *Lmhc*, *Lmhc nm*, *Lcol11a 2/1*, *Lmyo5a*, *Lmyo18a*, *Lstau* and *Lunc93a*), were assessed using an ovotestis reference pool sample, equal to that used in the amplification efficiency experiments of the endogenous control genes. The standard curve experiments for the proximal GOIs (*Ldia1*, *Ldia2 3' UTR*, *Ldia2 ORF*, *Lfat1* and *Lfry*) were calculated using a single-cell embryo reference sample. This sample was created by pooling equal amounts of cDNA generated from an RNA sample extracted from single-cell embryos from multiple *DD* individuals and another sample pooled from multiple *dd* individuals, using the RNA extraction and cDNA synthesis protocol already described. The *Ldia2 3' UTR* primer pair was also assessed for amplification efficiency using a foot tissue reference sample. The reference foot sample was created by pooling equal volumes of cDNA from all ten foot samples together.

Table 18 Primer sequence information for amplification of the nine functional GOIs including: primer name and associated protein from the UniRef90 hit of its most closely related protein product; gene abbreviation (Abv.) used throughout this analysis; Primer sequence in the 5' to 3' direction; Primer length (P.L) & amplicon length (A.L) in nucleotides; primer melting temperature (Tm) and the difference between melting temperature within each primer pair (Tm diff); the estimate of mispriming to any sequence (Any th) and specifically mispriming at the 3' end (3' th); and the predicted intron size between the two primers. * primer lies on an exon boundary. †full intron information unknown due to the transcriptomic sequence crossing two genomic contigs, the minimum intron size is presented.

Primer Name	Associated Protein	UniRef90 hit	Abv.	Sequence 5'-3'	P.L	A.L	Tm (°C)	Tm diff	%GC	Any th	3' th	Predicted Intron (bp)
ARPI_1-2ab_F	Actin-related protein 2/3 subunit 1a	UniRef90_K1R488	*Larp2/3 1a*	CTGAAAATAGCCTTGTTGCAGC	22	115	58.75	1.25	45.45	0	0	341
ARPI_1-2b_R				CCAGACTCCTTTTCCTGGGAC	21		60.00		57.14	0	0	
ARPII_1-3a_F	Actin-related protein 2/3 subunit 3	UniRef90_C3KIX3	*Larp2/3 3*	AGCCAGCTAACAAGGGAGAAG	21	129	59.72	0.20	52.38	0	0	>900†
ARPII_1-3a_R				AGCATAGCCACCATTTGCTTG*	21		59.52		47.62	0	0	
COL2A_3-4a_F	Collagen type XI alpha 2/1	UniRef90_G3HQS2	*Lcol11a 2/1*	TGGTCGACTTGGAAAGGATGG	21	110	60.00	0.23	52.38	16.46	0	728
COL2A_3-4a_R				CTCTGTGTCCTTTCTCTCCTGG	22		59.77		54.55	0	0	
MHCI_1-2a_F	Myosin heavy chain	UniRef90_Q9NJ19	*Lmhc*	TCAGATTGAGGAGGCCAACG	20	125	59.75	0.50	55	0	0	210
MHCI_1-2a_R				TCTCCAACTCGTGTGTGCTG	20		60.25		55	0	0	
MHCII_2-3a_F	Myosin heavy chain non-muscle	UniRef90_Q45R40	*Lmhc nm*	GCTACAGACAACAAGGGCTTC	21	111	59.19	0.65	52.38	0	0	338
MHCII_2-3a_R				ACAAATCAATGCCATCCGTGTC	22		59.84		45.45	0	0	
MV_F2	Myosin Va	UniRef90_F6K356	*Lmyo5a*	TTCAGCCCAGTATTGTCCCC*	20	115	59.38	0.89	55	0	0	1658
MV_R2				TCCTCTGTTTCCCTGGCATTG	21		60.27		52.38	0	0	
Staufen_3-4a_F	RNA binding protein Staufen	UniRef90_E2QDA4	*Lstau*	CTTGCGCAGAAACATGCCTG	20	116	60.73	0.52	55	6.12	6.12	125
Staufen_3-4a_R				TCCCCTCTCCTTCTGTCACC	20		60.25		60	0	0	
UMVIII_F2	Unconventional myosin -XVIIIa	UniRef90_K1QV80	*Lmyo18a*	GTCCAGCAGTCCTTTGAGAAC	21	129	58.85	0.32	52.38	0	0	497
UMVIII_R2				AAACTGGGGCTTGTTGTTGG	20		59.17		50	0	0	
UNC-93_F	unc-93 homolog a	UniRef90_K1P6Z5	*Lunc93a*	GAAGGAGGTCAGGGCGATG	19	115	59.86	0.23	63.16	0	0	>1.8kb†
UNC-93_R				GCTGCTTTGTAGACTCTGTAACG	23		59.63		47.83	0	0	

Table 19 Primer sequence information for amplification of the four proximal GOIs including: primer name and associated protein according to Blastx top hits; gene abbreviation (Abv.) used throughout this analysis; Primer sequence in the 5' to 3' direction; Primer length (P.L) & amplicon length (A.L) in nucleotides; primer melting temperature (Tm) and the difference between melting temperature within each primer pair (Tm diff); the estimate of mispriming to any sequence (Any th) and specifically mispriming at the 3' end (3' th); and the predicted intron size between the two primers.

Primer Name	Associated Protein	Abv.	5'-3' Sequence	P.L	A.L	Tm (°C)	Tm Diff	%GC	Any th	3' th	Predicted Intron (bp)
qPCR_PARA_3'UTR_F1	diaphanous formin	*Ldia1 3' UTR*	AGTGGTGTGGGCAAAAGATG	20	117	58.67	0.05	50	0	0	n/a
qPCR_PARA_3'UTR_R1			TATTCTGTTGATGCACGGCC	20		58.62		50	0	0	
qPCR_FOR_3'UTR_F1	diaphanous formin	*Ldia2 3' UTR*	GGGAGTTCAAGTTCAAGCCTATC	23	122	59.06	0.98	47.83	0	0	n/a
qPCR_FOR_3'UTR_R1			GGCAAGCTACGACTCTTCTC	20		58.08		55	0	0	
qPCR_FOR_ORF_F1	diaphanous formin	*Ldia2 ORF*	GGGTGACAATGAAGTGGACC	20	126	58.47	0.58	55	0	0	713
qPCR_FOR_ORF_R1			ACATGCATCTGTAACATCTGCC	22		59.05		45.45	11.53	0	
qPCR_CAD_F1	protocadherin FAT1	*Lfat1*	TGCCCATGTTGCTAAGTTCAG	21	126	58.84	0.49	47.62	6.1	0	1345
qPCR_CAD_R1			CCTCTATCCCAGTTCGACGG	20		59.33		60	0	0	
qPCR_FURRY_F1	Furry (gene)	*Lfry*	ACTTACCCTGCTCAAATGCC	20	121	58.16	1.25	50	0	0	715
qPCR_FURRY_R1			ATGTTTCTTGTGCTGCCGTC	20		59.41		50	0	0	

qPCR

cDNA samples were diluted to an appropriate working concentration (indicated through the results of the amplification efficiency experiments) using PCR grade water and divided into aliquots. These were not strictly single use but allowed storage of samples in multiple tubes to minimise freeze thaw cycles. One aliquot provided enough sample to perform six target gene experiments. All cDNA samples were vortexed before each run. Working concentrations of cDNA used in the qPCR varied between tissues, however within tissue analyses, all cDNA samples were used at the same dilution. All qPCR reactions were performed as described in Box 3. PCR grade water was used as a negative control for all mastermixes.

Plate setup

Inter-run calibration

A reference sample was created by pooling multiple ovotestis cDNA samples of all genotypes, hereafter referred to as 'OvoRef'. The OvoRef sample can loosely be considered as representing a heterozygote. The OvoRef sample was diluted to a working concentration of 1:30 and separated into smaller aliquots to minimise freeze-thaws. The single working dilution was made to a volume sufficient to be included on all experimental plates and provides an appropriate calibrator sample for calculating relative expression ratios across all tissues/plates.

All qPCR experiments were performed within 20 days. A new tube of SYBR green was defrosted and used within the day (no freeze-thaws) and light-exposure of mastermixes was kept to a minimum.

Embryo and foot tissue experiments

All samples, including the negative control and OvoRef sample were performed in triplicate repeat within the single-cell embryo and foot experiments.

The single-cell embryo experiment consisted of 17 samples plus an embryo reference sample (P1c), diluted to a working concentration of 1:15, necessitating the use of 54 wells per GOI. Therefore, a maximum of one GOI could be performed on all samples within a single 96 well plate. The foot experiment comprised only 10 samples, diluted to a working concentration of 1:30, necessitating the use of 30 wells per GOI. Subsequently the foot samples were run on the same plate as the embryo samples, requiring a total 84 wells. A single master-mix was created for each embryo and foot qPCR plate and therefore the same negative control (water) and OvoRef samples functioned for both experiments, requiring an additional 6 wells and therefore a total of 90 wells per GOI/plate (Figure 10a).

Three endogenous control genes were quantified in the embryo and foot tissue experiments, namely *Lhis2a*, *Lube2* and *Lywhaz* (as described in Chapter 2).

Due to the reduced amount of single cell embryo sample only eight GOIs could be included in addition to the three endogenous control genes in the embryo experiment. The following GOIs were selected to be quantified in the embryo and foot tissue: *Larp2/3 1a*; *Larp2/3 3*; *Ldia1 3' UTR*; *Ldia2 3' UTR*; *Ldia2 ORF*; *Lfat1*, *Lfry* & *Lmhc*.

Therefore, the remaining six GOIs (*Lcol11a 2/1*; *Lmhc nm*; *Lmyo5a*; *Lmyo18a*; *Lstau* & *Lunc93a*) were assessed in the foot tissue alone. The 10 foot samples, the negative control and the OvoRef sample were again performed in triplicate repeat, requiring a total of 36 wells per GOI. Therefore, two GOIs could be assessed within one 96 well plate (Figure 10b).

Ovotestis experiments

The ovotestis experiment comprised of 37 samples, diluted to a working concentration of 1:30. The ovotestis experimental samples, in addition to the negative control were performed in duplicate repeat, whereas the OvoRef sample was performed in quadruplicate repeat, thus requiring a total of 80 wells per GOI. Subsequently only one GOI was included per 96 well plate (Figure 10c). Three endogenous control genes were quantified in the ovotestis tissue, namely *Lhis2a*, *Lube2* and *Lrpl14* (as described in Chapter 2).

a

Em/Fo	1	2	3	4	5	6	7	8	9	10	11	12
A	11289	11289	11289	11282	11282	11282	11358	11358	11358	11349	11349	11349
B	11292	11292	11292	11283	11283	11283	11359	11359	11359	11350	11350	11350
C	11293	11293	11293	11284	11284	11284	11360	11360	11360	11351	11351	11351
D	11295	11295	11295	11287	11287	11287	11361	11361	11361	11352	11352	11352
E	11297	11297	11297	11301	11301	11301	11363	11363	11363	11353	11353	11353
F	11298	11298	11298	11303	11303	11303	1 cell pool	1 cell pool	1 cell pool	11354	11354	11354
G	x	x	x	x	x	x	11347	11347	11347	11356	11356	11356
H	OvoRef	OvoRef	OvoRef	H_2O	H_2O	H_2O	11348	11348	11348	11357	11357	11357

b

Fo	1	2	3	4	5	6	7	8	9	10	11	12
A	11347	11348	11349	11350	11351	11352	11353	11354	11356	11357	OvoRef	H_2O
B	11347	11348	11349	11350	11351	11352	11353	11354	11356	11357	OvoRef	H_2O
C	11347	11348	11349	11350	11351	11352	11353	11354	11356	11357	OvoRef	H_2O
D	11347	11348	11349	11350	11351	11352	11353	11354	11356	11357	OvoRef	H_2O
E	11347	11348	11349	11350	11351	11352	11353	11354	11356	11357	OvoRef	H_2O
F	11347	11348	11349	11350	11351	11352	11353	11354	11356	11357	OvoRef	H_2O
G	x	x	x	x	x	x	x	x	x	x	x	x
H	x	x	x	x	x	x	x	x	x	x	x	x

c

Ov	1	2	3	4	5	6	7	8	9	10	11	12
A	10622	10626	10627	10629	10630	10631	10633	10636	10638	10639	10640	10642
B	10622	10626	10627	10629	10630	10631	10633	10636	10638	10639	10640	10642
C	11347	11348	11349	11350	11351	11352	11353	11354	11356	11357	OvoRef	OvoRef
D	11347	11348	11349	11350	11351	11352	11353	11354	11356	11357	OvoRef	OvoRef
E	8515	8548	8554	8555	8559	8562	8582	8583	8806	8808	8996	9005
F	8515	8548	8554	8555	8559	8562	8582	8583	8806	8808	8996	9005
G	9007	9013	9014	H_2O	x	x	x	x	x	x	x	x
H	9007	9013	9014	H_2O	x	x	x	x	x	x	x	x

Figure 10 qPCR experimental plate setup. The embryo and foot combined experiments (a), the remaining foot experiments, including two GOIs per plate (indicted by the different background colour) (b) and the ovotestis experiments (c). Unused wells are indicated by 'x'.

Data Analysis

Cq data export

Cq values were exported for each well of the experiment using the 7500 software. Average Cq values derived from triplicate or duplicate repeats of each sample were used in analyses. Ideally only samples with standard deviation of <0.5 were used in the analysis. This occasionally involved removing perceived outliers, as described in Chapter 2, from the dataset. However, in some instances it was not possible to reliably designate a value for removal and subsequently a small number of average Cq values are included with high (>0.5) standard deviation (Table 21 - Table 28).

Ovotestis sample 10631 was removed from all analyses due to numerous failures to amplify a product.

The Cq data obtained from embryo samples included in the *Lmhc* qPCR were omitted from analysis due to high standard deviation and a failure to amplify a product in the majority of samples.

Expression Ratios

Normalised expression ratio or 'normalised relative quantity' (NRQ) values were calculated from the average Cq value of each sample using the Pfaffl method (Pfaffl 2001, Pfaffl, Tichopad et al. 2004, Hellemans, Mortier et al. 2007) (Equation 2). For each sample, firstly the relative quantity per target gene (ΔCq target) was calculated by subtracting the average Cq value of the sample from that of the calibrator sample. This delta Cq (ΔCq) value was then corrected for amplification efficiency (E) by multiplying delta Cq to the base percentage amplification efficiency (represented as a value between 1 and 2). The efficiency corrected relative quantities were then normalised to the endogenous control genes by dividing by the geometric mean (geoM) of the efficiency corrected delta Cq values calculated for each of the control genes (ΔCq ref) in the same manner as described above.

Equation 2 Formula according to Pfaffl's method to calculate normalised expression ratios relative to a calibrator sample whilst incorporating the amplification efficiency of each target. Formula explanation in main text.

$$\text{Normalised expression ratio} = \frac{(\text{Etarget})^{\Delta\text{Cq target(calibrator-sample)}}}{\text{geoM(Eref)}^{\Delta\text{Cq ref(calibrator-sample)}}}$$

The calculations were performed in Microsoft Excel 2010. All values were corrected for primer efficiencies using the average efficiency calculated via standard curve experiments described previously (Table 20).

NRQ values were calculated separately for the genotype analysis and tissue analysis. These were generated from the same raw Cq data, but relative to a different calibrator sample and normalised to a different combination of endogenous controls.

Genotype Analysis

A *DD* sample with a relatively high Cq value was used as the calibrator sample for the genotype comparison analysis within each tissue. Sample 11295 functioned as the calibrator for the embryo genotype analysis. Sample 11350 was the calibrator for the foot genotype analysis, and sample 8548 was used for the ovotestis genotype analysis. Consequently, all NRQs in the genotype analyses represent an expression ratio relative to a conspecific individual *DD* sample. All three endogenous control genes quantified within each tissue, contributed to the geometric mean of the endogenous control genes.

Tissue Analysis

The ovotestis reference sample, OvoRef, provided the calibrator for the tissue comparison analysis. The NRQ value for each experimental sample was calculated relative to the average Cq value only of the OvoRef sample which was quantified on the same experimental plate. The NRQs in the tissue analysis represent an expression ratio relative to an ovotestis sample of mixed genotype. NRQ values were normalised to the geometric mean of the two endogenous control genes quantified in all tissues, namely *Lhis2a* and *Lube2*.

Differential expression analysis

All statistical calculations were performed in the basic R package (http//cran.r-project.org, R version 2.15.3). Graphs were produced using the addition of graphics packages ggplot2, 0.9.3.1, (Wickham 2009) and gcookbook, version 1.0, (Chang 2013). All statistical tests were performed on NRQ values log transformed to the base 10. No probability corrections were performed to accommodate for multiple comparisons. Summary statistics were generated in the R package. Additional calculations were performed in Microsoft Excel 2010 to calculate the standard error of the means.

Genotype Analysis

Boxplots and histograms were created of the log transformed (base10) NRQs (LOG NRQ) calculated for each gene of interest grouped according to genotype.

Embryo

Non-parametric pairwise comparisons of the group means of LOGNRQ values for each genotype were performed using the Wilcoxon-Mann-Whitney test. Pairwise comparisons were made between all genotypes for the seven GOIs assessed, resulting in a total of 21 pairwise comparisons.

Ovotestis

Non-parametric pairwise comparisons of the group means of LOGNRQ values for each genotype were performed using the Wilcoxon-Mann-Whitney test. Pairwise comparisons were made between all genotypes for the 14 GOIs assessed, resulting in a total of 42 pairwise comparisons.

Foot

Non-parametric pairwise comparisons of the group means of LOGNRQ values for each genotype were performed using the Wilcoxon-Mann-Whitney test. Pairwise comparisons were made between both genotypes for the 14 GOIs assessed, resulting in a total of 14 pairwise comparisons.

Tissue Analysis

Boxplots and histograms were created of the LOG NRQs calculated for each gene of interest grouped by either genotypic group or tissue, or both genotypic group and tissue.

Within each genotype, non-parametric pairwise comparisons of the group means of LOGNRQ values for each tissue were performed using the Wilcoxon-Mann-Whitney test. The number of comparisons per genotype is described below.

Within the *DD* samples, pairwise comparisons were made between all three tissues assessed for seven genes of interest and between the foot and ovotestis tissues for the additional seven genes assessed in those tissues; a total of 28 comparisons.

Comparisons of the foot tissue cannot be performed for the *Dd* genotype group due to a lack of representative samples. Therefore, pairwise comparisons of the *Dd* samples were made only between the embryo and ovotestis tissue for the seven genes of interest assessed in both tissues; a total of seven comparisons.

Within the *dd* samples, pairwise comparisons were made between all three tissues assessed for seven genes of interest and between the foot and ovotestis tissues for the additional seven genes assessed in those tissues; a total of 28 comparisons.

An overall total of 63 pairwise comparisons in the tissue analysis and 77 pairwise comparisons in the genotype analysis resulted in a grand total of 140 pairwise comparisons.

Results

General QC

Primer specificity

Conventional non-quantitative PCRs of cDNA and genomic DNA analysed via fluorescent agarose gels, showed amplified products to be of the expected size and additionally demonstrated the difference in amplicon size generated from a cDNA or genomic DNA template when using exon-spanning primers. There was no visible amplification of multiple products from any of the samples (Figure 11, Figure 12).

Some of the primer pairs could not amplify a product from a genomic template, namely *Larp2/3 3, Lfat1* and *Lmyo5a*. Other primer pairs could amplify a genomic product, although the resulting amplicon was of a substantial size difference, as seen in *Ldia2 ORF*, *Lfry*, *Larp2/3 1a*, *Lcol11a 2/1*, *Lmhc*, *Lmhc nm*, *Lmyo18a*, *Lstau* and *Lunc93a*. The only primer pairs to amplify a genomic product at the same size as a transcriptomic product were *Ldia1 3' UTR* and *Ldia2 3' UTR* (Figure 11, Figure 12).

The majority of melt temperature (Tm) curves of the qPCR reactions showed distinct peaks for experimental samples. A smaller peak at a lower Tm was often visible in the negative controls of some genes. Figure 13, Figure 14 and Figure 15 show representative Tm curves for each GOI. The presence and height of the lower Tm peaks varied between runs; however, the peak was generally at consistent, reduced temperature at a lower intensity and is assumed to represent primer dimer.

The only qPCRs to produce wide and variable Tm peaks, were the *Ldia2 3' UTR* and *Ldia2 ORF* (Figure 15). The wide Tm peaks were only seen in the *dd* embryo samples and not in the *DD* or *Dd* embryo samples, or in any of the foot or ovotestis samples. These samples also occasionally showed smaller peaks at higher temperatures than the specific amplicon. A number of the *dd* technical replicates of *Ldia2* were removed after being flagged by the software for the presence of multiple Tm peaks. All of the *dd* samples for *Ldia2 ORF* and most of the *dd* samples for *Ldia2 3' UTR* were reduced to duplicate repeat after data cleaning. Sample 11287 was included still flagged as producing multiple Tm peaks for *Ldia2* 3' UTR (Table 22).

The amplicons of a number of primer pairs were sequenced by Sanger sequencing to further verify specificity. This was important for *Ldia2 3' UTR* and *ORF* primers because of the multiple Tm peaks seen in some of the sinistral homozygote samples. Sequencing has shown the product to be specific

to the *Ldia2* gene. The spurious peaks are assumed to be various primer dimers resulting from the low concentration of target material present in the *dd* samples.

Amplification efficiency

All primer pairs demonstrated amplification efficiency between 1.775 and 1.986 with R^2 values exceeding 0.98. All primers demonstrated acceptable amplification efficiency in dilutions of up to 1:75/1.33% of full concentration (Table 20). The working concentrations of 1:15 and 1:30 used in the qPCR experiments fall well within these limits.

Sample quality

All samples underwent an intronic PCR reaction to check for amplification of genomic DNA specific products. Every embryo tissue sample used within this experiment failed to produce a PCR product from the intronic PCR. Conversely every ovotestis and foot sample did produce a clear intronic PCR product (presented in S5). No multiple products were seen in any of the exon-spanning test PCRs (Figure 11, Figure 12).

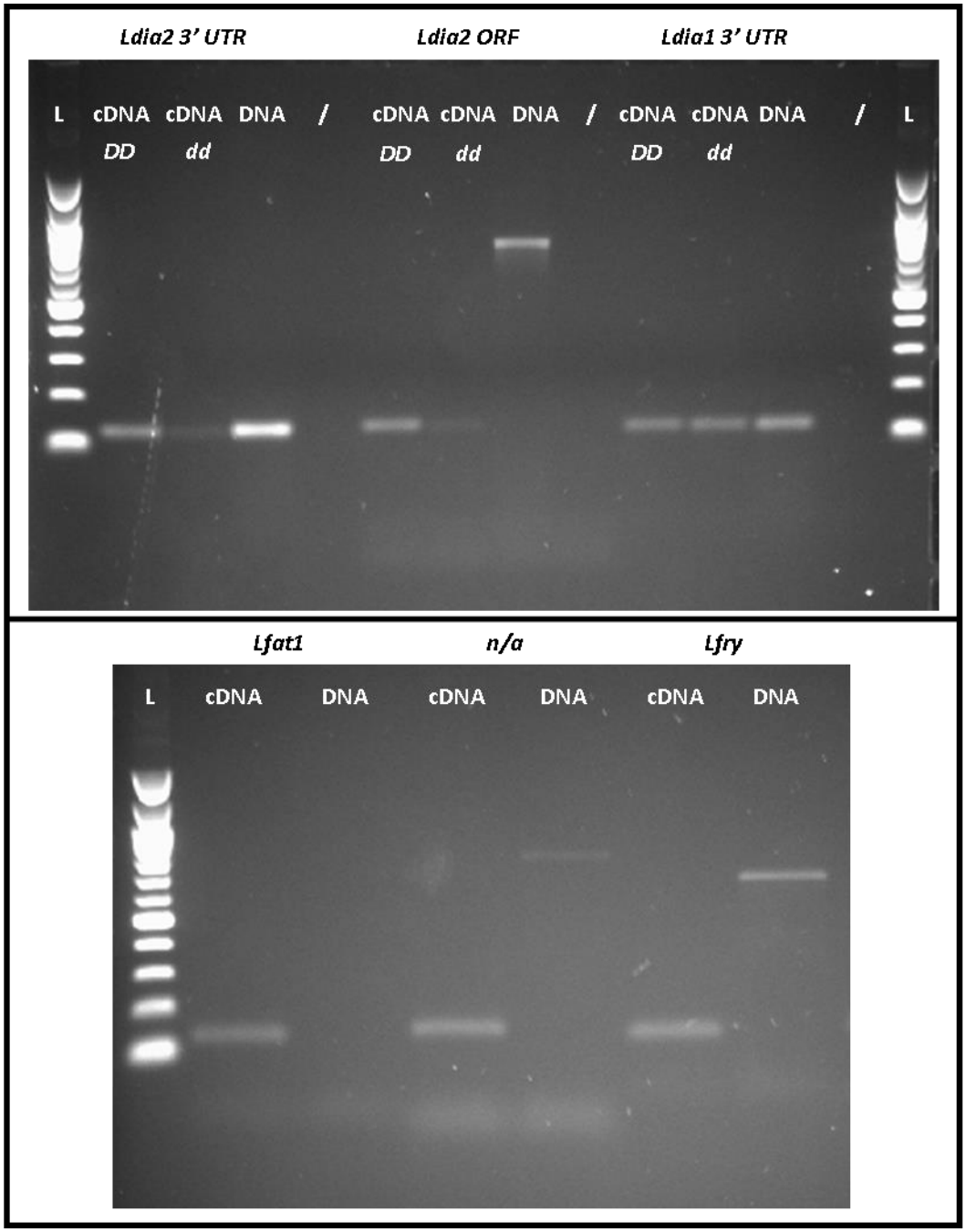


Figure 11 Composite UV visualisations of PCR products from each of the five proximal GOIs from cDNA (cDNA) and genomic DNA (DNA) templates, size fractionated through gel electrophoresis. The size of products is inferred from the DNA marker of known size (L). The PCR products of another pair of primers not used in this experiment (n/a) also appear on the gel.

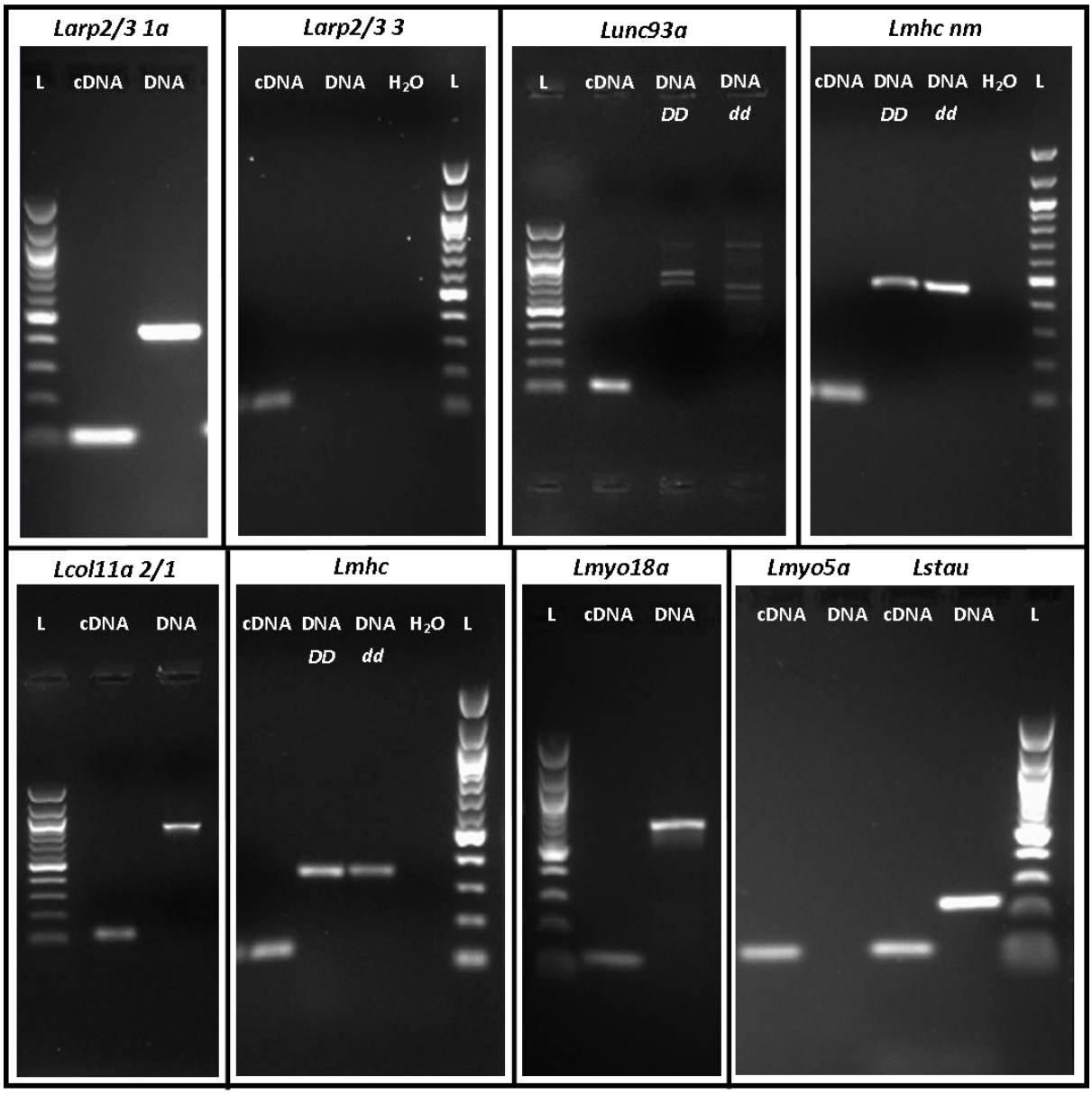


Figure 12 Composite UV visualisations of PCR products from each of the nine functional GOIs from cDNA and genomic DNA templates, size fractionated through gel electrophoresis. The size of products is indicated by the DNA marker of known size (L). Some gel images include the negative control (H_2O), some PCRs included cDNA or DNA samples from both homozygote genotypes *DD* and *dd*.

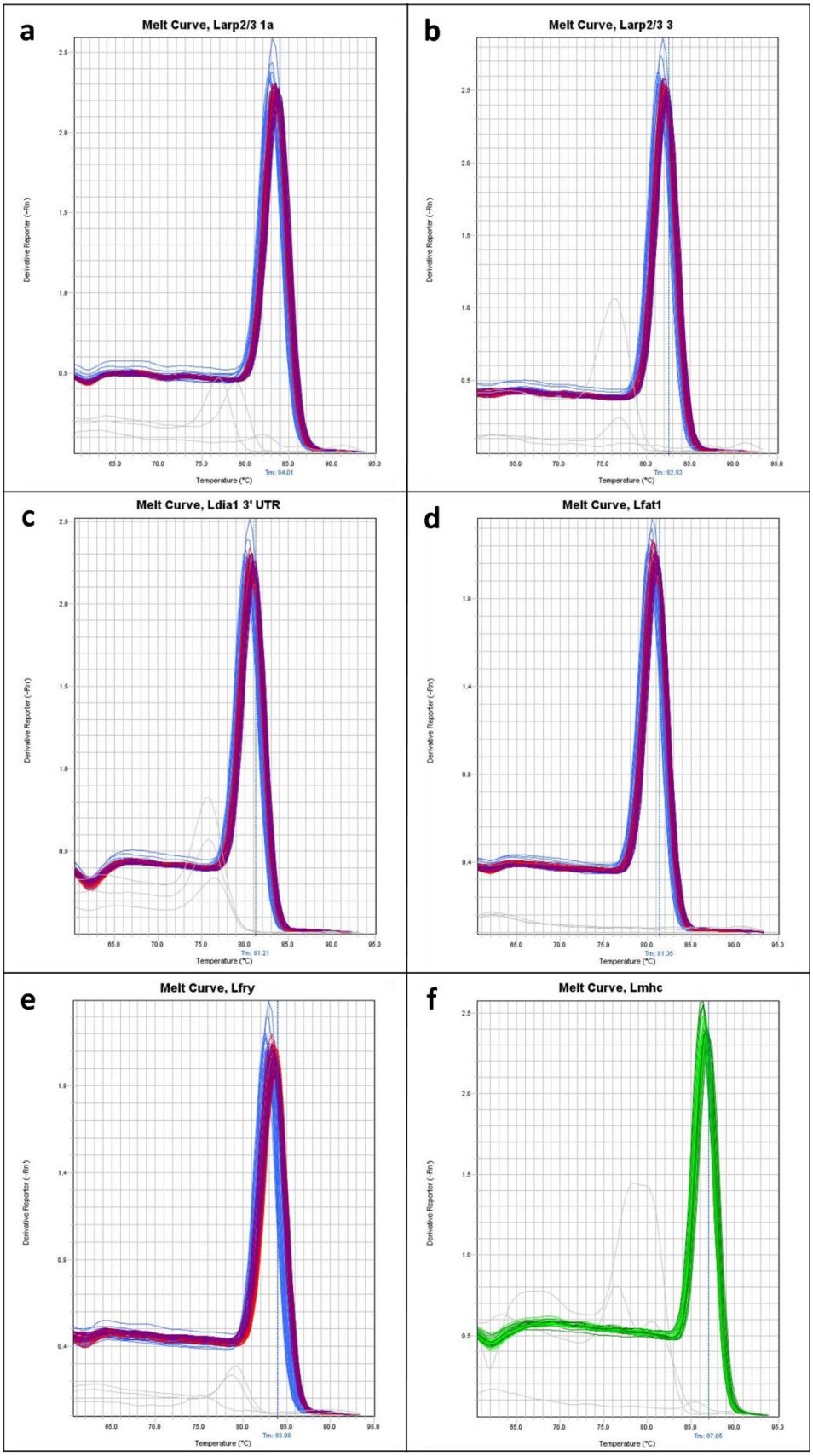


Figure 13 Representative temperature melt curves of qPCR amplification of *Larp2/3 1a* (a); *Larp2/3 3* (b); *Ldia1 3' UTR* (c); *Lfat1* (d); *Lfry* (e) and *Lmhc* (f). Tm curves a-e were produced from *DD* (blue), *Dd* (purple) and *dd* (red) embryo samples. Tm curve f was produced from *DD* (light green) and *dd* (dark green) foot samples. Negative controls are shown in grey.

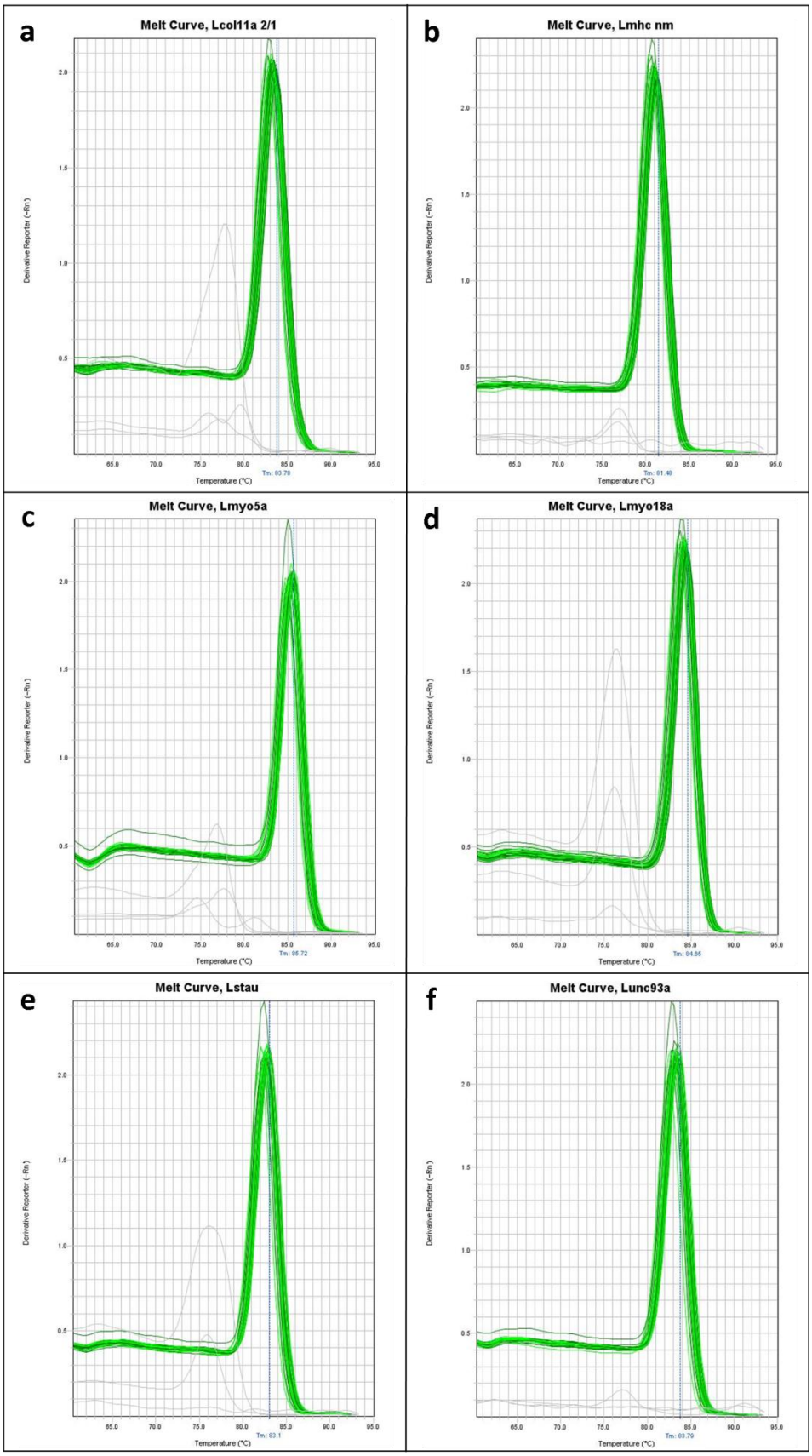


Figure 14 Representative temperature melt curves of qPCR amplification of *Lcol11a 2/1* (a); *Lmhc nm* (b); *Lmyo5a* (c); *Lmyo18a* (d); *Lstau* (e) and *Lunc93a* (f). Tm curves were produced from *DD* (light green) and *dd* (dark green) foot samples. Negative controls are shown in grey.

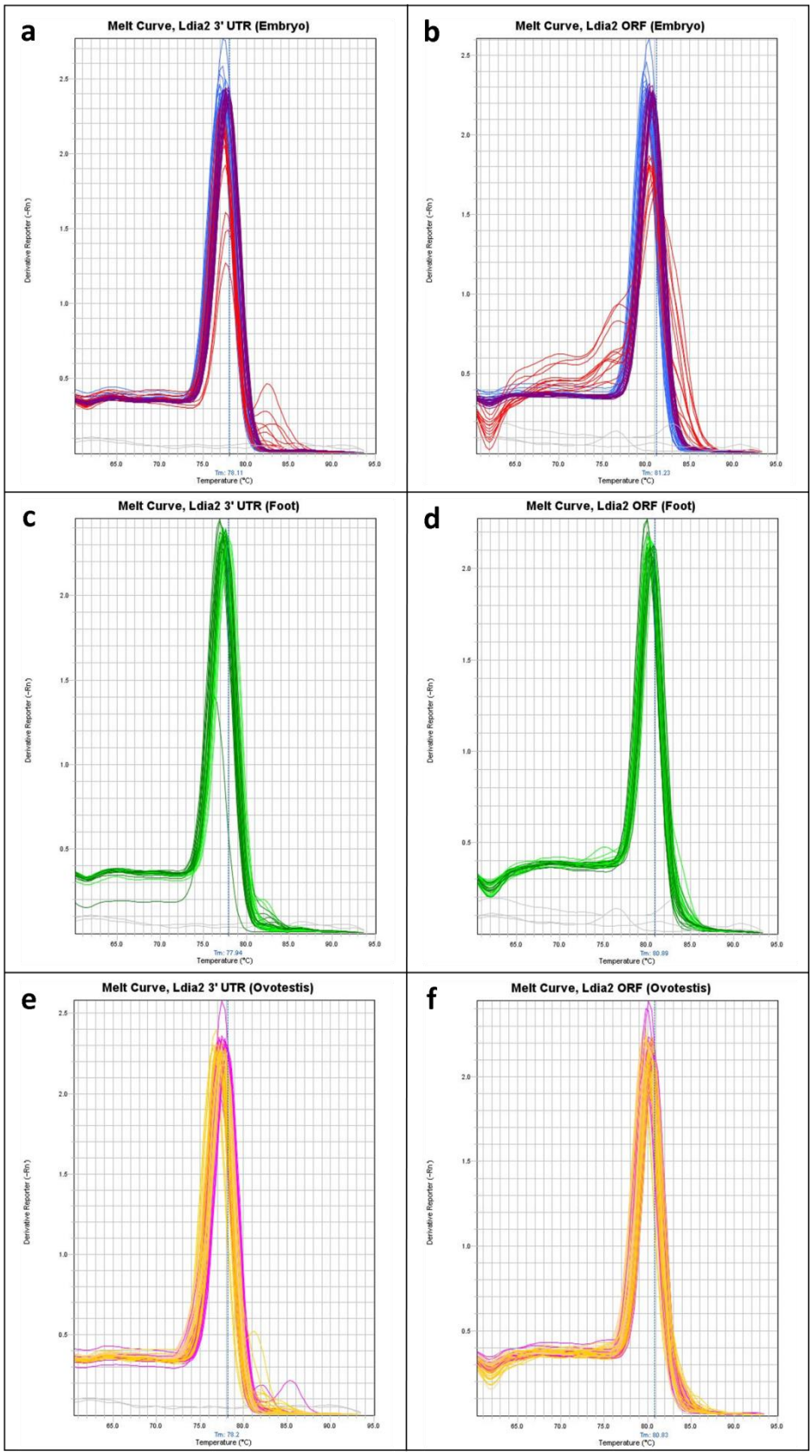


Figure 15 Representative temperature melt curves of qPCR amplification of *Ldia2 3' UTR* (a, c, e) and *Ldia2 ORF* (b, d, f). Tm curves a & b were produced from *DD* (blue), *Dd* (purple) and *dd* (red) embryo samples. Tm curves e & f were produced from *DD* (light green) and *dd* (dark green) foot samples. Tm curves e & f were produced from *DD* (magenta), *Dd* (peach) and *dd* (yellow) ovotestis samples Negative controls are shown in grey.

Table 20 Amplification efficiency estimates of each primer pair for the 14 GOIs assessed, represented by their gene abbreviation (Abv.). The average efficiency is quoted as the amount each template will increase per qPCR cycle (between 1 and 2). The minimum dilution is presented as a percentage of the undiluted original cDNA concentration required in the qPCR reaction. Additionally, the number of runs included to generate the average amplification efficiency is quoted and the tissue the experiments were performed on: Ovotestis reference sample (O); embryo reference sample (E), foot reference sample (F).

Primer Pair	Abv.	Efficiency (R2 > 0.98)	Minimum Dilution (%)	No. of runs included
ARPI_1-2b	*Larp2/3 1a*	1.847	0.67	3°
ARPII_1-3a	*Larp2/3 3*	1.775	0.67	2^{O}
COL2A_3-4a	*Lcol11a 2/1*	1.890	0.67	3°
MHCI_1-2a	*Lmhc*	1.892	0.13	3°
MHCII_2-3a	*Lmhc nm*	1.924	0.67	4°
MV_F2R2	*Lmyo5a*	1.946	0.13	2^{O}
UMVIII_F2R2	*Lmyo18a*	1.913	0.13	2^{O}
Staufen_3-4a	*Lstau*	1.957	0.67	2^{O}
UNC-93_FR	*Lunc93a*	1.978	0.13	2^{O}
PARA_3'_UTR	*Ldia1 3' UTR*	1.986	0.27	2^{E}
FOR_3'_UTR	*Ldia2 3' UTR*	1.912	0.27	$4^{E, F}$
FOR_ORF	*Ldia2 ORF*	1.948	1.33	2^{E}
CAD_F1R1	*Lfat1*	1.838	0.59	2^{E}
FURRY_F1R1	*Lfry*	1.876	0.13	2^{E}

Q RT PCR

Raw Cq data

The 7500 software used to design and run the qPCR experiments, automatically flagged a well when it perceived an issue that may compromise the quality of the Cq data. Flags included, multiple Tm peaks observed during the melt curve stage or high (>0.5) standard deviation (SD) between technical replicates. The majority of flagged wells were removed from the analysis in an attempt to minimise erroneous noise in the dataset. However, in some instances this was not deemed appropriate and therefore some flagged data points were included in the average Cq values.

Embryo Cq data

Average Cq values exported for each embryo sample for the three endogenous control genes and seven GOIs assessed are presented in Table 21 & Table 22 respectively, with their associated SD and the number of replicates included in the average. All average Cq values were calculated from 3 replicate Cq values with the exception of the *dd* samples in *Ldia2 3' UTR* and *Ldia2 ORF*.

In the *Ldia2 3' UTR* experiment all *dd* samples yielded very high Cq values (>32) with high levels of SD. Individual data points exhibiting substantially different values than the other two replicates were perceived as outliers and removed, resulting in the reduction of SD to <0.5 and the number of

replicates to two. This occurred in all *dd* samples except 11301, which maintained high SD at 0.804. Two samples were additionally flagged for multiple Tm peaks; two replicates of 11287 & one replicate of 11301. The removal of the one flagged replicate for 11301 cleared the flags for multiple Tm peaks. However, it was not considered appropriate to remove two wells from 11287, leaving only one representative Cq value, and as such one flag remains for the presence of multiple Tm peaks in 11287 (Table 22, Figure 15).

The *dd* samples in the *Ldia2 ORF* experiment showed higher Cq values than the *DD* or *Dd* samples (Table 22). Only one well was flagged by the software for multiple Tm peaks and subsequently omitted from the dataset leaving sample 11301 with only two replicates. However, when looking at the Tm plot for *Ldia2 ORF* the melt curves created for the *dd* samples exhibited a number of wide peaks indicating variation in the specific size of amplicons (Figure 15). Replicates showing very uneven Tm peaks were omitted from the dataset, which resulted in all *dd* samples being represented by only two technical replicates. Clean, single Tm peaks were generated from all of the *DD* and *Dd* samples (Figure 15). Additionally, sample 11283 bore high SD between its remaining replicates (0.524).

Table 21 Average Cq values (Cq) and associated standard deviation (SD) calculated from technical replicates (n) of 17 embryo samples and the ovotestis reference sample (OvoRef) for three endogenous control genes; *Lhis2a*, *Lube2* & *Lywhaz*. Including sample ID and genotype (Geno).[C] sample used as calibrator *amplification observed in negative controls

Embryo samples		*Lhis2a*			*Lube2*			*Lywhaz**		
ID	**Geno**	**Cq**	**SD**	**n**	**Cq**	**SD**	**n**	**Cq**	**SD**	**n**
11289	*DD*	20.93	0.04	3	24.14	0.13	3	23.20	0.10	3
11292	*DD*	20.61	0.02	3	23.61	0.07	3	22.69	0.03	3
11293	*DD*	20.36	0.02	3	23.91	0.03	3	22.96	0.02	3
11295[C]	*DD*	21.09	0.10	3	24.58	0.16	3	23.82	0.05	3
11297	*DD*	20.27	0.12	3	23.67	0.15	3	22.74	0.12	3
11298	*DD*	19.98	0.07	3	23.48	0.14	3	22.30	0.03	3
11358	*Dd*	20.12	0.01	3	23.19	0.08	3	22.12	0.06	3
11359	*Dd*	19.63	0.03	3	22.65	0.06	3	21.55	0.03	3
11360	*Dd*	19.35	0.01	3	22.38	0.05	3	21.35	0.05	3
11361	*Dd*	19.59	0.01	3	22.52	0.03	3	21.65	0.02	3
11363	*Dd*	18.98	0.04	3	21.87	0.06	3	20.94	0.02	3
11282	*dd*	20.09	0.05	3	23.12	0.02	3	22.23	0.04	3
11283	*dd*	21.36	0.06	3	24.73	0.05	3	23.49	0.03	3
11284	*dd*	20.24	0.02	3	23.36	0.09	3	22.39	0.01	3
11287	*dd*	21.02	0.05	3	24.37	0.13	3	23.57	0.12	3
11301	*dd*	21.20	0.08	3	24.73	0.11	3	23.63	0.07	3
11303	*dd*	20.74	0.01	3	23.87	0.06	3	22.86	0.02	3
OvoRef	*D/d*	21.65	0.04	3	21.47	0.02	3	19.57	0.06	3

Table 22 Average Cq values (Cq) and associated standard deviation (SD) calculated from technical replicates (n) of 17 embryo samples and the ovotestis reference sample (OvoRef) for seven GOIs. Including sample ID and genotype (Geno).[C] sample used as calibrator °multiple Tm peaks recorded †high SD observed between replicates.

Embryo samples		*Larp2/3 1a*			*Larp2/3 3*			*Ldia1 3' UTR*			*Ldia2 3' UTR*			*Ldia2 ORF*			*Lfat1*			*Lfry*		
ID	**Geno**	**Cq**	**SD**	**n**	**Cq**	**SD**	**n**	**Cq**	**SD**	**n**	**Cq**	**SD**	**n**	**Cq**	**SD**	**n**	**Cq**	**SD**	**n**	**Cq**	**SD**	**n**
11289	*DD*	24.00	0.18	3	20.56	0.03	3	23.23	0.08	3	26.44	0.18	3	22.29	0.08	3	28.05	0.16	3	22.30	0.11	3
11292	*DD*	23.28	0.08	3	20.01	0.19	3	22.12	0.15	3	25.65	0.09	3	21.57	0.02	3	26.96	0.04	3	21.79	0.09	3
11293	*DD*	23.49	0.02	3	20.31	0.31	3	22.84	0.11	3	26.35	0.07	3	22.13	0.14	3	27.22	0.11	3	22.29	0.05	3
11295[C]	*DD*	24.42	0.02	3	20.89	0.12	3	23.44	0.06	3	26.78	0.11	3	22.64	0.05	3	27.91	0.05	3	22.80	0.23	3
11297	*DD*	23.40	0.14	3	20.27	0.20	3	22.47	0.12	3	25.86	0.15	3	21.89	0.24	3	26.96	0.15	3	21.93	0.04	3
11298	*DD*	22.84	0.09	3	19.79	0.24	3	22.36	0.09	3	25.52	0.02	3	21.19	0.21	3	26.80	0.16	3	21.57	0.02	3
11358	*Dd*	22.61	0.11	3	19.74	0.13	3	22.45	0.08	3	26.68	0.14	3	22.64	0.08	3	26.77	0.15	3	21.72	0.11	3
11359	*Dd*	21.99	0.12	3	19.28	0.08	3	21.56	0.06	3	25.61	0.05	3	21.47	0.15	3	26.01	0.04	3	20.86	0.13	3
11360	*Dd*	22.00	0.19	3	19.39	0.07	3	21.47	0.04	3	25.55	0.10	3	21.83	0.33	3	25.64	0.03	3	20.63	0.12	3
11361	*Dd*	22.32	0.09	3	19.48	0.03	3	21.62	0.10	3	25.62	0.11	3	21.99	0.13	3	25.91	0.09	3	20.71	0.01	3
11363	*Dd*	21.69	0.10	3	18.96	0.20	3	20.77	0.09	3	25.05	0.03	3	21.15	0.13	3	25.00	0.07	3	20.15	0.14	3
11282	*dd*	22.92	0.02	3	19.34	0.08	3	21.83	0.06	3	32.51	0.08	3	26.67	0.18	2	26.67	0.08	3	21.68	0.17	3
11283	*dd*	24.53	0.13	3	21.01	0.22	3	23.43	0.14	3	36.73	0.43	2	27.26	0.52†	2	27.86	0.10	3	22.86	0.15	3
11284	*dd*	23.26	0.06	3	19.98	0.19	3	21.89	0.05	3	32.24	0.07	2	26.64	0.07	2	26.65	0.11	3	21.65	0.04	3
11287	*dd*	24.33	0.10	3	21.12	0.17	3	22.91	0.25	3	33.87	0.03°	2	27.24	0.10	2	27.45	0.27	3	22.63	0.17	3
11301	*dd*	24.37	0.09	3	21.28	0.11	3	23.72	0.17	3	35.65	0.80†	2	28.68	0.02	2	28.02	0.27	3	23.00	0.35	3
11303	*dd*	23.74	0.05	3	20.59	0.14	3	22.61	0.05	3	33.22	0.07	2	28.02	0.40	2	27.26	0.04	3	22.18	0.20	3
OvoRef	*D/d*	18.52	0.03	3	16.96	0.02	3	18.82	0.02	3	27.05	0.06	3	22.34	0.27	3	23.39	0.04	3	19.71	0.06	3

Foot Cq data

Average Cq values exported for each foot sample for the three endogenous control genes and 14 GOIs assessed are presented in Table 23 & Table 24 respectively, with their associated SD and the number of replicates included in the average. All averages were calculated from three replicate Cq values, with the exception of the four individuals described below.

Firstly, in the *Ldia2 3' UTR* experiment, one technical replicate of sample 11351 was flagged as an outlier and removed from analysis. Another sample 11347 was flagged for high SD across the three replicates. No one replicate appeared to represent an outlier and as such all three replicates were included in the average with high SD (0.643). It is important to note that the Cq values for all of the foot samples in *Ldia2 3' UTR* were fairly high (>28) and as such may be more prone to fluctuations in Cq.

In the *Ldia2 ORF* experiment, sample 11348 was flagged for high SD across its three replicates. The removal of a perceived outlier reduced SD to 0.358 and the average was calculated from the remaining two replicates.

In the *Lunc93a* experiment, one technical replicate of sample 11356 was flagged by the software as an outlier and subsequently removed from the dataset. Sample 11347 was flagged for high SD across the three replicates. The removal of one perceived outlier reduced SD to 0.022 and the average was calculated from the remaining two replicates.

Table 23 Average Cq values (Cq) and associated standard deviation (SD) calculated from technical replicates (n) of 10 foot samples and the ovotestis reference sample (OvoRef) for three endogenous control genes; *Lhis2a*, *Lube2* & *Lywhaz*. Including sample ID and genotype (Geno).[C] sample used as calibrator, *amplification observed in negative controls

Foot samples		*Lhis2a*			*Lube2*			*Lywhaz**		
ID	**Geno**	**Cq**	**SD**	**n**	**Cq**	**SD**	**n**	**Cq**	**SD**	**n**
11347	*DD*	26.05	0.04	3	27.12	0.06	3	24.14	0.05	3
11350[C]	*DD*	21.88	0.08	3	22.96	0.04	3	20.58	0.03	3
11351	*DD*	24.06	0.03	3	25.53	0.04	3	22.67	0.04	3
11352	*DD*	22.60	0.02	3	23.94	0.02	3	21.30	0.01	3
11357	*DD*	21.94	0.01	3	23.16	0.04	3	20.90	0.04	3
11348	*dd*	23.93	0.02	3	24.56	0.11	3	22.12	0.06	3
11349	*dd*	23.35	0.06	3	24.32	0.05	3	22.17	0.02	3
11353	*dd*	23.99	0.03	3	24.60	0.06	3	22.29	0.05	3
11354	*dd*	22.90	0.02	3	23.40	0.05	3	21.30	0.02	3
11356	*dd*	23.80	0.01	3	24.90	0.01	3	22.27	0.04	3
OvoRef	*D/d*	21.65	0.04	3	21.47	0.02	3	19.57	0.06	3

Table 24 Average Cq values (Cq) and associated standard deviation (SD) calculated from technical replicates (n) of 10 foot samples and the ovotestis reference sample (OvoRef) for 14 GOIs. Including sample ID and genotype (Geno).[C] sample used as calibrator *amplification observed in negative controls †high SD observed between replicates.

Foot, GOI 1-7		***Larp2/3 1a***			***Larp2/3 3a****			***Ldia1 3' UTR***			***Ldia2 3' UTR***			***Ldia2 ORF***			***Lfat1***			***Lfry***		
ID	***Geno***	**Cq**	**SD**	***n***	**Cq**	**SD**	***n***	**Cq**	**SD**	***n***	**Cq**	**SD**	***n***	**Cq**	**SD**	***n***	**Cq**	**SD**	***n***	**Cq**	**SD**	***n***
11347	*DD*	23.76	0.12	3	23.26	0.14	3	23.42	0.19	3	30.60	0.64†	3	26.23	0.14	3	27.58	0.12	3	23.98	0.13	3
11350[C]	*DD*	20.87	0.04	3	18.64	0.20	3	19.85	0.03	3	28.83	0.24	3	24.49	0.12	3	24.92	0.02	3	21.48	0.08	3
11351	*DD*	22.27	0.06	3	21.38	0.15	3	21.71	0.19	3	30.06	0.01	2	25.74	0.39	3	26.21	0.05	3	22.73	0.08	3
11352	*DD*	21.31	0.07	3	19.58	0.26	3	20.93	0.06	3	29.39	0.21	3	25.11	0.37	3	25.82	0.14	3	22.35	0.06	3
11357	*DD*	20.56	0.00	3	19.15	0.21	3	20.18	0.05	3	28.28	0.08	3	24.35	0.09	3	24.68	0.06	3	21.35	0.04	3
11348	*dd*	21.54	0.04	3	20.67	0.08	3	21.27	0.07	3	29.65	0.09	3	24.88	0.36	2	25.54	0.08	3	22.72	0.05	3
11349	*dd*	21.65	0.05	3	20.26	0.30	3	21.39	0.21	3	29.56	0.35	3	25.21	0.04	3	26.04	0.06	3	22.76	0.05	3
11353	*dd*	21.63	0.09	3	20.67	0.21	3	21.54	0.11	3	29.20	0.49	3	24.56	0.31	3	25.68	0.07	3	22.53	0.13	3
11354	*dd*	21.05	0.08	3	19.53	0.10	3	20.70	0.03	3	28.78	0.28	3	24.35	0.09	3	24.86	0.02	3	21.63	0.09	3
11356	*dd*	21.68	0.05	3	20.83	0.09	3	21.73	0.10	3	29.24	0.27	3	24.63	0.18	3	25.87	0.04	3	22.31	0.10	3
OvoRef	*D/d*	18.52	0.03	3	16.96	0.02	3	18.82	0.02	3	27.05	0.06	3	22.34	0.27	3	23.39	0.04	3	19.71	0.06	3
Foot, GOI 8-14		***Lcol11 2a****			***Lmhc****			***Lmhc nm****			***Lmyo5a****			***Lmyo18a****			***Lstau****			***Lunc93a***		
ID	***Geno***	**Cq**	**SD**	***n***	**Cq**	**SD**	***n***	**Cq**	**SD**	***n***	**Cq**	**SD**	***n***	**Cq**	**SD**	***n***	**Cq**	**SD**	***n***	**Cq**	**SD**	***n***
11347	*DD*	22.21	0.18	3	21.22	0.11	3	18.90	0.03	3	21.80	0.06	3	26.78	0.05	3	21.58	0.03	3	25.64	0.02	2
11350[C]	*DD*	20.58	0.10	3	18.54	0.05	3	16.05	0.11	3	20.42	0.02	3	24.61	0.10	3	19.82	0.03	3	25.98	0.31	3
11351	*DD*	20.42	0.05	3	19.61	0.06	3	17.17	0.15	3	20.70	0.05	3	25.78	0.13	3	20.48	0.03	3	25.68	0.15	3
11352	*DD*	20.69	0.11	3	19.34	0.05	3	16.87	0.07	3	20.90	0.03	3	26.18	0.02	3	20.39	0.06	3	25.57	0.39	3
11357	*DD*	19.57	0.07	3	18.75	0.04	3	16.84	0.07	3	20.13	0.24	3	23.94	0.19	3	19.68	0.04	3	24.80	0.18	3
11348	*dd*	19.63	0.17	3	18.84	0.09	3	17.13	0.33	3	20.69	0.16	3	24.70	0.08	3	20.14	0.14	3	24.71	0.03	3
11349	*dd*	19.55	0.14	3	18.40	0.06	3	17.37	0.20	3	21.41	0.04	3	25.44	0.12	3	20.49	0.04	3	25.09	0.23	3
11353	*dd*	19.61	0.03	3	19.20	0.11	3	17.40	0.21	3	20.21	0.02	3	25.42	0.15	3	20.14	0.10	3	24.84	0.23	3
11354	*dd*	19.53	0.03	3	19.19	0.05	3	17.39	0.11	3	20.01	0.02	3	24.50	0.06	3	20.16	0.19	3	25.46	0.28	3
11356	*dd*	20.60	0.04	3	19.48	0.15	3	17.67	0.05	3	20.43	0.08	3	24.66	0.08	3	20.19	0.07	3	25.04	0.30	3
OvoRef	*D/d*	20.30	0.07	3	19.60	0.06	3	14.72	0.07	3	19.60	0.13	3	23.13	0.32	3	17.97	0.05	3	20.26	0.01	2

Ovotestis Cq data

Average Cq values exported for each ovotestis sample for the three endogenous control genes assessed are presented in Table 25. Due to the larger amount of samples in the ovotestis experiment the average Cq data for the 14 GOIs assessed are split across three tables (Table 26, Table 27, Table 28). Each average Cq value is presented with the associated SD and number of replicates included in the average. All individual sample averages were calculated from two replicate Cq values, with the exception of one individual in the *Lrpl14* experiment.

Because the ovotestis samples were represented by only two technical replicates it was generally not possible to identify outliers. Therefore, a number of sample averages were included with high SD. However, in the *Lrpl14* experiment, a technical replicate of sample 11347 clearly exhibited a substandard reaction evident from the amplification curve (data not shown) and was subsequently removed from analysis. The Cq value for 11347 in *Lrpl14* consequently only represents one reaction and therefore has no SD (Table 25).

The OvoRef sample averages included Cq values from four replicates and as such any outliers were easily identified. The only occurrence of an outlier in the OvoRef sample was observed in the *Lmhc* experiment and subsequently removed (Table 27).

Table 25 Average Cq values (Cq) and associated standard deviation (SD) calculated from technical replicates (n) of 36 ovotestis samples and the reference sample (OvoRef) for three endogenous control genes; *Lhis2a*, *Lube2* & *Lrpl14*. Including sample ID and genotype (Geno). [C] sample used as calibrator *amplification observed in negative controls †high SD observed between replicates.

Ovotestis		*Lhis2a**			*Lube2*			*Lrpl14**		
ID	**Geno**	**Cq**	**SD**	**n**	**Cq**	**SD**	**n**	**Cq**	**SD**	**n**
8515	*DD*	23.18	0.04	2	23.50	0.11	2	20.74	0.28	2
8548[C]	*DD*	23.18	0.16	2	25.20	0.24	2	21.88	0.07	2
8582	*DD*	23.99	0.04	2	24.34	0.21	2	21.42	0.15	2
8583	*DD*	23.06	0.03	2	23.66	0.10	2	21.16	0.17	2
9014	*DD*	23.93	0.03	2	24.87	0.04	2	21.51	0.09	2
10627	*DD*	23.14	0.11	2	23.88	0.23	2	20.23	0.30	2
10633	*DD*	22.93	0.00	2	23.34	0.05	2	19.09	0.27	2
10636	*DD*	22.57	0.02	2	23.56	0.04	2	19.21	0.25	2
10638	*DD*	23.92	0.03	2	24.45	0.08	2	20.11	0.06	2
11347	*DD*	22.36	0.03	2	22.74	0.27	2	21.09	n/a	1
11350	*DD*	21.54	0.03	2	21.88	0.07	2	19.88	0.10	2
11351	*DD*	20.52	0.01	2	21.73	0.01	2	19.68	0.41	2
11352	*DD*	23.17	0.01	2	23.82	0.01	2	21.34	0.01	2
11357	*DD*	21.87	0.00	2	22.27	0.08	2	20.01	0.41	2
8554	*Dd*	23.84	0.00	2	24.91	0.00	2	21.24	0.00	2
8555	*Dd*	23.24	0.01	2	23.71	0.05	2	20.93	0.05	2
8559	*Dd*	23.78	0.01	2	23.60	0.02	2	20.55	0.04	2
8562	*Dd*	21.80	0.02	2	23.92	0.25	2	20.40	0.14	2
9013	*Dd*	21.98	0.03	2	23.45	0.04	2	20.81	0.17	2
10622	*Dd*	23.46	0.11	2	23.54	0.03	2	20.39	0.17	2
10629	*Dd*	23.41	0.03	2	24.01	0.00	2	20.07	0.21	2
10639	*Dd*	23.39	0.01	2	24.06	0.09	2	19.98	0.01	2
8806	*dd*	22.38	0.01	2	23.65	0.04	2	20.80	0.14	2
8808	*dd*	22.02	0.03	2	23.42	0.03	2	20.00	0.01	2
8996	*dd*	24.10	0.05	2	24.79	0.15	2	20.94	0.14	2
9005	*dd*	22.59	0.03	2	23.44	0.07	2	20.89	0.12	2
9007	*dd*	22.97	0.00	2	23.75	0.03	2	20.92	0.07	2
10626	*dd*	21.28	0.05	2	21.94	0.00	2	18.42	0.15	2
10630	*dd*	23.27	0.04	2	23.63	0.10	2	19.69	0.58†	2
10640	*dd*	24.86	0.10	2	25.26	0.08	2	20.80	0.06	2
10642	*dd*	22.28	0.04	2	24.40	0.09	2	20.18	0.42	2
11348	*dd*	22.46	0.01	2	23.00	0.09	2	20.95	0.03	2
11349	*dd*	24.09	0.03	2	25.17	0.08	2	22.26	0.42	2
11353	*dd*	22.03	0.04	2	22.22	0.09	2	20.93	0.05	2
11354	*dd*	20.80	0.06	2	21.87	0.12	2	20.15	0.10	2
11356	*dd*	21.72	0.05	2	21.82	0.00	2	20.34	0.05	2
OvoRef	*D/d*	20.55	0.05	4	21.39	0.08	4	18.88	0.39	4

Table 26 Average Cq values (Cq) and associated standard deviation (SD) calculated from technical replicates (n) of 36 ovotestis samples and the ovotestis reference sample (OvoRef) for five GOIs (*Larp2/3 1a, Larp2/3 3, Ldia1 3' UTR, Ldia2 3' UTR, Ldia2 ORF*). Including sample ID and genotype (Geno).[C] sample used as calibrator *amplification observed in negative controls †high SD observed between replicates.

Ovo, GOI 1-5		*Larp2/3 1a*			*Larp2/3 3**			*Ldia1 3'UTR*			*Ldia2 3' UTR*			*Ldia2 ORF**		
ID	**Geno**	**Cq**	**SD**	**n**	**Cq**	**SD**	**n**	**Cq**	**SD**	**n**	**Cq**	**SD**	**n**	**Cq**	**SD**	**n**
8515	*DD*	24.95	0.00	2	20.31	0.06	2	21.35	0.07	2	28.68	0.01	2	29.27	0.11	2
8548[C]	*DD*	26.38	0.04	2	21.45	0.15	2	23.08	0.17	2	30.18	0.05	2	30.27	0.16	2
8582	*DD*	25.68	0.34	2	21.32	0.12	2	22.32	0.10	2	29.17	0.56†	2	29.67	0.19	2
8583	*DD*	25.42	0.08	2	20.68	0.13	2	21.58	0.09	2	29.34	0.30	2	29.22	0.21	2
9014	*DD*	26.53	0.07	2	22.40	0.07	2	23.17	0.09	2	28.85	0.11	2	29.91	0.03	2
10627	*DD*	24.87	0.40	2	21.55	0.20	2	22.29	0.09	2	29.22	0.33	2	29.49	0.16	2
10633	*DD*	23.72	0.44	2	20.63	0.05	2	21.60	0.05	2	29.62	0.22	2	29.91	0.37	2
10636	*DD*	24.57	0.13	2	20.79	0.00	2	21.87	0.07	2	27.95	0.19	2	28.66	0.03	2
10638	*DD*	25.24	0.03	2	21.84	0.10	2	22.58	0.11	2	29.46	0.15	2	29.91	0.23	2
11347	*DD*	25.18	0.31	2	19.42	0.11	2	21.12	0.12	2	28.28	0.08	2	29.14	0.48	2
11350	*DD*	24.64	0.27	2	18.55	0.05	2	20.53	0.01	2	27.39	0.01	2	28.12	0.01	2
11351	*DD*	24.15	0.20	2	18.65	0.02	2	20.12	0.07	2	27.04	0.13	2	28.06	0.10	2
11352	*DD*	26.06	0.02	2	20.68	0.02	2	22.00	0.18	2	28.18	0.05	2	28.90	0.21	2
11357	*DD*	24.99	0.88†	2	19.01	0.00	2	20.95	0.03	2	27.77	0.03	2	28.66	0.03	2
8554	*Dd*	26.02	0.26	2	21.38	0.05	2	22.70	0.09	2	29.61	0.20	2	29.71	0.10	2
8555	*Dd*	25.45	0.22	2	20.85	0.07	2	21.60	0.04	2	29.02	0.26	2	29.51	0.08	2
8559	*Dd*	25.27	0.27	2	20.94	0.04	2	21.59	0.02	2	29.49	0.08	2	30.23	0.13	2
8562	*Dd*	25.16	0.36	2	20.06	0.17	2	21.82	0.12	2	28.66	0.14	2	29.27	0.23	2
9013	*Dd*	25.69	0.01	2	21.33	0.14	2	22.78	0.10	2	27.93	0.02	2	28.75	0.30	2
10622	*Dd*	24.34	0.46	2	21.66	0.20	2	21.97	0.02	2	30.02	0.22	2	30.20	0.37	2
10629	*Dd*	24.43	0.64†	2	20.85	0.06	2	21.88	0.03	2	30.15	0.22	2	30.92	0.03	2
10639	*Dd*	25.95	0.19	2	23.12	0.22	2	23.48	0.02	2	28.99	0.38	2	29.61	0.24	2
8806	*dd*	25.44	0.04	2	20.60	0.13	2	21.87	0.10	2	28.89	0.12	2	29.47	0.10	2
8808	*dd*	23.97	0.01	2	20.64	0.42	2	21.59	0.06	2	28.98	0.53†	2	29.47	0.12	2
8996	*dd*	25.76	0.03	2	22.02	0.02	2	22.39	0.06	2	30.24	0.04	2	30.14	0.33	2
9005	*dd*	25.46	0.02	2	20.71	0.03	2	21.22	0.08	2	29.08	0.04	2	29.73	0.21	2
9007	*dd*	25.54	0.27	2	21.38	0.00	2	21.95	0.04	2	29.77	0.07	2	30.15	0.07	2
10626	*dd*	23.37	0.22	2	19.27	0.10	2	20.54	0.07	2	28.35	0.02	2	29.00	0.08	2
10630	*dd*	24.88	0.57†	2	21.49	0.01	2	22.16	0.24	2	30.23	0.03	2	30.52	0.21	2
10640	*dd*	25.44	0.20	2	22.86	0.05	2	23.43	0.00	2	31.79	0.27	2	31.82	0.02	2
10642	*dd*	24.03	0.81†	2	20.11	0.18	2	22.64	0.01	2	30.42	0.42	2	30.34	0.34	2
11348	*dd*	25.53	0.55†	2	20.19	0.01	2	21.10	0.13	2	28.74	0.31	2	29.33	0.34	2
11349	*dd*	27.54	0.02	2	21.88	0.01	2	22.65	0.11	2	30.36	0.27	2	30.69	0.06	2
11353	*dd*	24.94	0.07	2	19.35	0.01	2	20.58	0.04	2	28.08	0.11	2	28.56	0.15	2
11354	*dd*	25.05	0.06	2	19.94	0.06	2	20.95	0.06	2	26.74	0.31	2	27.70	0.14	2
11356	*dd*	25.00	0.14	2	18.71	0.06	2	20.01	0.11	2	28.65	0.29	2	28.98	0.11	2
OvoRef	*D/d*	23.89	0.32	4	18.12	0.12	4	19.72	0.07	4	26.96	0.10	4	27.72	0.06	4

Table 27 Average Cq values (Cq) and associated standard deviation (SD) calculated from technical replicates (n) of 36 ovotestis samples and the ovotestis reference sample (OvoRef) for five GOIs (*Lfat1, Lfry, Lcol11a 2/1, Lmhc, Lmhc nm*), including sample ID and genotype (Geno).[C] sample used as calibrator *amplification observed in negative controls †high SD observed between replicates.

Ovo, GOI 6-10		*Lfat1*			*Lfry*			*Lcoll11 2a**			*Lmhc*			*Lmhc nm*		
ID	**Geno**	**Cq**	**SD**	**n**	**Cq**	**SD**	**n**	**Cq**	**SD**	**n**	**Cq**	**SD**	**n**	**Cq**	**SD**	**n**
8515	*DD*	20.96	0.04	2	25.82	0.09	2	24.04	0.21	2	19.98	0.01	2	16.96	0.01	2
8548[C]	*DD*	22.45	0.19	2	27.27	0.12	2	25.19	0.18	2	21.06	0.29	2	17.76	0.18	2
8582	*DD*	22.16	0.05	2	26.54	0.00	2	23.90	0.01	2	20.00	0.07	2	17.55	0.09	2
8583	*DD*	21.60	0.07	2	25.83	0.14	2	22.60	0.03	2	20.72	0.20	2	17.25	0.13	2
9014	*DD*	23.12	0.28	2	27.69	0.22	2	25.04	0.18	2	21.88	0.03	2	18.05	0.17	2
10627	*DD*	22.25	0.07	2	26.52	0.05	2	23.79	0.11	2	19.69	0.03	2	17.75	0.04	2
10633	*DD*	22.43	0.19	2	26.20	0.10	2	21.65	0.02	2	18.58	0.01	2	17.11	0.13	2
10636	*DD*	22.30	0.09	2	26.65	0.15	2	23.56	0.03	2	19.64	0.10	2	16.77	0.03	2
10638	*DD*	23.04	0.13	2	27.25	0.07	2	24.70	0.02	2	19.72	0.06	2	18.10	0.11	2
11347	*DD*	21.36	0.07	2	25.80	0.03	2	22.53	0.04	2	18.97	0.06	2	16.47	0.04	2
11350	*DD*	20.49	0.09	2	24.72	0.06	2	21.88	0.09	2	20.13	0.35	2	15.72	0.14	2
11351	*DD*	20.60	0.07	2	24.75	0.02	2	20.45	0.02	2	17.96	0.28	2	15.48	0.08	2
11352	*DD*	22.96	0.22	2	26.54	0.03	2	22.26	0.02	2	19.78	0.01	2	17.32	0.16	2
11357	*DD*	20.63	0.03	2	25.54	0.02	2	23.21	0.22	2	20.22	0.30	2	16.25	0.06	2
8554	*Dd*	22.55	0.01	2	26.89	0.05	2	23.61	0.20	2	20.81	0.13	2	17.85	0.07	2
8555	*Dd*	21.66	0.01	2	25.93	0.01	2	25.17	0.02	2	22.05	0.41	2	17.79	0.02	2
8559	*Dd*	22.22	0.07	2	26.52	0.03	2	24.38	0.05	2	21.24	0.06	2	17.75	0.03	2
8562	*Dd*	21.76	0.03	2	26.44	0.03	2	24.30	0.04	2	21.67	0.17	2	17.26	0.09	2
9013	*Dd*	22.53	0.17	2	26.52	0.11	2	23.39	0.00	2	20.90	0.11	2	16.81	0.12	2
10622	*Dd*	22.57	0.06	2	26.48	0.00	2	24.33	0.02	2	19.56	0.04	2	17.10	0.11	2
10629	*Dd*	22.96	0.31	2	26.79	0.16	2	22.84	0.13	2	19.62	1.10†	2	16.97	0.03	2
10639	*Dd*	23.26	0.02	2	27.39	0.12	2	24.89	0.03	2	19.83	0.07	2	17.76	0.15	2
8806	*dd*	22.21	0.08	2	26.16	0.05	2	23.04	0.25	2	20.51	0.01	2	17.17	0.08	2
8808	*dd*	21.79	0.02	2	25.75	0.05	2	22.57	0.20	2	19.63	0.10	2	16.90	0.02	2
8996	*dd*	22.48	0.10	2	26.98	0.03	2	24.26	0.00	2	22.08	0.31	2	17.75	0.08	2
9005	*dd*	21.70	0.08	2	26.19	0.12	2	23.79	0.15	2	21.83	0.04	2	17.29	0.00	2
9007	*dd*	22.09	0.22	2	26.57	0.10	2	22.92	0.00	2	21.51	0.02	2	17.14	0.22	2
10626	*dd*	20.76	0.05	2	24.91	0.05	2	21.48	0.00	2	17.80	0.01	2	14.95	0.03	2
10630	*dd*	22.48	0.01	2	26.95	0.00	2	24.12	0.21	2	19.26	0.00	2	17.32	0.16	2
10640	*dd*	23.65	0.06	2	28.63	0.07	2	24.82	0.12	2	19.66	0.13	2	18.78	0.05	2
10642	*dd*	23.10	0.19	2	27.82	0.16	2	25.03	0.08	2	20.28	0.01	2	16.49	0.17	2
11348	*dd*	22.06	0.11	2	25.94	0.03	2	22.01	0.34	2	20.29	0.13	2	16.85	0.06	2
11349	*dd*	23.11	0.11	2	27.45	0.18	2	21.92	0.42	2	21.50	0.01	2	17.89	0.06	2
11353	*dd*	21.10	0.10	2	25.42	0.20	2	21.73	0.05	2	21.23	0.42	2	15.78	0.16	2
11354	*dd*	21.59	0.05	2	24.99	0.03	2	22.34	0.04	2	20.50	0.21	2	15.68	0.04	2
11356	*dd*	20.84	0.07	2	25.05	0.01	2	23.11	0.32	2	19.34	0.05	2	15.85	0.02	2
OvoRef	*D/d*	19.99	0.15	4	24.27	0.03	4	20.95	0.09	4	17.49	0.37	3	14.72	0.14	4

Table 28 Average Cq values (Cq) and associated standard deviation (SD) calculated from technical replicates (n) of 36 ovotestis samples and the ovotestis reference sample (OvoRef) for four GOIs (*Lmyo5a, Lmyo18a, Lstau, Lunc93a*), including sample ID and genotype (Geno).[C] sample used as calibrator *amplification observed in negative controls †high SD observed between replicates.

Ovo, GOI 11-14		*Lmyo5a*			*Lmyo18a*			*Lstau*			*Lunc93a**		
ID	**Geno**	**Cq**	**SD**	**n**	**Cq**	**SD**	**n**	**Cq**	**SD**	**n**	**Cq**	**SD**	**n**
8515	*DD*	25.71	0.02	2	25.92	0.06	2	23.69	0.49	2	22.23	0.05	2
8548[C]	*DD*	26.37	0.01	2	26.39	0.15	2	23.94	0.07	2	22.50	0.20	2
8582	*DD*	25.78	0.03	2	26.77	0.16	2	23.01	0.48	2	21.73	0.07	2
8583	*DD*	25.51	0.02	2	25.57	0.10	2	22.03	0.12	2	20.73	0.04	2
9014	*DD*	26.04	0.09	2	27.27	0.17	2	21.80	0.23	2	24.51	0.21	2
10627	*DD*	25.94	0.06	2	27.27	0.01	2	22.64	0.12	2	22.87	0.11	2
10633	*DD*	25.21	0.03	2	26.87	0.11	2	21.75	0.05	2	21.62	0.03	2
10636	*DD*	25.65	0.26	2	26.40	0.08	2	21.33	0.00	2	21.58	0.05	2
10638	*DD*	26.28	0.01	2	28.17	0.04	2	22.95	0.04	2	23.41	0.02	2
11347	*DD*	25.31	0.05	2	26.22	0.02	2	22.94	0.19	2	21.51	0.18	2
11350	*DD*	24.19	0.01	2	25.29	0.09	2	22.37	0.18	2	20.88	0.03	2
11351	*DD*	24.08	0.10	2	25.14	0.08	2	21.91	0.39	2	21.38	0.07	2
11352	*DD*	26.08	0.08	2	27.26	0.10	2	23.37	0.42	2	22.74	0.11	2
11357	*DD*	24.88	0.03	2	26.18	0.06	2	23.09	0.89†	2	21.56	0.02	2
8554	*Dd*	25.80	0.04	2	26.55	0.01	2	23.38	0.21	2	22.52	0.28	2
8555	*Dd*	25.92	0.18	2	25.26	0.05	2	22.42	0.33	2	21.74	0.09	2
8559	*Dd*	25.85	0.09	2	26.21	0.05	2	22.46	0.38	2	22.55	0.39	2
8562	*Dd*	25.07	0.11	2	26.06	0.02	2	22.73	0.28	2	22.22	0.02	2
9013	*Dd*	24.90	0.21	2	25.52	0.03	2	19.43	0.33	2	22.68	0.17	2
10622	*Dd*	25.38	1.63†	2	27.02	0.10	2	21.77	0.20	2	22.65	0.04	2
10629	*Dd*	26.24	0.14	2	26.85	0.08	2	22.60	0.01	2	22.42	0.07	2
10639	*Dd*	26.08	0.05	2	27.10	0.27	2	21.32	0.22	2	24.79	0.09	2
8806	*dd*	25.13	0.07	2	25.84	0.04	2	21.09	0.32	2	22.24	0.09	2
8808	*dd*	25.07	0.15	2	25.57	0.11	2	21.24	0.47	2	21.68	0.03	2
8996	*dd*	26.04	0.14	2	26.88	0.09	2	23.13	0.33	2	23.23	0.05	2
9005	*dd*	25.73	0.01	2	26.56	0.06	2	23.05	1.09†	2	23.17	0.39	2
9007	*dd*	25.29	0.02	2	26.06	0.16	2	21.01	0.10	2	22.48	0.00	2
10626	*dd*	24.52	0.04	2	24.83	0.11	2	20.76	0.01	2	20.62	0.15	2
10630	*dd*	26.33	0.06	2	27.80	0.08	2	22.63	0.02	2	22.55	0.04	2
10640	*dd*	27.37	0.18	2	27.99	0.11	2	24.47	0.06	2	23.52	0.27	2
10642	*dd*	26.35	0.16	2	27.20	0.00	2	22.41	0.05	2	23.52	0.03	2
11348	*dd*	25.00	0.09	2	26.26	0.01	2	22.34	0.15	2	22.17	0.49	2
11349	*dd*	26.25	0.08	2	27.55	0.03	2	24.04	0.10	2	22.47	0.17	2
11353	*dd*	24.05	0.19	2	25.93	0.04	2	21.86	0.14	2	22.03	0.25	2
11354	*dd*	23.26	0.39	2	25.39	0.21	2	19.36	0.11	2	23.33	0.12	2
11356	*dd*	24.75	0.07	2	25.85	0.02	2	22.71	0.62†	2	21.14	0.23	2
OvoRef	*D/d*	24.05	0.06	4	24.55	0.11	4	20.37	0.09	4	20.05	0.21	4

Genotype Analysis

All normalised relative quantity (NRQ) values within the genotype analyses are relative to an individual *DD* calibrator sample of the same tissue type. Additionally, the relative expression data is mostly presented and analysed using NRQ values log transformed to the base 10 (LOG NRQ). This is important to note when interpreting the observed log-fold changes in gene expression.

The relative quantities (RQ) of the three endogenous control genes and the subsequent geometric mean RQ used to calculate the NRQ values in the embryo, foot and ovotestis tissue analyses are presented in Table 29, Table 30 & Table 31 respectively. A summary of the genotypic group means of NRQs for each GOI assessed within the three separate genotype analyses is presented in Table 32. No statistical analyses were performed on the non-transformed NRQ values, therefore only the geometric group mean is presented.

The LOG NRQs for each of the embryo tissue samples for the seven GOIs assessed are presented in Table 33. The LOG NRQs for each of the foot tissue samples for the 14 GOIs assessed are presented in Table 34. Due to the larger number of samples included in the ovotestis experiment, the LOG NRQs for each of the ovotestis tissue samples for the 14 GOIs assessed are presented across three tables (Table 35, Table 36, Table 37). Each of these tables presents the individual sample count data and the genotypic group means used within the statistical analyses presented in Table 38, Table 39 and the boxplot graphs (Figure 16 - Figure 29). Histogram plots of the ovotestis genotype analysis data and summary statistics of each genotype analysis are presented in the SI (S10).

Table 29 Relative quantity (RQ) values per embryo sample for each of the three endogenous control genes assessed (*Lhis2a*, *Lube2* & *Lywhaz*) and resulting geometric mean (GeoMean), including sample ID (ID) and genotype (Geno). [C] sample used as calibrator.

Embryo sample description		RQ values			
ID	**Geno**	***Lhis2a***	***Lube2***	***Lywhaz***	**GeoMean**
11289	*DD*	1.113	1.339	1.491	1.305
11292	*DD*	1.376	1.887	2.083	1.755
11293	*DD*	1.622	1.549	1.752	1.639
11295[C]	*DD*	1	1	1	1
11297	*DD*	1.716	1.819	2.018	1.847
11298	*DD*	2.082	2.061	2.683	2.258
11358	*Dd*	1.894	2.484	3.018	2.422
11359	*Dd*	2.639	3.538	4.388	3.447
11360	*Dd*	3.168	4.213	4.972	4.048
11361	*Dd*	2.706	3.854	4.090	3.494
11363	*Dd*	4.055	5.896	6.528	5.384
11282	*dd*	1.943	2.600	2.816	2.423
11283	*dd*	0.832	0.905	1.236	0.976
11284	*dd*	1.754	2.222	2.531	2.145
11287	*dd*	1.046	1.146	1.172	1.120
11301	*dd*	0.928	0.908	1.126	0.982
11303	*dd*	1.257	1.595	1.864	1.552

Table 30 Relative quantity (RQ) values per foot sample for each of the three endogenous control genes assessed (*Lhis2a*, *Lube2* & *Lywhaz*) and resulting geometric mean (GeoMean), including sample ID (ID) and genotype (Geno). [C] sample used as calibrator.

Foot sample description		RQ values			
ID	**Geno**	***Lhis2a***	***Lube2***	***Lywhaz***	**GeoMean**
11347	*DD*	0.062	0.066	0.098	0.074
11350[C]	*DD*	1	1	1	1
11351	*DD*	0.234	0.186	0.257	0.223
11352	*DD*	0.620	0.528	0.629	0.590
11357	*DD*	0.958	0.875	0.815	0.881
11348	*dd*	0.255	0.352	0.368	0.321
11349	*dd*	0.376	0.411	0.356	0.381
11353	*dd*	0.246	0.341	0.328	0.302
11354	*dd*	0.506	0.749	0.628	0.620
11356	*dd*	0.278	0.281	0.334	0.297

Table 31 Relative quantity (RQ) values per ovotestis sample for each of the three endogenous control genes assessed (*Lhis2a*, *Lube2* & *Lrpl14*) and resulting geometric mean (GeoMean), including sample ID (ID) and genotype (Geno). [C] sample used as calibrator.

Ovotestis sample description		RQ values			
ID	**Geno**	***Lhis2a***	***Lube2***	***Lrpl14***	**GeoMean**
8515	*DD*	0.999	3.036	2.094	1.852
8548[C]	*DD*	1	1	1	1
8582	*DD*	0.585	1.754	1.352	1.115
8583	*DD*	1.085	2.742	1.591	1.679
9014	*DD*	0.606	1.238	1.272	0.985
10627	*DD*	1.022	2.369	2.909	1.917
10633	*DD*	1.182	3.375	6.068	2.893
10636	*DD*	1.501	2.928	5.604	2.910
10638	*DD*	0.611	1.628	3.139	1.462
11347	*DD*	1.721	4.981	1.668	2.427
11350	*DD*	2.960	8.765	3.632	4.550
11351	*DD*	5.835	9.663	4.154	6.164
11352	*DD*	1.005	2.458	1.417	1.518
11357	*DD*	2.382	6.786	3.345	3.781
8554	*Dd*	0.643	1.208	1.514	1.056
8555	*Dd*	0.958	2.645	1.854	1.674
8559	*Dd*	0.668	2.840	2.370	1.651
8562	*Dd*	2.498	2.313	2.607	2.470
9013	*Dd*	2.215	3.134	1.998	2.403
10622	*Dd*	0.831	2.958	2.618	1.860
10629	*Dd*	0.857	2.183	3.212	1.818
10639	*Dd*	0.869	2.106	3.411	1.841
8806	*dd*	1.703	2.749	2.009	2.111
8808	*dd*	2.156	3.203	3.377	2.857
8996	*dd*	0.543	1.305	1.835	1.092
9005	*dd*	1.481	3.153	1.892	2.067
9007	*dd*	1.148	2.584	1.865	1.769
10626	*dd*	3.537	8.435	9.314	6.526
10630	*dd*	0.941	2.791	4.106	2.210
10640	*dd*	0.327	0.962	2.014	0.859
10642	*dd*	1.819	1.686	3.001	2.096
11348	*dd*	1.615	4.222	1.820	2.315
11349	*dd*	0.545	1.018	0.786	0.758
11353	*dd*	2.138	7.033	1.849	3.029
11354	*dd*	4.862	8.820	3.065	5.084
11356	*dd*	2.631	9.122	2.711	4.022

Table 32 Normalised relative quantities (NRQ) of each GOI, presented as a geometric mean per genotypic group (Geno) within the genotype analysis for each tissue, including number of samples within each group (n).

Genotype Analysis			NRQ													
Tissue	**Geno**	**n**	***Larp2/31a***	***Larp2/33***	***Ldia1 3' UTR***	***Ldia2 3' UTR***	***Ldia2 ORF***	***Lfat1***	***Lfry***	***Lcol11 2a***	***Lmhc***	***Lmhcnm***	***Lmyo5a***	***Lmyo18a***	***Lstau***	***Lunc93a***
Embryo	*DD*	6	1.064	0.878	1.019	0.984	1.002	0.906	0.976	x	x	x	x	x	x	x
	Dd	5	1.127	0.645	0.987	0.554	0.477	0.953	0.960	x	x	x	x	x	x	x
	dd	6	0.985	0.843	1.132	0.006	0.029	0.999	0.937	x	x	x	x	x	x	x
Foot	*DD*	5	1.508	0.943	1.010	1.755	1.626	1.477	1.474	2.407	1.411	1.249	2.022	1.498	1.766	3.515
	dd	5	1.837	0.993	0.983	2.025	2.317	1.798	1.535	4.507	1.998	1.131	2.490	2.189	2.069	5.220
Ovotestis	*DD*	14	1.005	0.814	1.186	1.288	0.956	0.674	0.924	1.697	1.013	0.765	0.849	0.451	1.148	0.631
	Dd	8	1.092	0.617	1.000	1.031	0.773	0.561	0.841	1.107	0.698	0.699	0.894	0.584	2.027	0.487
	dd	14	0.970	0.714	1.206	0.797	0.656	0.571	0.811	1.731	0.694	0.820	0.878	0.447	1.499	0.471

Table 33 Log-transformed normalised relative quantities (LOG NRQ) per embryo sample for each of the 7 GOIs assessed and resulting arithmetic mean LOG NRQ (M) and standard error of the mean (SEM) per group according to genotype (Geno), including number of individuals within each group (n). [C] sample used as calibrator.

Embryo Genotype Analysis			*Larp2/3 1a*			*Larp2/3 3*			*Ldia1 3'UTR*			*Ldia2 3' UTR*			*Ldia2 ORF*			*Lfat1*			*Lfry*		
ID	**Geno**	***n***	**LOG NRQ**	**M**	**SEM**	**LOG NRQ**	**M**	**SEM**	**LOG NRQ**	**M**	**SEM**	**LOG NRQ**	**M**	**SEM**	**LOG NRQ**	**M**	**SEM**	**LOG NRQ**	**M**	**SEM**	**LOG NRQ**	**M**	**SEM**
11289			0.00			-0.04			-0.05			-0.02			-0.01			-0.15			0.02		
11292			0.06			-0.03			0.15			0.07			0.07			0.01			0.03		
11293	*DD*	6	0.03	0.03	0.01	-0.07	-0.06	0.02	-0.04	0.01	0.03	-0.09	-0.01	0.02	-0.06	0.00	0.02	-0.03	-0.04	0.02	-0.07	-0.01	0.02
11295[C]			0.00			0.00			0.00			0.00			0.00			0.00			0.00		
11297			0.00			-0.12			0.02			-0.01			-0.05			-0.02			-0.03		
11298			0.07			-0.09			-0.03			0.00			0.07			-0.06			-0.02		
11358			0.10			-0.10			-0.09			-0.36			-0.38			-0.08			-0.09		
11359			0.11			-0.15			0.02			-0.21			-0.20			-0.04			-0.01		
11360	*Dd*	5	0.04	0.05	0.02	-0.24	-0.19	0.03	-0.02	-0.01	0.03	-0.26	-0.26	0.03	-0.37	-0.32	0.03	-0.01	-0.02	0.02	-0.01	-0.02	0.02
11361			0.02			-0.20			0.00			-0.22			-0.35			-0.02			0.03		
11363			0.00			-0.26			0.06			-0.24			-0.30			0.04			-0.01		
11282			0.01			-0.01			0.09			-2.00			-1.55			-0.06			-0.08		
11283			-0.02			-0.02			0.01			-2.79			-1.33			0.02			-0.01		
11284	*dd*	6	-0.02	-0.01	0.01	-0.11	-0.07	0.02	0.13	0.05	0.03	-1.87	-2.20	0.15	-1.49	-1.54	0.07	0.00	0.00	0.02	-0.02	-0.03	0.01
11287			-0.02			-0.11			0.11			-2.04			-1.38			0.07			0.00		
11301			0.02			-0.09			-0.08			-2.49			-1.74			-0.02			-0.05		
11303			-0.01			-0.12			0.06			-2.00			-1.75			-0.02			-0.02		

Table 34 Log-transformed normalised relative quantities (LOG NRQ) per foot sample for each of the 14 GOIs assessed and resulting arithmetic mean LOG NRQ (M) and standard error of the mean (SEM) per group according to genotype (Geno), including number of individuals within each group (n). [C] sample used as calibrator.

Foot, GOI 1-7			*Larp2/3 1a*			*Larp2/3 3a*			*Ldia1 3' UTR*			*Ldia2 3' UTR*			*Ldia2 ORF*			*Lfat1*			*Lfry*		
ID	**Geno**	***n***	**LOG NRQ**	**M**	**SEM**	**LOG NRQ**	**M**	**SEM**	**LOG NRQ**	**M**	**SEM**	**LOG NRQ**	**M**	**SEM**	**LOG NRQ**	**M**	**SEM**	**LOG NRQ**	**M**	**SEM**	**LOG NRQ**	**M**	**SEM**
11347			0.36			-0.02			0.07			0.63			0.63			0.43			0.45		
11350[C]			0.00			0.00			0.00			0.00			0.00			0.00			0.00		
11351	*DD*	5	0.28	0.18	0.06	-0.03	-0.03	0.01	0.09	0.00	0.03	0.31	0.24	0.11	0.29	0.21	0.11	0.31	0.17	0.09	0.31	0.17	0.09
11352			0.11			0.00			-0.09			0.07			0.05			-0.01			-0.01		
11357			0.14			-0.07			-0.05			0.21			0.09			0.12			0.09		
11348			0.32			-0.01			0.07			0.26			0.38			0.33			0.16		
11349			0.21			0.02			-0.04			0.22			0.21			0.12			0.07		
11353	*dd*	5	0.32	0.26	0.03	0.01	0.00	0.01	0.02	-0.01	0.02	0.42	0.31	0.05	0.50	0.36	0.06	0.32	0.25	0.04	0.23	0.19	0.04
11354			0.16			-0.01			-0.05			0.22			0.25			0.22			0.17		
11356			0.31			-0.02			-0.03			0.41			0.49			0.28			0.30		
Foot, GOI 8-14			***Lcol11 2a***			***Lmhc***			***Lmhc nm***			***Lmyo5a***			***Lmyo18a***			***Lstau***			***Lunc93a***		
ID	**Geno**	***n***	**LOG NRQ**	**M**	**SEM**	**LOG NRQ**	**M**	**SEM**	**LOG NRQ**	**M**	**SEM**	**LOG NRQ**	**M**	**SEM**	**LOG NRQ**	**M**	**SEM**	**LOG NRQ**	**M**	**SEM**	**LOG NRQ**	**M**	**SEM**
11347			0.68			0.39			0.32			0.73			0.52			0.62			1.23		
11350[C]			0.00			0.00			0.00			0.00			0.00			0.00			0.00		
11351	*DD*	5	0.69	0.38	0.14	0.35	0.15	0.09	0.33	0.10	0.10	0.57	0.31	0.14	0.32	0.18	0.13	0.46	0.25	0.12	0.74	0.55	0.21
11352			0.20			0.01			0.00			0.09			-0.21			0.06			0.35		
11357			0.33			0.00			-0.17			0.14			0.25			0.10			0.40		
11348			0.76			0.41			0.19			0.42			0.47			0.40			0.87		
11349			0.71			0.46			0.04			0.13			0.19			0.22			0.69		
11353	*dd*	5	0.79	0.65	0.06	0.34	0.30	0.08	0.14	0.05	0.06	0.58	0.40	0.08	0.29	0.34	0.06	0.43	0.32	0.06	0.86	0.72	0.09
11354			0.50			0.03			-0.17			0.33			0.24			0.11			0.36		
11356			0.52			0.27			0.07			0.53			0.51			0.42			0.81		

Table 35 Log-transformed normalised relative quantities (LOG NRQ) per ovotestis sample for five GOIs (*Larp2/3 1a, Larp2/3 3, Ldia1 3' UTR, Ldia2 3' UTR, Ldia2 ORF*) assessed and resulting arithmetic mean LOG NRQ (M) and standard error of the mean (SEM) per group according to genotype (Geno), including number of individuals within each group (n). [C] sample used as calibrator.

Ovo, GOI 1-5			*Larp2/3 1a*			*Larp2/3 3a*			*Ldia1 3' UTR*			*Ldia2 3' UTR*			*Ldia2 ORF*		
ID	**Geno**	***n***	**LOG NRQ**	**M**	**SEM**	**LOG NRQ**	**M**	**SEM**	**LOG NRQ**	**M**	**SEM**	**LOG NRQ**	**M**	**SEM**	**LOG NRQ**	**M**	**SEM**
8515			0.11			0.02			0.25			0.16			0.02		
8548[C]			0.00			0.00			0.00			0.00			0.00		
8582			0.14			-0.02			0.18			0.24			0.13		
8583			0.03			-0.03			0.22			0.01			0.08		
9014			-0.03			-0.23			-0.02			0.38			0.11		
10627			0.12			-0.31			-0.05			-0.01			-0.06		
10633	*DD*	14	0.25	0.00	0.04	-0.26	-0.09	0.04	-0.02	0.07	0.03	-0.30	0.11	0.05	-0.36	-0.02	0.04
10636			0.02			-0.30			-0.10			0.17			0.00		
10638			0.14			-0.26			-0.01			0.04			-0.06		
11347			-0.07			0.12			0.20			0.15			-0.06		
11350			-0.19			0.06			0.10			0.13			-0.04		
11351			-0.19			-0.09			0.09			0.10			-0.15		
11352			-0.09			0.01			0.14			0.38			0.22		
11357			-0.21			0.03			0.06			0.10			-0.11		
8554			0.07			-0.01			0.09			0.14			0.14		
8555			0.03			-0.07			0.22			0.10			0.00		
8559			0.08			-0.09			0.23			-0.02			-0.21		
8562	*Dd*	8	-0.07	0.04	0.06	-0.04	-0.21	0.08	-0.02	0.00	0.08	0.04	0.01	0.06	-0.10	-0.11	0.07
9013			-0.19			-0.35			-0.29			0.25			0.06		
10622			0.28			-0.32			0.06			-0.22			-0.25		
10629			0.26			-0.11			0.10			-0.25			-0.45		
10639			-0.15			-0.68			-0.38			0.07			-0.08		
8806			-0.07			-0.11			0.04			0.04			-0.09		
8808			0.19			-0.25			-0.01			-0.12			-0.23		
8996			0.13			-0.18			0.17			-0.05			0.00		
9005			-0.07			-0.13			0.24			-0.01			-0.16		
9007			-0.02			-0.23			0.09			-0.13			-0.21		
10626			-0.01			-0.27			-0.06			-0.30			-0.45		
10630	*dd*	14	0.06	-0.01	0.05	-0.35	-0.15	0.04	-0.07	0.08	0.04	-0.36	-0.10	0.05	-0.42	-0.18	0.04
10640			0.32			-0.29			-0.04			-0.39			-0.38		
10642			0.31			0.01			-0.19			-0.39			-0.34		
11348			-0.14			-0.05			0.22			0.04			-0.09		
11349			-0.19			0.01			0.25			0.07			0.00		
11353			-0.10			0.04			0.26			0.11			0.01		
11354			-0.35			-0.33			-0.07			0.26			0.04		
11356			-0.24			0.08			0.31			-0.17			-0.23		

Table 36 Log-transformed normalised relative quantities (LOG NRQ) per ovotestis sample for five GOIs (*Lfat1, Lfry, Lcol11a 2/1, Lmhc, Lmhc nm*) assessed and resulting arithmetic mean LOG NRQ (M) and standard error of the mean (SEM) per group according to genotype (Geno), including number of individuals within each group (n). [C] sample used as calibrator.

Ovo, GOI 6-10			*Lfat1*			*Lfry*			*Lcol11a 2/1*			*Lmhc*			*Lmhc nm*		
ID	**Geno**	***n***	**LOG NRQ**	**M**	**SEM**	**LOG NRQ**	**M**	**SEM**	**LOG NRQ**	**M**	**SEM**	**LOG NRQ**	**M**	**SEM**	**LOG NRQ**	**M**	**SEM**
8515			0.13			0.13			0.05			0.03			-0.04		
8548[C]			0.00			0.00			0.00			0.00			0.00		
8582			0.03			0.15			0.31			0.25			0.01		
8583			0.00			0.17			0.49			-0.13			-0.08		
9014			-0.17			-0.11			0.05			-0.22			-0.08		
10627			-0.23			-0.08			0.10			0.10			-0.28		
10633	*DD*	14	-0.46	-0.17	0.05	-0.17	-0.03	0.04	0.52	0.23	0.06	0.23	0.01	0.06	-0.28	-0.12	0.03
10636			-0.42			-0.29			-0.01			-0.07			-0.18		
10638			-0.32			-0.16			-0.03			0.21			-0.26		
11347			-0.10			0.02			0.35			0.19			-0.02		
11350			-0.14			0.04			0.26			-0.40			-0.08		
11351			-0.30			-0.10			0.52			0.07			-0.14		
11352			-0.32			0.02			0.63			0.17			-0.06		
11357			-0.10			-0.10			-0.03			-0.34			-0.15		
8554			-0.05			0.08			0.41			0.05			-0.05		
8555			-0.01			0.14			-0.22			-0.50			-0.23		
8559			-0.16			-0.01			0.01			-0.27			-0.21		
8562	*Dd*	8	-0.21	-0.25	0.06	-0.16	-0.08	0.05	-0.14	0.04	0.09	-0.56	-0.16	0.10	-0.25	-0.16	0.03
9013			-0.40			-0.17			0.12			-0.34			-0.11		
10622			-0.30			-0.05			-0.03			0.15			-0.08		
10629			-0.39			-0.13			0.39			0.14			-0.04		
10639			-0.48			-0.30			-0.18			0.08			-0.27		
8806			-0.26			-0.02			0.27			-0.17			-0.16		
8808			-0.28			-0.04			0.27			-0.06			-0.21		
8996			-0.05			0.04			0.22			-0.32			-0.04		
9005			-0.12			-0.02			0.07			-0.53			-0.18		
9007			-0.15			-0.05			0.38			-0.37			-0.07		
10626			-0.37			-0.17			0.21			0.09			-0.02		
10630	*dd*	14	-0.35	-0.24	0.04	-0.26	-0.09	0.04	-0.05	0.24	0.08	0.15	-0.16	0.08	-0.22	-0.09	0.03
10640			-0.25			-0.30			0.17			0.45			-0.23		
10642			-0.49			-0.47			-0.28			-0.11			0.04		
11348			-0.26			0.00			0.52			-0.15			-0.11		
11349			-0.05			0.07			1.03			0.00			0.08		
11353			-0.12			0.02			0.48			-0.53			0.08		
11354			-0.48			-0.08			0.08			-0.55			-0.12		
11356			-0.18			0.00			-0.03			-0.13			-0.06		

Table 37 Log-transformed normalised relative quantities (LOG NRQ) per ovotestis sample for four GOIs (*Lmyo5a, Lmyo18a, Lstau, Lunc93a*) assessed and resulting arithmetic mean LOG NRQ (M) and standard error of the mean (SEM) per group according to genotype (Geno), including number of individuals within each group (n). [c] sample used as calibrator.

Ovo, GOI 11-14			*Lmyo5a*			*Lmyo18a*			*Lstau*			*Lunc93a*		
ID	**Geno**	***n***	**LOG NRQ**	**M**	**SEM**	**LOG NRQ**	**M**	**SEM**	**LOG NRQ**	**M**	**SEM**	**LOG NRQ**	**M**	**SEM**
8515	*DD*	14	-0.08	-0.07	0.03	-0.13	-0.35	0.06	-0.20	0.06	0.07	-0.19	-0.20	0.07
8548			0.00			0.00			0.00			0.00		
8582			0.12			-0.15			0.22			0.18		
8583			0.02			0.01			0.33			0.30		
9014			0.10			-0.24			0.63			-0.59		
10627			-0.16			-0.53			0.10			-0.39		
10633			-0.13			-0.60			0.18			-0.20		
10636			-0.26			-0.47			0.30			-0.19		
10638			-0.14			-0.67			0.12			-0.44		
11347			-0.08			-0.34			-0.10			-0.09		
11350			-0.03			-0.35			-0.20			-0.18		
11351			-0.13			-0.44			-0.20			-0.46		
11352			-0.10			-0.43			-0.02			-0.25		
11357			-0.15			-0.52			-0.33			-0.30		
8554	*Dd*	8	0.14	-0.05	0.04	-0.07	-0.23	0.07	0.14	0.31	0.11	-0.03	-0.31	0.10
8555			-0.09			0.10			0.22			0.00		
8559			-0.07			-0.17			0.21			-0.23		
8562			-0.02			-0.30			-0.04			-0.31		
9013			0.04			-0.13			0.93			-0.43		
10622			0.02			-0.45			0.36			-0.31		
10629			-0.22			-0.39			0.13			-0.24		
10639			-0.18			-0.47			0.50			-0.94		
8806	*dd*	14	0.03	-0.06	0.05	-0.17	-0.35	0.04	0.51	0.18	0.07	-0.25	-0.33	0.07
8808			-0.08			-0.23			0.33			-0.21		
8996			0.06			-0.18			0.20			-0.25		
9005			-0.13			-0.36			-0.06			-0.51		
9007			0.06			-0.15			0.61			-0.24		
10626			-0.28			-0.38			0.11			-0.26		
10630			-0.33			-0.74			0.04			-0.36		
10640			-0.23			-0.38			-0.09			-0.24		
10642			-0.32			-0.55			0.12			-0.62		
11348			0.03			-0.33			0.10			-0.27		
11349			0.15			-0.21			0.09			0.13		
11353			0.19			-0.35			0.12			-0.34		
11354			0.19			-0.42			0.63			-0.95		
11356			-0.14			-0.45			-0.25			-0.20		

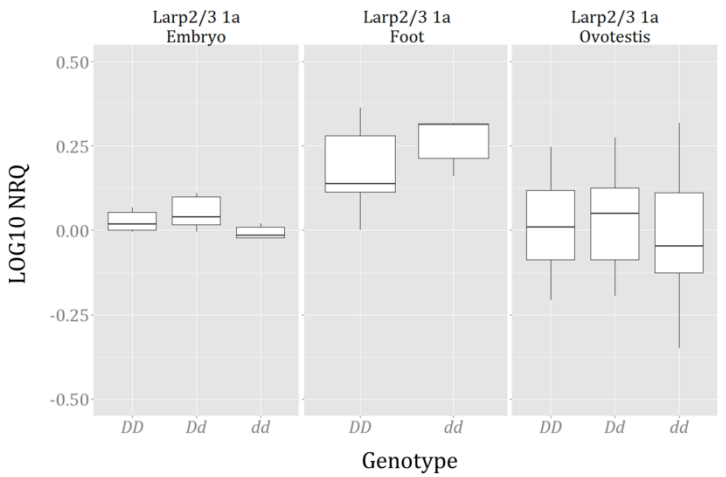


Figure 16 Composite boxplot showing Log scale NRQ values (LOG10 NRQ) for *Larp2/3 1a* in embryo, foot and ovotestis tissue, compared between genotypes *DD*, *Dd* & *dd*, calculated relative to a conspecific *DD* individual.

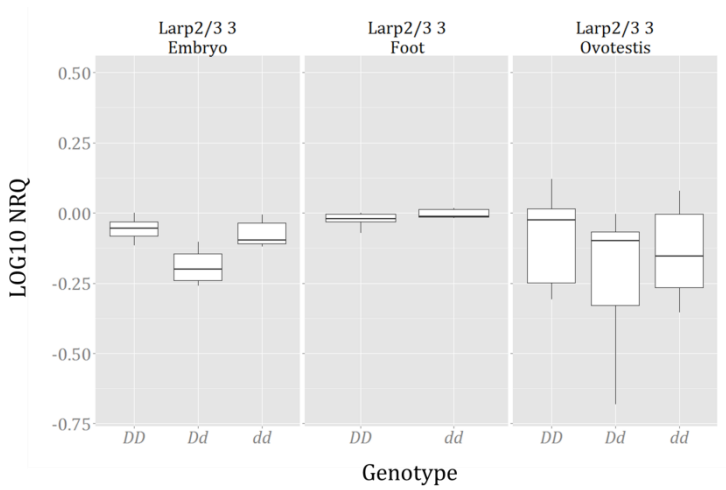


Figure 17 Composite boxplot showing Log scale NRQ values (LOG10 NRQ) for *Larp2/3 3* in embryo, foot and ovotestis tissue, compared between genotypes *DD*, *Dd* & *dd*, calculated relative to a conspecific *DD* individual.

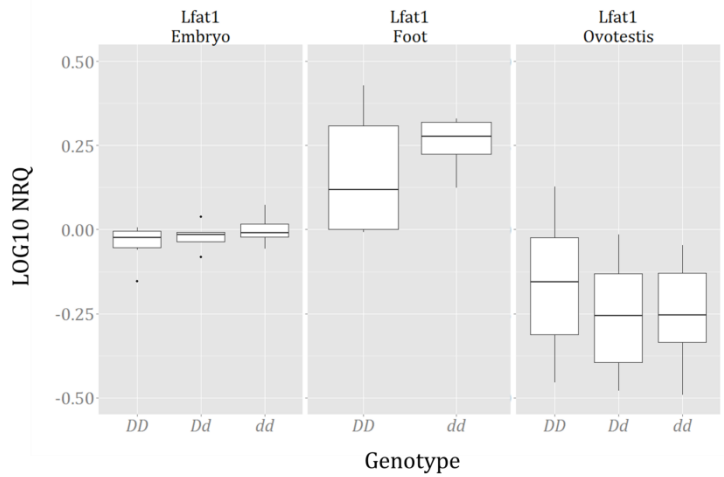


Figure 18 Composite boxplot showing Log scale NRQ values (LOG10 NRQ) for *Lfat1* in embryo, foot and ovotestis tissue, compared between genotypes *DD*, *Dd* & *dd*, calculated relative to a conspecific *DD* individual.

Figure 19 Composite boxplot showing Log scale NRQ values (LOG10 NRQ) for *Lfry* in embryo, foot and ovotestis tissue, compared between genotypes *DD*, *Dd* & *dd*, calculated relative to a conspecific *DD* individual.

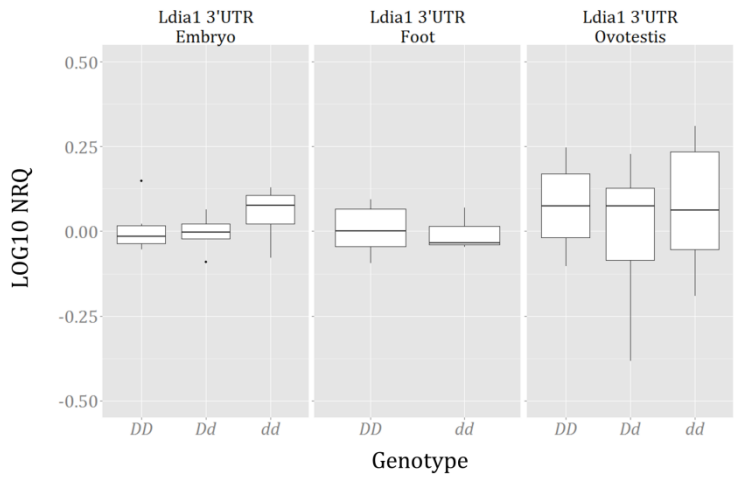


Figure 20 Composite boxplot showing Log scale NRQ values (LOG10 NRQ) for *Ldia1 3' UTR* in embryo, foot and ovotestis tissue, compared between genotypes *DD*, *Dd* & *dd*, calculated relative to a conspecific *DD* individual.

Figure 21 Composite boxplot showing Log scale NRQ values (LOG10 NRQ) for *Ldia2 3' UTR* in embryo, foot and ovotestis tissue, compared between genotypes *DD*, *Dd* & *dd*, calculated relative to a conspecific *DD* individual.

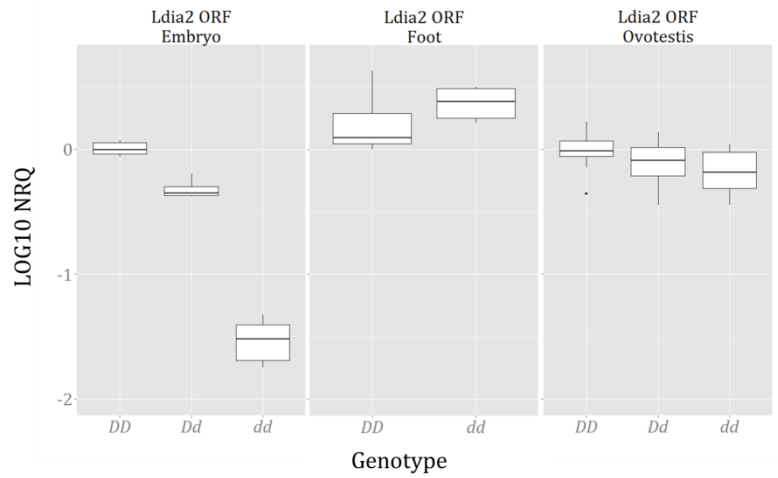


Figure 22 Composite boxplot showing Log scale NRQ values (LOG10 NRQ) for *Ldia2 ORF* in embryo, foot and ovotestis tissue, compared between genotypes *DD*, *Dd* & *dd*, calculated relative to a conspecific *DD* individual.

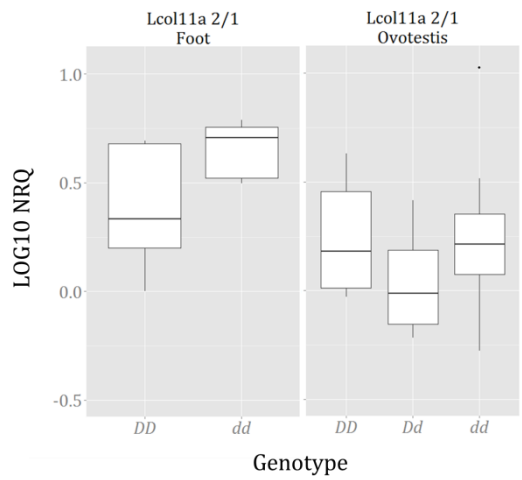


Figure 23 Composite boxplot showing Log scale NRQ values (LOG10 NRQ) for *Lcol11a 2/1* in foot and ovotestis tissue, compared between genotypes *DD* & *dd*, calculated relative to a conspecific *DD* individual.

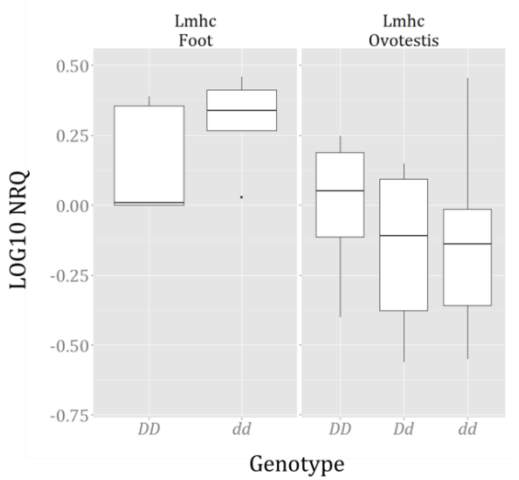


Figure 24 Composite boxplot showing Log scale NRQ values (LOG10 NRQ) for *Lmhc* in foot and ovotestis tissue, compared between genotypes *DD* & *dd*, calculated relative to a conspecific *DD* individual.

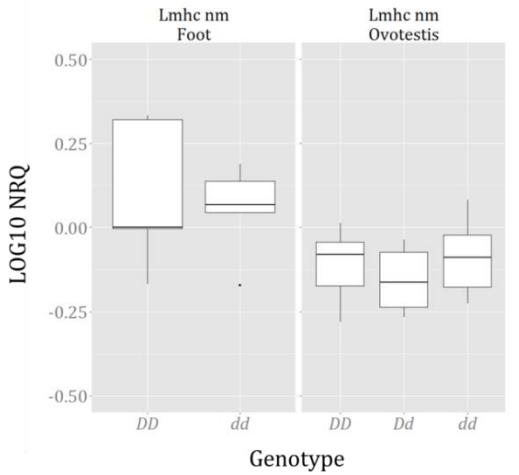


Figure 25 Composite boxplot showing Log scale NRQ values (LOG10 NRQ) for *Lmhc nm* in foot and ovotestis tissue, compared between genotypes *DD* & *dd*, calculated relative to a conspecific *DD* individual.

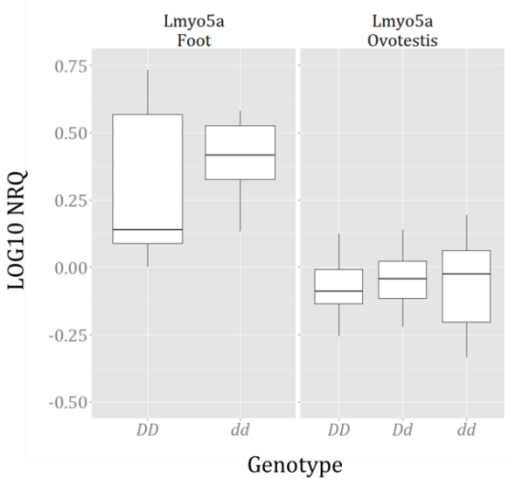


Figure 26 Composite boxplot showing Log scale NRQ values (LOG10 NRQ) for *Lmyo5a* in foot and ovotestis tissue, compared between genotypes *DD* & *dd*, calculated relative to a conspecific *DD* individual.

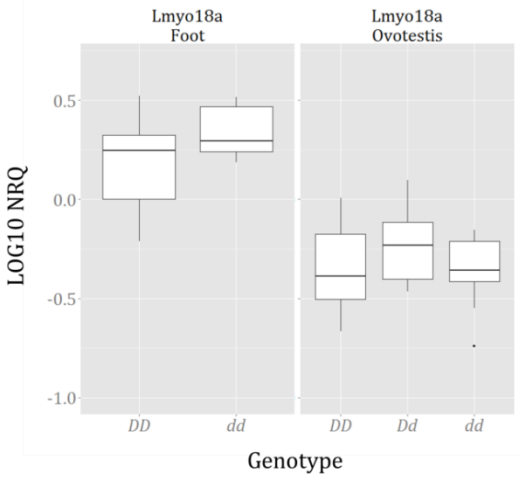


Figure 27 Composite boxplot showing Log scale NRQ values (LOG10 NRQ) for *Lmyo18a* in foot and ovotestis tissue, compared between genotypes *DD* & *dd*, calculated relative to a conspecific *DD* individual.

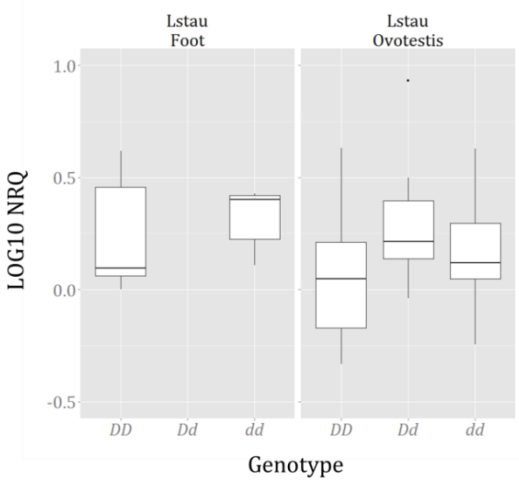


Figure 28 Composite boxplot showing Log scale NRQ values (LOG10 NRQ) for *Lstau* in foot and ovotestis tissue, compared between genotypes *DD* & *dd*, calculated relative to a conspecific *DD* individual.

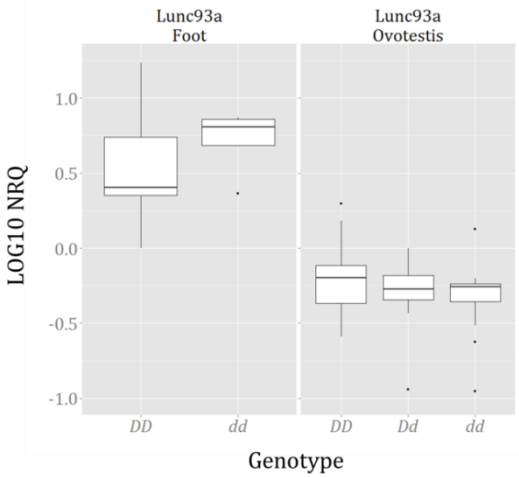


Figure 29 Composite boxplot showing Log scale NRQ values (LOG10 NRQ) for *Lunc93a* in foot and ovotestis tissue, compared between genotypes *DD* & *dd*, calculated relative to a conspecific *DD* individual.

Table 38 Wilcoxon rank test results for pairwise comparisons between genotypes *DD*, *Dd* and *dd* within embryo, foot and ovotestis tissue for seven GOIs. The total number of individuals within each genotype analysis is quoted (n) in addition to the number of individuals within each genotypic group (n, *DD*; n, *Dd*; n, *dd*). The Wilcoxon rank value (W) is presented with the associated probability value (p). Statistical significance (sig) is highlighted via * <0.05, ** <0.01.

Genotype Analysis GOI 1-7, Wilcox.test																	
GOI	**Tissue**	***n***	***DD-dd***					***DD-Dd***					***Dd-dd***				
			n, DD	***n, dd***	**W**	**p**	**sig**	***n, DD***	***n, Dd***	**W**	**p**	**sig**	***n, Dd***	***n, dd***	**W**	**p**	**sig**
Larp2/3 1a	**Embryo**	17	6	6	30	0.065		6	5	10	0.429		5	6	27	0.030	*
	Foot	10	5	5	7	0.310		n/a									
	Ovotestis	36	14	14	106	0.735		14	8	50	0.714		8	14	65	0.570	
Larp2/3 3	**Embryo**	17	6	6	23	0.485		6	5	29	0.009	**	5	6	3	0.030	*
	Foot	10	5	5	6	0.222		n/a									
	Ovotestis	36	14	14	117	0.401		14	8	81	0.095		8	14	50	0.714	
Ldia1 3'UTR	**Embryo**	17	6	6	12	0.394		6	5	15	1		5	6	7	0.178	
	Foot	10	5	5	13	1		n/a									
	Ovotestis	36	14	14	95	0.910		14	8	61	0.764		8	14	47	0.570	
Ldia2 3' UTR	**Embryo**	17	6	6	36	0.002	**	6	5	30	0.004	**	5	6	30	0.004	**
	Foot	10	5	5	8	0.421		n/a									
	Ovotestis	36	14	14	155	0.008	**	14	8	73	0.267		8	14	74	0.238	
Ldia2 ORF	**Embryo**	17	6	6	36	0.002	**	6	5	30	0.004	**	5	6	30	0.004	**
	Foot	10	5	5	7	0.310		n/a									
	Ovotestis	36	14	14	152	0.012	*	14	8	74	0.238		8	14	68	0.441	
Lfat1	**Embryo**	17	6	6	11	0.310		6	5	14	0.931		5	6	12	0.662	
	Foot	10	5	5	8	0.421		n/a									
	Ovotestis	36	14	14	123	0.265		14	8	70	0.365		8	14	53	0.868	
Lfry	**Embryo**	17	6	6	25	0.310		6	5	16	0.931		5	6	19	0.537	
	Foot	10	5	5	11	0.841		n/a									
	Ovotestis	36	14	14	112	0.541		14	8	69	0.402		8	14	56	1	

Table 39 Wilcoxon rank test results for pairwise comparisons between genotypes *DD*, *Dd* and *dd* within foot and ovotestis tissue for seven GOIs. The total number of individuals within each genotype analysis is quoted (n) in addition to the number of individuals within each genotypic group (n, *DD*; n, *Dd*; n, *dd*). The Wilcoxon rank value (W) is presented with the associated probability value (p). Statistical significance (sig) is highlighted via * <0.05, ** <0.01.

Genotype Analysis GOI 8-14, Wilcox.test																	
GOI	**Tissue**	***n***	***DD-dd***					***DD-Dd***					***Dd-dd***				
			n, DD	***n, dd***	**W**	**p**	**sig**	***n, DD***	***n, Dd***	**W**	**p**	**sig**	***n, Dd***	***n, dd***	**W**	**p**	**sig**
Lcoll11a 2/1	**Foot**	10	5	5	4	0.095		n/a									
	Ovotestis	36	14	14	100	0.946		14	8	81	0.095		8	14	35	0.165	
Lmhc	**Foot**	10	5	5	6	0.222		n/a									
	Ovotestis	36	14	14	138	0.069		14	8	75	0.212		8	14	58	0.920	
Lmhc nm	**Foot**	10	5	5	13	1		n/a									
	Ovotestis	36	14	14	85	0.571		14	8	68	0.441		8	14	33	0.127	
Lmyo18a	**Foot**	10	5	5	10	0.691		n/a									
	Ovotestis	36	14	14	95	0.910		14	8	39	0.267		8	14	73	0.267	
Lmyo5a	**Foot**	10	5	5	10	0.691		n/a									
	Ovotestis	36	14	14	90	0.735		14	8	47	0.570		8	14	55	0.973	
Lstau	**Foot**	10	5	5	10	0.691		n/a									
	Ovotestis	36	14	14	76	0.329		14	8	29	0.070		8	14	77	0.165	
Lunc93a	**Foot**	10	5	5	8	0.421		n/a									
	Ovotestis	36	14	14	130	0.150		14	8	66	0.525		8	14	61	0.764	

Diaphanous formin, *Ldia2 3' UTR*

Embryo

The LOG NRQs recorded in the *Ldia2 3' UTR* embryo experiment were found to be highly statistically significant between all genotype groups (all p values <0.005, Table 38). The heterozygote *Dd* samples exhibited almost exactly half the expression level seen in the *DD* samples (Table 32). The *dd* samples showed a dramatically reduced expression level at 0.6% of that of the *DD* calibrator. Alternatively considered the *DD* sample exhibited 167 times higher expression of *Ldia2* 3' UTR compared to the *dd* samples (Table 32).

Ovotestis

The same pattern of gene expression was observed in the ovotestis tissue however to a much lesser extent and with increased variation (Figure 21). The only significant difference in relative gene expression was identified between *DD* and *dd* samples, yet it was still found to be highly significant with a p value of 0.007 (Table 38). The *dd* samples were observed to express 79.7% of the level of *Ldia2* 3' UTR expression observed in the *DD* sample, or alternatively the *DD* sample expressed 1.25 times higher expression of *Ldia2* 3' UTR compared to the *dd* samples (Table 32).

Foot

The only comparisons possible for the foot tissue were between the homozygote genotypes. No significant difference was found between the genotypes. The boxplot shows the Log transformed relative expression of the *dd* group to be slightly higher than that of the *DD* group (Figure 21); however, this small scale difference is negated by the variation seen in the *DD* samples.

Diaphanous formin, *Ldia2 ORF*

Embryo

The open reading frame (ORF) of the diaphanous formin showed a very similar expression pattern as the 3'UTR in the embryo experiment. The Log transformed relative expression levels of *Ldia2* ORF in all genotypic groups were found to be highly significantly different (p values <0.005, Table 38). The largest difference again, was seen between the two homozygote sample groups, with the *dd* sample group exhibiting 2.9% of the level of expression seen in the *DD* sample. The level of expression in the heterozygote group exhibited just under half of the expression level of the homozygote *DD* (Table 32, Figure 22).

Ovotestis

Similarly to the *Ldia2 3' UTR* experiment, the *Ldia2* ORF ovotestis expression pattern revealed a less pronounced difference than that seen in the embryo samples (Figure 22). Again, a statistically significant difference in Log-transformed expression level was only observed between the homozygote groups (p value = 0.01, Table 38). The *dd* samples exhibited 65.6% of the expression level recorded in the *DD* sample (Table 32).

Foot

No significant difference was found between the Log-transformed relative levels of expression of *Ldia2* ORF of homozygote genotypes in the foot tissue. Again the trend seen from the boxplot reveals a small increase in the average expression ratio of the *dd* sample group compared to the *DD* sample group, yet this is not statistically significant (Figure 22, Table 38).

Actin related protein 2/3 subunit 1a, *Larp2/3 1a*

Embryo

In the embryo tissue, *Larp2/3 1a* showed an increase in expression of both the *DD* and the *Dd* samples compared to the *dd* sample group (Table 32, Figure 16). This difference was only found to be statistically significant between the heterozygote *Dd* and the *dd* sample groups (p value=0.030, Table 38) and not between the homozygote groups (p value = 0.065, Table 38). The boxplot appears to show the Log-transformed expression ratio of the heterozygote, *Dd* also to be increased compared to the *DD* group; however, this is not statistically significant (Figure 16, Table 38).

No significant expression differentiation was found between any of the genotypes within the other tissues.

Actin related protein 2/3 subunit 3, *Larp2/3 3*

Embryo

Larp2/3 3 in the embryo tissue showed a decrease in expression in the heterozygote, *Dd* samples compared to both the homozygote sample groups (Figure 17). The heterozygote samples expressed approximately 70% of the expression level observed in the homozygote groups (Table 32). This difference was found to be statistically significant between both groups, however the significance of the sinistral, *dd*, homozygote sample group (p value=0.030) was not as strong as that of the dextral, *DD*, homozygote group (p value =0.009, Table 38).

Ovotestis

Although not statistically significant (p value = 0.095), the boxplot of the ovotestis tissue analysis also reveals a reduction in the average Log-transformed expression ratio of *Larp2/3 3* of the heterozygote group, compared to both homozygote groups.

Foot

The foot analysis demonstrated no difference in relative expression between genotypic groups; however, the foot analysis compared only homozygote samples.

Tissue Analysis

The NRQ values compared in the tissue analysis are all relative to the same ovotestis reference sample 'OvoRef', which can loosely be considered to represent a heterozygote (see Methods). Again the relative expression data is mostly presented and analysed using NRQ values log-transformed to the base 10 (LOG NRQ). This is important to note when interpreting the observed log-fold changes in gene expression. The relative quantities of the OvoRef sample were not included in the analyses and so are not presented in the NRQ data tables.

The relative quantities (RQ) of the two endogenous control genes, including the geometric mean RQ used to calculate the NRQ values in the embryo, foot and ovotestis tissue analyses are presented in Table 40 and Table 41. A summary of the genotype and tissue specific group means of NRQs for each of the 14 GOIs assessed included in the tissue comparison analysis is presented in Table 42. No statistical analyses were performed on the non-transformed NRQ values, therefore only the geometric group mean is presented.

The LOG NRQs for the seven GOIs quantified in all three tissues are presented across the two tables; Table 43 and Table 44. The LOG NRQs for the remaining seven GOIs quantified in only the foot and ovotestis tissues are presented across the two tables; Table 45 and Table 46. These tables include the individual sample count data and the genotype-specific group means according to tissue, that were used within the statistical analyses (Table 47, Table 48) and the boxplot graphs (Figure 30 - Figure 43). Summary statistics are presented in the SI (S8).

Table 40 Relative quantity (RQ) values per embryo and foot sample, plus the ovotestis reference sample (OvoRef), for both of the endogenous control genes assessed in the tissue analysis (*Lhis2a*, *Lube2*) and resulting geometric mean (GeoMean), including sample ID (ID) and genotype (Geno). [C] sample used as calibrator.

Sample Description			RQ values		
ID	**Geno**	**Tissue**	***Lhis2a***	***Lube2***	**GeoMean**
11289	*DD*	Embryo	1.619	0.175	0.532
11292	*DD*	Embryo	2.001	0.247	0.703
11293	*DD*	Embryo	2.359	0.203	0.691
11295	*DD*	Embryo	1.454	0.131	0.436
11297	*DD*	Embryo	2.496	0.238	0.771
11298	*DD*	Embryo	3.028	0.270	0.904
11358	*Dd*	Embryo	2.755	0.325	0.946
11359	*Dd*	Embryo	3.837	0.463	1.333
11360	*Dd*	Embryo	4.606	0.551	1.593
11361	*Dd*	Embryo	3.935	0.504	1.408
11363	*Dd*	Embryo	5.896	0.771	2.132
11282	*dd*	Embryo	2.826	0.340	0.980
11283	*dd*	Embryo	1.211	0.118	0.378
11284	*dd*	Embryo	2.551	0.291	0.861
11287	*dd*	Embryo	1.520	0.150	0.477
11301	*dd*	Embryo	1.350	0.119	0.400
11303	*dd*	Embryo	1.828	0.209	0.618
11347	*DD*	Foot	0.054	0.025	0.036
11350	*DD*	Foot	0.860	0.378	0.570
11351	*DD*	Foot	0.201	0.070	0.119
11352	*DD*	Foot	0.533	0.199	0.326
11357	*DD*	Foot	0.824	0.331	0.522
11348	*dd*	Foot	0.219	0.133	0.171
11349	*dd*	Foot	0.323	0.155	0.224
11353	*dd*	Foot	0.211	0.129	0.165
11354	*dd*	Foot	0.435	0.283	0.351
11356	*dd*	Foot	0.239	0.106	0.159
OvoRef[C]	*D/d*	Ovotestis	1	1	1

Table 41 Relative quantity (RQ) values per ovotestis sample, plus the ovotestis reference sample (OvoRef), for both of the endogenous control genes assessed in the tissue analysis (*Lhis2a*, *Lube2*) and resulting geometric mean (GeoMean), including sample ID (ID) and genotype (Geno). [C] sample used as calibrator.

Sample Description			RQ values		
ID	**Geno**	**Tissue**	***Lhis2a***	***Lube2***	**GeoMean**
8515	*DD*	Ovotestis	0.175	0.251	0.210
8548	*DD*	Ovotestis	0.175	0.083	0.120
8582	*DD*	Ovotestis	0.102	0.145	0.122
8583	*DD*	Ovotestis	0.190	0.227	0.208
9014	*DD*	Ovotestis	0.106	0.102	0.104
10627	*DD*	Ovotestis	0.179	0.196	0.187
10633	*DD*	Ovotestis	0.207	0.279	0.240
10636	*DD*	Ovotestis	0.263	0.242	0.252
10638	*DD*	Ovotestis	0.107	0.135	0.120
11347	*DD*	Ovotestis	0.301	0.412	0.352
11350	*DD*	Ovotestis	0.518	0.725	0.613
11351	*DD*	Ovotestis	1.021	0.800	0.904
11352	*DD*	Ovotestis	0.176	0.203	0.189
11357	*DD*	Ovotestis	0.417	0.562	0.484
8554	*Dd*	Ovotestis	0.113	0.100	0.106
8555	*Dd*	Ovotestis	0.168	0.219	0.192
8559	*Dd*	Ovotestis	0.117	0.235	0.166
8562	*Dd*	Ovotestis	0.437	0.191	0.289
9013	*Dd*	Ovotestis	0.388	0.259	0.317
10622	*Dd*	Ovotestis	0.145	0.245	0.189
10629	*Dd*	Ovotestis	0.150	0.181	0.165
10639	*Dd*	Ovotestis	0.152	0.174	0.163
8806	*dd*	Ovotestis	0.298	0.228	0.260
8808	*dd*	Ovotestis	0.377	0.265	0.316
8996	*dd*	Ovotestis	0.095	0.108	0.101
9005	*dd*	Ovotestis	0.259	0.261	0.260
9007	*dd*	Ovotestis	0.201	0.214	0.207
10626	*dd*	Ovotestis	0.619	0.698	0.657
10630	*dd*	Ovotestis	0.165	0.231	0.195
10640	*dd*	Ovotestis	0.057	0.080	0.067
10642	*dd*	Ovotestis	0.318	0.140	0.211
11348	*dd*	Ovotestis	0.283	0.349	0.314
11349	*dd*	Ovotestis	0.095	0.084	0.090
11353	*dd*	Ovotestis	0.374	0.582	0.467
11354	*dd*	Ovotestis	0.851	0.730	0.788
11356	*dd*	Ovotestis	0.460	0.755	0.590
OvoRef	*D/d*	Ovotestis	1	1	1

Table 42 Normalised relative quantities (NRQ) of each GOI, presented as a geometric mean per genotypic group (Geno) and tissue within the tissue analysis, including number of samples within each group (n). Each value is relative to the ovotestis reference sample, 'OvoRef'.

Tissue Analysis			NRQ													
Tissue	**Geno**	**n**	***Larp2/31a***	***Larp2/33***	***Ldia1 3' UTR***	***Ldia2 3' UTR***	***Ldia2 ORF***	***Lfat1***	***Lfry***	***Lcol11 2a***	***Lmhc***	***Lmhc nm***	***Lmyo5a***	***Lmyo18a***	***Lstau***	***Lunc93a***
Embryo	*DD*	6	0.069	0.224	0.103	2.835	1.973	0.140	0.336	x	x	x	x	x	x	x
Embryo	*Dd*	5	0.077	0.175	0.105	1.676	0.988	0.154	0.348	x	x	x	x	x	x	x
Embryo	*dd*	6	0.065	0.219	0.118	0.019	0.058	0.158	0.330	x	x	x	x	x	x	x
Foot	*DD*	5	0.652	0.656	0.910	1.012	0.708	1.062	0.880	3.681	5.056	0.956	2.147	1.045	0.930	0.130
Foot	*dd*	5	0.784	0.682	0.875	1.153	0.997	1.277	0.905	6.809	7.072	0.855	2.611	1.508	1.077	0.190
Ovotestis	*DD*	14	1.935	1.068	1.049	1.418	1.553	1.339	1.239	1.011	0.926	0.933	1.612	1.215	0.931	1.059
Ovotestis	*Dd*	8	2.256	0.869	0.950	1.218	1.348	1.196	1.210	0.708	0.684	0.915	1.821	1.687	1.765	0.878
Ovotestis	*dd*	14	1.820	0.914	1.040	0.855	1.039	1.105	1.060	1.005	0.618	0.975	1.624	1.174	1.185	0.770

Table 43 Log-transformed normalised relative quantities (LOG NRQ) for each sample included within the tissue analysis of the four GOIs: *Larp2/3 1a*, *Larp2/3 3*, *Lfat1* & *Lfry*, with resulting arithmetic mean LOG NRQ (M) and standard error of the mean (SEM) per group according to genotype (Geno) within the specific tissue. Also presented is the number of individuals within each group (n). The calibrator sample 'OvoRef' was not included in analyses.

Sample and group description				*Larp2/3 1a*			*Larp2/3 3*			*Lfat1*			*Lfry*		
ID	Geno	Tissue	n	LOG NRQ	M	SEM	LOG NRQ	M	SEM	LOG NRQ	M	SEM	LOG NRQ	M	SEM
11289	*DD*	Embryo	6	-1.19	-1.16	0.02	-0.62	-0.65	0.02	-0.96	-0.85	0.02	-0.43	-0.47	0.02
11292				-1.11			-0.61			-0.79			-0.42		
11293				-1.16			-0.68			-0.85			-0.55		
11295				-1.21			-0.62			-0.84			-0.49		
11297				-1.19			-0.71			-0.83			-0.50		
11298				-1.11			-0.66			-0.86			-0.46		
11347		Foot	5	0.04	-0.19	0.08	-0.13	-0.18	0.02	0.33	0.03	0.10	0.27	-0.06	0.10
11350				-0.38			-0.18			-0.16			-0.24		
11351				-0.07			-0.18			0.18			0.10		
11352				-0.26			-0.17			-0.16			-0.24		
11357				-0.26			-0.26			-0.06			-0.17		
10627		Ovotestis	14	0.47	0.29	0.06	-0.13	0.03	0.02	0.13	0.13	0.04	0.11	0.09	0.03
10633				0.66			-0.01			-0.03			0.09		
10636				0.42			-0.07			-0.01			-0.05		
10638				0.56			-0.01			0.11			0.11		
8515				0.40			0.13			0.42			0.26		
8548				0.25			0.09			0.27			0.10		
8582				0.44			0.11			0.34			0.29		
8583				0.28			0.04			0.26			0.25		
9014				0.28			-0.09			0.15			0.05		
11347				0.11			0.13			0.09			0.03		
11350				0.01			0.10			0.08			0.09		
11351				-0.02			-0.09			-0.12			-0.09		
11352				0.15			0.08			-0.06			0.10		
11357				0.02			0.09			0.14			-0.03		
11358	*Dd*	Embryo	5	-1.06	-1.12	0.02	-0.67	-0.76	0.03	-0.87	-0.81	0.02	-0.53	-0.46	0.02
11359				-1.05			-0.70			-0.82			-0.44		
11360				-1.13			-0.81			-0.80			-0.45		
11361				-1.16			-0.78			-0.82			-0.42		
11363				-1.17			-0.83			-0.76			-0.45		
10622		Ovotestis	8	0.60	0.35	0.07	-0.16	-0.06	0.08	0.04	0.08	0.06	0.12	0.08	0.05
10629				0.64			0.10			0.00			0.09		
10639				0.24			-0.46			-0.08			-0.07		
8554				0.41			0.16			0.30			0.26		
8555				0.30			0.04			0.28			0.26		
8559				0.41			0.08			0.19			0.17		
8562				0.20			0.06			0.07			-0.05		
9013				0.02			-0.30			-0.17			-0.12		

11282				-1.16			-0.58			-0.86			-0.53		
11283				-1.18			-0.59			-0.76			-0.44		
11284		Embryo	6	-1.20	-1.18	0.01	-0.69	-0.66	0.02	-0.80	-0.80	0.02	-0.47	-0.48	0.01
11287				-1.23			-0.72			-0.75			-0.48		
11301				-1.16			-0.68			-0.83			-0.50		
11303				-1.18			-0.70			-0.81			-0.47		
11348				-0.04			-0.16			0.20			-0.06		
11349				-0.18			-0.17			-0.05			-0.19		
11353		Foot	5	-0.05	-0.11	0.04	-0.14	-0.17	0.01	0.18	0.11	0.05	0.01	-0.04	0.05
11354				-0.22			-0.19			0.07			-0.07		
11356				-0.04			-0.17			0.14			0.09		
11348				0.07			-0.01			-0.04			0.05		
11349	*dd*			0.08			0.11			0.22			0.18		
11353				0.05			0.02			0.04			0.02		
11354				-0.20			-0.35			-0.32			-0.09		
11356				-0.07			0.08			0.00			0.02		
10626				0.32			-0.10			-0.02			0.01		
10630		Ovotestis	14	0.45	0.26	0.07	-0.13	-0.04	0.03	0.05	0.04	0.04	-0.02	0.03	0.03
10640				0.76			-0.01			0.20			-0.02		
10642				0.64			0.18			-0.15			-0.30		
8806				0.17			-0.04			0.00			0.07		
8808				0.48			-0.13			0.02			0.09		
8996				0.50			0.02			0.33			0.25		
9005				0.17			-0.06			0.13			0.06		
9007				0.24			-0.13			0.13			0.05		

Table 44 Log-transformed normalised relative quantities (LOG NRQ) for each sample included within the tissue analysis of the three GOIs: *Ldia1 3' UTR*, *Ldia2 3' UTR* & *Ldia2 ORF*, with resulting arithmetic mean LOG NRQ (M) and standard error of the mean (SEM) per group according to genotype (Geno)within the specific tissue. Also presented is the number of individuals within each group (n). The calibrator sample 'OvoRef' was not included in analyses.

Sample and group description				*Ldia1 3'UTR*			*Ldia2 3' UTR*			*Ldia2 ORF*		
ID	Geno	Tissue	n	LOGNRQ	M	SEM	LOGNRQ	M	SEM	LOGNRQ	M	SEM
11289	*DD*	Embryo	6	-1.04	-0.99	0.03	0.45	0.45	0.02	0.29	0.30	0.03
11292				-0.83			0.55			0.38		
11293				-1.04			0.36			0.22		
11295				-1.02			0.44			0.27		
11297				-0.97			0.45			0.24		
11298				-1.01			0.48			0.37		
11347		Foot	5	0.06	-0.04	0.04	0.44	0.01	0.12	0.31	-0.15	0.13
11350				-0.06			-0.26			-0.38		
11351				0.06			0.08			-0.06		
11352				-0.14			-0.17			-0.32		
11357				-0.13			-0.06			-0.30		
10627		Ovotestis	14	-0.04	0.02	0.02	0.09	0.15	0.04	0.22	0.19	0.04
10633				0.06			-0.13			-0.02		
10636				-0.04			0.32			0.32		
10638				0.07			0.22			0.29		
8515				0.19			0.19			0.23		
8548				-0.08			0.01			0.18		
8582				0.14			0.29			0.35		
8583				0.13			0.01			0.25		
9014				-0.05			0.45			0.35		
11347				0.03			0.08			0.04		
11350				-0.03			0.09			0.10		
11351				-0.08			0.02			-0.05		
11352				0.04			0.38			0.38		
11357				-0.05			0.09			0.04		
11358	*Dd*	Embryo	5	-1.06	-0.98	0.03	0.13	0.22	0.03	-0.06	-0.01	0.04
11359				-0.94			0.28			0.13		
11360				-0.99			0.22			-0.06		
11361				-0.98			0.26			-0.05		
11363				-0.91			0.23			0.01		
10622		Ovotestis	8	0.05	-0.02	0.08	-0.14	0.09	0.05	0.00	0.13	0.06
10629				0.14			-0.12			-0.14		
10639				-0.33			0.22			0.24		
8554				0.09			0.23			0.40		
8555				0.16			0.14			0.20		
8559				0.22			0.07			0.05		
8562				-0.09			0.06			0.09		
9013				-0.41			0.23			0.20		

11282				-0.89			-1.53			-1.25		
11283				-0.95			-2.30			-1.00		
11284		Embryo	6	-0.85	-0.93	0.03	-1.40	-1.73	0.14	-1.18	-1.23	0.07
11287				-0.90			-1.60			-1.10		
11301				-1.06			-2.02			-1.44		
11303				-0.92			-1.53			-1.44		
11348				0.04			0.04			0.03		
11349				-0.12			-0.06			-0.18		
11353		Foot	5	-0.03	-0.06	0.03	0.18	0.06	0.05	0.14	0.00	0.07
11354				-0.11			-0.03			-0.13		
11356				-0.07			0.18			0.13		
11348				0.09			0.00			0.04		
11349	*dd*			0.17			0.09			0.19		
11353				0.07			0.01			0.09		
11354				-0.26			0.17			0.11		
11356				0.14			-0.25			-0.14		
10626				-0.06			-0.21			-0.19		
10630		Ovotestis	14	-0.02	0.02	0.04	-0.21	-0.07	0.04	-0.10	0.02	0.03
10640				0.06			-0.19			-0.02		
10642				-0.20			-0.30			-0.08		
8806				-0.06			0.04			0.08		
8808				-0.06			-0.07			-0.01		
8996				0.20			0.07			0.29		
9005				0.14			-0.01			0.00		
9007				0.02			-0.11			-0.02		

Table 45 Log-transformed normalised relative quantities (LOG NRQ) for each sample included within the tissue analysis of the three GOIs: *Lcol11a 2/1*, *Lmhc* & *Lmhc nm*, with resulting arithmetic mean LOG NRQ (M) and standard error of the mean (SEM) per group according to genotype (Geno)within the specific tissue. Also presented is the number of individuals within each group (n). The calibrator sample 'OvoRef' was not included in analyses.

Sample and group description				*Lcol11a*			*Lmhc*			*Lmhc nm*		
ID	**Geno**	**Tissue**	**n**	**LOG NRQ**	**M**	**SEM**	**LOG NRQ**	**M**	**SEM**	**M**	**Mean**	**SEM**
11347	*DD*	Foot	10	0.91	0.57	0.15	0.99	0.70	0.10	0.25	-0.02	0.11
11350				0.17			0.54			-0.13		
11351				0.89			0.92			0.23		
11352				0.38			0.56			-0.12		
11357				0.48			0.52			-0.32		
10627		Ovotestis	14	-0.06	0.00	0.06	0.12	-0.03	0.07	-0.13	-0.03	0.02
10633				0.42			0.32			-0.06		
10636				-0.12			0.00			0.02		
10638				-0.12			0.30			-0.04		
8515				-0.18			-0.01			0.04		
8548				-0.25			-0.07			0.06		
8582				0.10			0.22			0.11		
8583				0.23			-0.21			-0.04		
9014				-0.15			-0.23			0.04		
11347				0.01			0.04			-0.04		
11350				-0.04			-0.52			-0.07		
11351				0.18			-0.09			-0.17		
11352				0.36			0.09			-0.02		
11357				-0.31			-0.44			-0.12		
11348	*dd*	Foot	10	0.95	0.83	0.06	0.98	0.85	0.08	0.08	-0.07	0.07
11349				0.86			0.98			-0.10		
11353				0.97			0.89			0.02		
11354				0.67			0.57			-0.30		
11356				0.71			0.83			-0.04		
11348		Ovotestis	14	0.21	0.00	0.08	-0.27	-0.21	0.10	-0.10	-0.01	0.03
11349				0.78			-0.06			0.15		
11353				0.11			-0.70			0.03		
11354				-0.28			-0.73			-0.17		
11356				-0.37			-0.28			-0.09		
10626				0.03			0.10			0.12		
10630				-0.17			0.22			-0.03		
10640				0.10			0.57			0.02		
10642				-0.45			-0.10			0.17		
8806				0.00			-0.25			-0.11		
8808				0.05			-0.09			-0.12		
8996				0.08			-0.28			0.13		
9005				-0.20			-0.62			-0.15		
9007				0.14			-0.43			0.00		

Table 46 Log-transformed normalised relative quantities (LOG NRQ) for each sample included within the tissue analysis of the four GOIs: *Lmyo5a*, *Lmyo18a*, *Lstau* & Lunc93, with resulting arithmetic mean LOG NRQ (M) and standard error of the mean (SEM) per group according to genotype (Geno)within the specific tissue. Also presented is the number of individuals within each group (n). The calibrator sample 'OvoRef' was not included in analyses.

Sample and group description				*Lmyo5a*			*Lmyo18a*			*Lstau*			*Lunc93a*		
ID	Geno	Tissue	n	LOG NRQ	M	SEM	LOG NRQ	M	SEM	LOG NRQ	M	SEM	LOG NRQ	M	SEM
11347				0.80			0.41			0.39			-0.15		
11350				0.01			-0.17			-0.30			-1.45		
11351		Foot	10	0.61	0.33	0.16	0.18	0.02	0.14	0.19	-0.03	0.13	-0.68	-0.89	0.22
11352				0.11			-0.37			-0.22			-1.09		
11357				0.13			0.05			-0.22			-1.06		
10627				0.18			-0.04			0.07			-0.11		
10633				0.28			-0.03			0.22			0.15		
10636				0.13			0.08			0.32			0.15		
10638				0.28			-0.10			0.17			-0.07		
8515	*DD*			0.20			0.29			-0.29			0.03		
8548				0.25			0.40			-0.12			0.20		
8582		Ovotestis	14	0.41	0.21	0.03	0.29	0.08	0.05	0.14	-0.03	0.08	0.42	0.03	0.06
8583				0.26			0.39			0.20			0.48		
9014				0.41			0.22			0.57			-0.34		
11347				0.09			-0.02			-0.30			0.02		
11350				0.17			0.00			-0.37			-0.03		
11351				0.03			-0.12			-0.40			-0.35		
11352				0.13			-0.04			-0.15			-0.07		
11357				0.08			-0.14			-0.48			-0.13		
11348				0.45			0.32			0.13			-0.55		
11349				0.13			0.00			-0.09			-0.78		
11353		Foot	10	0.61	0.42	0.09	0.14	0.18	0.07	0.15	0.03	0.07	-0.57	-0.72	0.10
11354				0.34			0.07			-0.19			-1.08		
11356				0.56			0.37			0.15			-0.62		
11348				0.23			0.02			-0.07			-0.13		
11349				0.41			0.20			-0.02			0.33		
11353				0.33			-0.06			-0.10			-0.25		
11354				0.33			-0.13			0.40			-0.87		
11356	*dd*			0.03			-0.14			-0.45			-0.09		
10626				0.04			0.10			0.07			0.01		
10630		Ovotestis	14	0.05	0.21	0.04	-0.21	0.07	0.05	0.05	0.07	0.07	-0.03	-0.11	0.08
10640				0.21			0.20			-0.02			0.15		
10642				0.01			-0.07			0.08			-0.35		
8806				0.27			0.22			0.38			-0.06		
8808				0.20			0.21			0.25			0.02		
8996				0.42			0.34			0.19			0.05		
9005				0.10			0.02			-0.20			-0.34		
9007				0.32			0.26			0.50			-0.03		

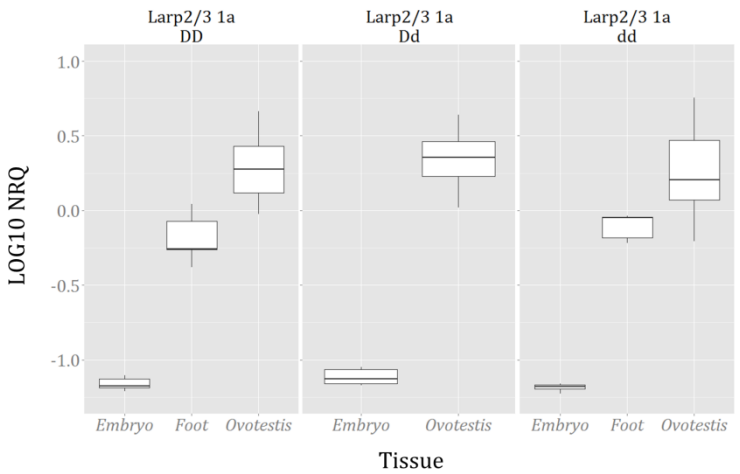


Figure 30 Composite boxplot showing Log scale NRQ values (LOG10 NRQ) for *Larp2/3 1a* in genotypes *DD*, *Dd* & *dd*, compared between embryo, foot and ovotestis tissue, calculated relative to the OvoRef calibrator sample.

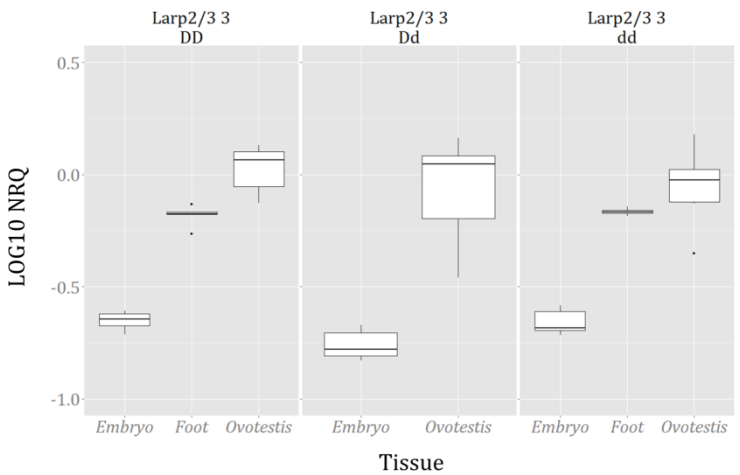


Figure 31 Composite boxplot showing Log scale NRQ values (LOG10 NRQ) for *Larp2/3 3* in genotypes *DD*, *Dd* & *dd*, compared between embryo, foot and ovotestis tissue, calculated relative to the OvoRef calibrator sample.

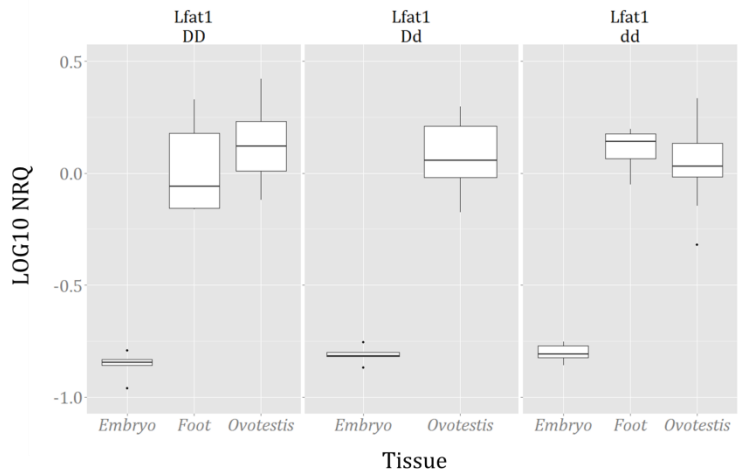


Figure 32 Composite boxplot showing Log scale NRQ values (LOG10 NRQ) for *Lfat1* in genotypes *DD*, *Dd* & *dd*, compared between embryo, foot and ovotestis tissue, calculated relative to the OvoRef calibrator sample.

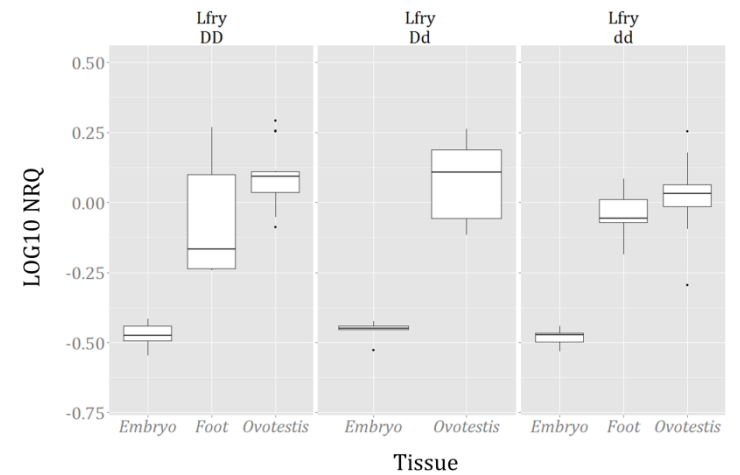


Figure 33 Composite boxplot showing Log scale NRQ values (LOG10 NRQ) for *Lfry* in genotypes *DD*, *Dd* & *dd*, compared between embryo, foot and ovotestis tissue, calculated relative to the OvoRef calibrator sample.

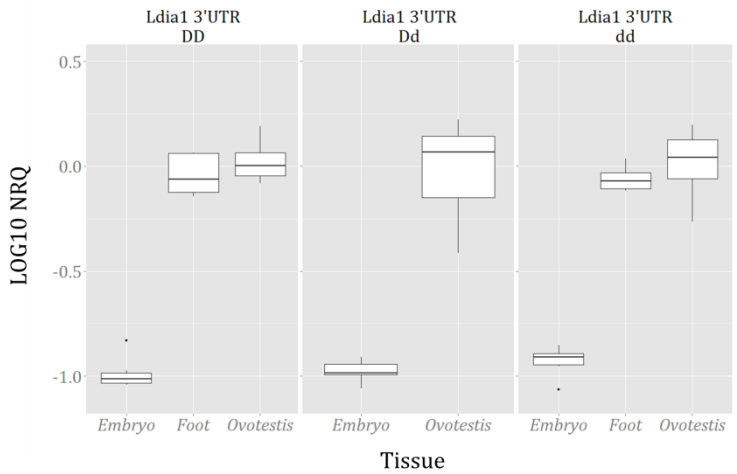


Figure 34 Composite boxplot showing Log scale NRQ values (LOG10 NRQ) for *Ldia1 3' UTR* in genotypes *DD*, *Dd* & *dd*, compared between embryo, foot and ovotestis tissue, calculated relative to the OvoRef calibrator sample.

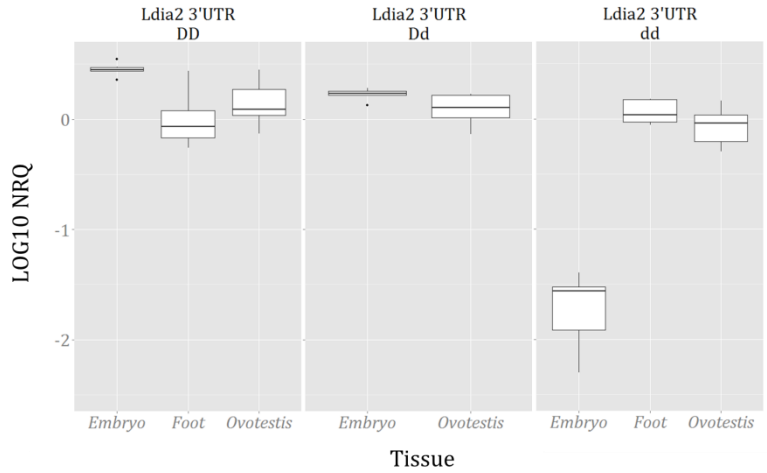


Figure 35 Composite boxplot showing Log scale NRQ values (LOG10 NRQ) for *Ldia2 3' UTR* in genotypes *DD*, *Dd* & *dd*, compared between embryo, foot and ovotestis tissue, calculated relative to the OvoRef calibrator sample.

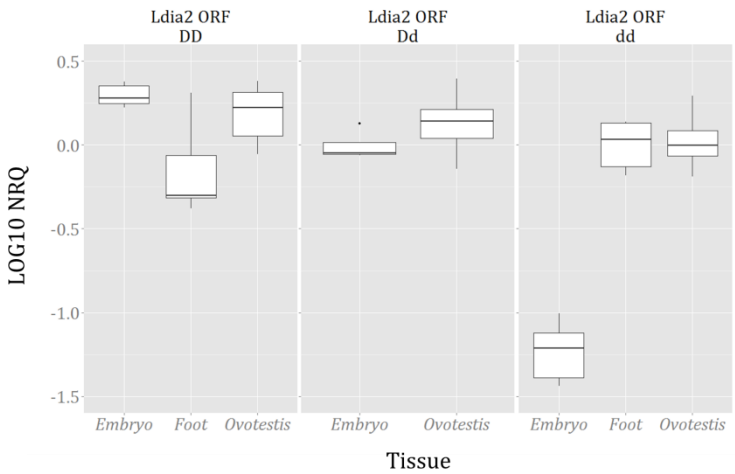


Figure 36 Composite boxplot showing Log scale NRQ values (LOG10 NRQ) for *Ldia2 ORF* in genotypes *DD*, *Dd* & *dd*, compared between embryo, foot and ovotestis tissue, calculated relative to the OvoRef calibrator sample.

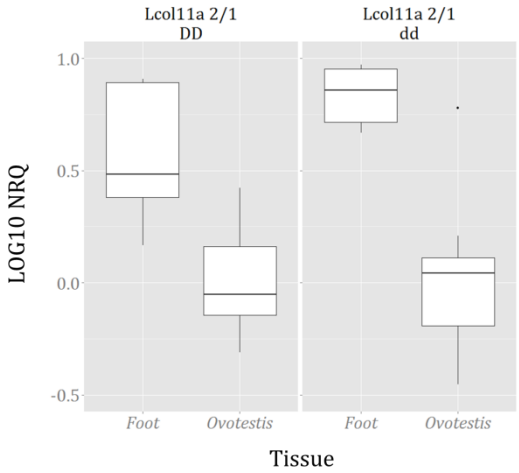


Figure 37 Composite boxplot showing Log scale NRQ values (LOG10 NRQ) for *Lcol11a 2/1* in genotypes *DD* & *dd*, compared between foot and ovotestis tissue, calculated relative to the OvoRef calibrator sample.

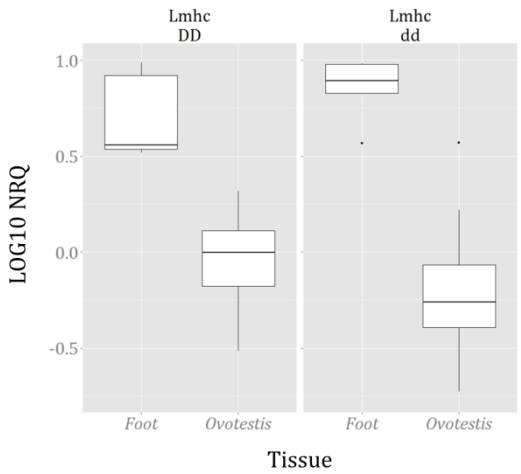


Figure 38 Composite boxplot showing Log scale NRQ values (LOG10 NRQ) for *Lmhc* in genotypes *DD* & *dd*, compared between foot and ovotestis tissue, calculated relative to the OvoRef calibrator sample.

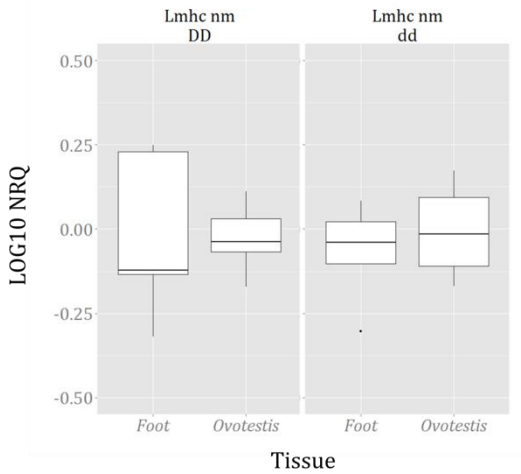


Figure 39 Composite boxplot showing Log scale NRQ values (LOG10 NRQ) for *Lmhc nm* in genotypes *DD* & *dd*, compared between foot and ovotestis tissue, calculated relative to the OvoRef calibrator sample.

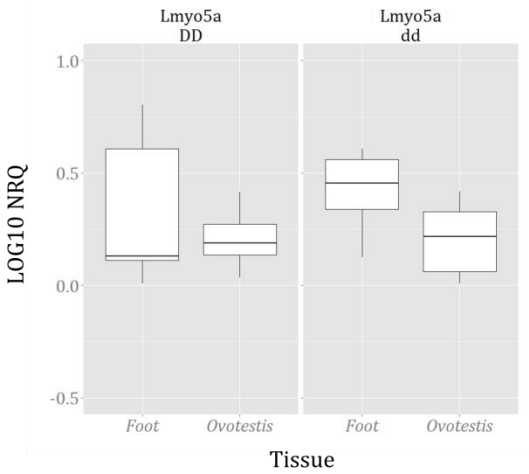


Figure 40 Composite boxplot showing Log scale NRQ values (LOG10 NRQ) for *Lmyo5a* in genotypes *DD* & *dd*, compared between foot and ovotestis tissue, calculated relative to the OvoRef calibrator sample.

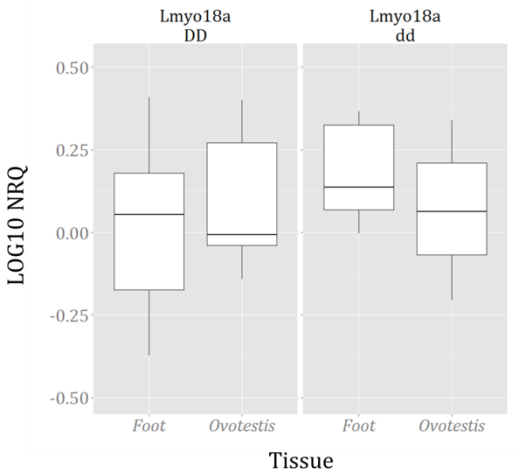


Figure 41 Composite boxplot showing Log scale NRQ values (LOG10 NRQ) for *Lmyo18a* in genotypes *DD* & *dd*, compared between foot and ovotestis tissue, calculated relative to the OvoRef calibrator sample.

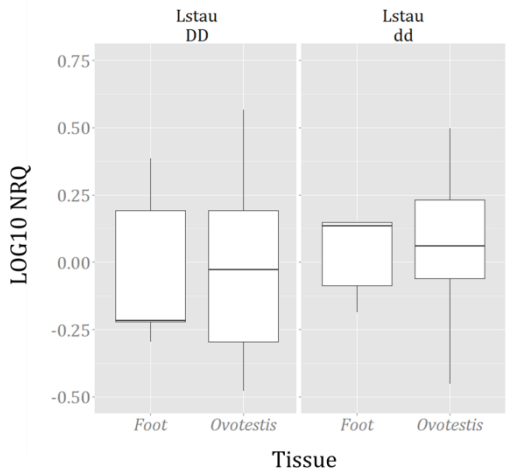


Figure 42 Composite boxplot showing Log scale NRQ values (LOG10 NRQ) for *Lstau* in genotypes *DD* & *dd*, compared between foot and ovotestis tissue, calculated relative to the OvoRef calibrator sample.

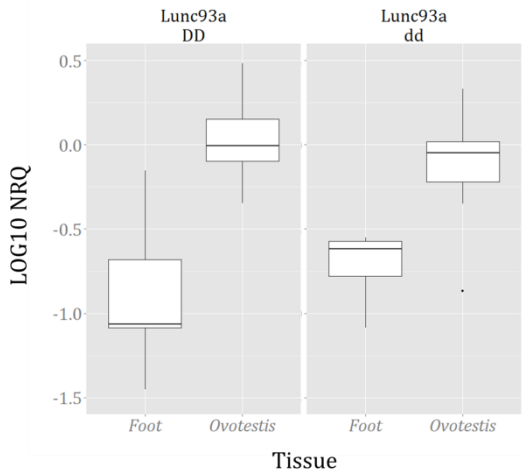


Figure 43 Composite boxplot showing Log scale NRQ values (LOG10 NRQ) for *Lunc93a* in genotypes *DD* & *dd*, compared between foot and ovotestis tissue, calculated relative to the OvoRef calibrator sample.

Table 47 Wilcoxon rank test results for pairwise comparisons between embryo, foot and ovotestis tissue within genotypes *DD*, *Dd* and *dd* for seven GOIs. The total number of individuals within each genotype group is quoted (n) in addition to the number of individuals within each tissue specific genotypic group (n, Embryo; n, Ovotestis; n, Foot). The Wilcoxon rank value (W) is presented with the associated probability value (p). Statistical significance (sig) is highlighted via * <0.05, ** <0.01, * <0.001. Probability values are presented to 3 decimal places, thus '0.000' represents <0.001.**

Tissue Analysis GOI 1-7																	
GOI	**Geno**	**n**	**Embryo / Ovotestis**					**Embryo / Foot**					**Ovotestis / Foot**				
			n, Embryo	**n, Ovotestis**	**W**	**P**	**sig**	**n, Embryo**	**n, Foot**	**W**	**P**	**sig**	**n, Ovotestis**	**n, Foot**	**W**	**P**	**sig**
Larp2/3 1a	*DD*	25	6	14	0	0.000	***	6	5	0	0.004	**	14	5	67	0.001	**
	Dd	13	5	8	0	0.002	**	n/a									
	dd	25	6	14	0	0.000	***	6	5	0	0.004	**	14	5	63	0.007	**
Larp2/3 3	*DD*	25	6	14	0	0.000	***	6	5	0	0.004	**	14	5	70	0.000	***
	Dd	13	5	8	0	0.002	**	n/a									
	dd	25	6	14	0	0.000	***	6	5	0	0.004	**	14	5	65	0.003	**
Ldia1 3'UTR	*DD*	25	6	14	0	0.000	***	6	5	0	0.004	**	14	5	48	0.257	
	Dd	13	5	8	0	0.002	**	n/a									
	dd	25	6	14	0	0.000	***	6	5	0	0.004	**	14	5	52	0.130	
Ldia2 3' UTR	*DD*	25	6	14	79	0.001	***	6	5	28	0.017	*	14	5	52	0.130	
	Dd	13	5	8	34	0.045	*	n/a									
	dd	25	6	14	0	0.000	***	6	5	0	0.004	***	14	5	18	0.130	
Ldia2 ORF	*DD*	25	6	14	59	0.179		6	5	26	0.052		14	5	60	0.019	*
	Dd	13	5	8	9	0.127		n/a									
	dd	25	6	14	0	0.000	***	6	5	0	0.004	**	14	5	35	1.000	
Lfat1	*DD*	25	6	14	0	0.000	***	6	5	0	0.004	**	14	5	46	0.343	
	Dd	13	5	8	0	0.002	**	n/a									
	dd	25	6	14	0	0.000	***	6	5	0	0.004	**	14	5	26	0.444	
Lfry	*DD*	25	6	14	0	0.000	***	6	5	0	0.004	**	14	5	49	0.219	
	Dd	13	5	8	0	0.002	**	n/a									
	dd	25	6	14	0	0.000	***	6	5	0	0.004	**	14	5	49	0.219	

Table 48 Wilcoxon rank test results for pairwise comparisons between foot and ovotestis tissue within genotypes *DD* and *dd* for seven GOIs. The total number of individuals within each genotype group is quoted (n) in addition to the number of individuals within each tissue specific genotypic group (n, Ovotestis; n, Foot). The Wilcoxon rank value (W) is presented with the associated probability value (p). Statistical significance (sig) is highlighted via * <0.05, ** <0.01, * <0.001. Probability values are presented to 3 decimal places, thus '0.000' represents <0.001.**

Tissue Analysis GOI 8-14							
GOI	**Geno**	**n**	**Ovotestis / Foot**				
			n, Ovotestis	**n, Foot**	**W**	**P**	**Sig**
Lcoll11 2a	*DD*	19	14	5	5	0.003	**
	dd	19	14	5	2	0.001	**
Lmhc	*DD*	19	14	5	0	0.000	***
	dd	19	14	5	1	0.000	***
Lmhc nm	*DD*	19	14	5	39	0.754	
	dd	19	14	5	41	0.622	
Lmyo5a	*DD*	19	14	5	36	0.964	
	dd	19	14	5	11	0.026	*
Lstau	*DD*	19	14	5	33	0.893	
	dd	19	14	5	38	0.823	
Lmyo18a	*DD*	19	14	5	39	0.754	
	dd	19	14	5	23	0.298	
Lunc93	*DD*	19	14	5	68	0.001	**
	dd	19	14	5	66	0.002	**

Comparison to embryonic tissue

The relative expression levels observed in five of the seven GOI targets assessed in the embryo, ovotestis and foot tissue were found to be significantly reduced in all genotypic groups of the embryo tissue compared to both the ovotestis tissue and the foot tissue (Figure 30, Figure 31, Figure 32, Figure 33, Figure 34, Table 47).

Only the diaphanous formin, *Ldia2*, targets demonstrated an alternative expression pattern. The *DD* embryo samples showed an increased relative expression of *Ldia2 3' UTR* and *Ldia2 ORF* compared to the ovotestis and foot tissue (Figure 35, Figure 36). This increase was found to be statistically significant for the 3'UTR target between all tissues, whereas the ORF target expression difference was only found to be statistically significant between the embryo and foot tissues (Table 47).

The *dd* embryo samples however, showed significantly underrepresented levels of *Ldia2* 3' UTR and *Ldia2* ORF compared to the ovotestis and foot tissue (Figure 35, Figure 36, Table 47), similarly to the expression patterns seen in the other GOIs.

The expression pattern of *Ldia2* observed in the *Dd* samples was a little more convoluted. *Ldia2 3' UTR* was found to be significantly over-expressed in the *Dd* embryo samples compared to the

ovotestis (the same pattern as seen in the *DD* embryos, Table 47, Figure 35). However, *Ldia2* ORF was shown to be under-expressed in the *Dd* embryo samples compared to the ovotestis, although this difference was not found to be significant (Figure 36, Table 47). No comparisons were available for the *Dd* individuals in the foot tissue.

Comparison of the ovotestis and foot tissues

Comparisons between the ovotestis and foot were less pronounced than those to the embryo tissue, and generally exhibited larger variation. Comparisons could only be made between homozygote genotypes.

A significant difference in the relative expression was identified for the *DD* samples in the *Ldia2 ORF*, yet not in the *Ldia2 3'* UTR target (Table 47). Both targets however, showed relative expression was greater in the ovotestis than in the foot (Figure 35, Figure 36).

The actin related proteins, *Larp2/3 1a* and *Larp2/3 3* both exhibited highly significant increased relative expression in the ovotestis tissue compared to the foot tissues (Table 47). Both genotype groups showed the same expression pattern (Figure 30, Figure 31).

Lcol11a 2/1 showed highly significant difference in expression between tissues. The foot tissue showed a greater than three-fold increase in expression compared to the ovotestis tissue (Table 42). Both genotype groups demonstrated the increased expression, although it was more pronounced in the *dd* samples (Table 48, Figure 37).

The ovotestis tissue showed a five-fold increase in expression of Lunc93a compared to the foot tissue (Table 42, Figure 43). This expression difference was found to be highly significant in both of the genotypic groups assessed (Table 48).

Lmyo5a was found to be significantly differentially expressed only in the *dd* tissue comparison. The *dd* samples exhibited an increased expression in the foot compared to the ovotestis, whereas the *DD* samples displayed the opposite pattern, although compromised by high variability in the foot samples and not statistically significant (Figure 40). The expression difference in the *dd* represents a two-fold increase in expression in the foot tissue.

Lmhc was found to be over-expressed in the foot tissue compared to the ovotestis, by more than a five-fold increase (Table 42). This expression difference was found to be highly statistically significant in both genotype comparisons, yet more pronounced in the *dd* samples (Table 48), although there were outliers present in both tissues in *dd* (Figure 38).

Discussion

Differential gene expression between genotypes

The general lack of differential expression between genotypes in the majority of GOIs suggests that there may be less pleiotropic effects associated with the *dd* individuals in early development than assumed based upon their reduced hatch rate (Davison, Barton et al. 2009). However, it is important to note that only the very early developmental stages were assessed here and there may be downstream differences in expression associated with genotype not detected here.

Diaphanous related formin, *Ldia2*

The dramatically reduced expression of *Ldia2* in the *dd* embryo samples of both the 3' UTR target and the ORF, coupled with the absence of any differential expression in the other candidate genes within the chirality locus, supports the hypothesis that the frameshift mutation in *Ldia2* is strongly associated with the genetic determinant of LR asymmetry in *L. stagnalis*.

The heterozygote, *Dd*, samples in both the ovotestis and embryo tissue exhibited an expression level almost exactly halfway between that observed in *DD* and *dd* for both 3' UTR and ORF. This suggests that both copies of the gene are being equally transcribed/regulated, as opposed to the silencing of one copy in the homozygote, *DD*. This expression pattern further supports the tight association of this gene with the chirality phenotype.

The single deletion in the sinistral version of *Ldia2* creates a frameshift mutation very early in the coding region of the gene. However, it was unknown whether or not this would lead to differences in the quantity of transcript because the frameshift would not necessarily prevent or hinder transcription, yet simply generate a missense transcript that would likely result in an inability to form the required protein. Nonsense mediated mRNA decay (NMD) is known to be triggered by the presence of a premature stop codon in the transcript, although this process is still poorly understood (Lykke-Anderson and Jensen 2015). The sequence alignment of the *Ldia2* gene in *L. stagnalis* has indicated that the frameshift does result in a premature stop codon within the first several amino acids (*alignments not presented*). Therefore, it seems likely that the reduced level of *Ldia2* in the *dd* samples is due to NMD of the recessive, *d*, transcript. This would fit the observed halved quantity seen in the *Dd* samples also.

The inclusion of the 3' UTR and the ORF of *Ldia2* may be able to provide support to this hypothesis based on the expected direction of NMD. In the ovotestis tissue, the 3' UTR generally showed a greater reduction in expression in the *Dd* and *dd* samples relative to *DD*, than the ORF (Table 35).

This was also seen in the *dd* samples in the embryo tissue (Table 33), although the accuracy of these fine scale differences may be compromised due to the high Cq values and resulting error rate observed in these samples. The pattern observed in the *Dd* embryo samples conversely showed a greater reduction in the ORF compared to the UTR (Table 33). The scale of these differences, are generally not very large and therefore the inferences are limited yet show potentially increased NDM in the 3' UTR compared to the ORF.

Comprehensive studies of NMD have not been performed in molluscs. However, in NMD study organisms, from yeast to mammals, decay is observed in both a 5' to 3' and 3' to 5' direction of the mRNA, originating from either the 3' end or exon-exon boundaries (Elliot and Ladomery 2011). The variation in starting position of NMD limits interpretation of the differences between the reduction *Ldia2* in the 3' UTR and ORF.

However, the frameshift in the sinistral *Ldia2* will be present in all tissues, and therefore the resulting NMD would be expected to be taking place in all tissues. The lack of quantitative differences in the foot tissue suggests some other form of regulation may be occurring in the embryo and ovotestis tissue. Alternatively, the process of NMD may be obscured in actively transcribing tissues. The embryo tissue only contains maternal mRNAs transcribed prior to laying. Therefore, in the absence of newly transcribed *Ldia2*, the process of NMD on the mutated gene has enough time to result in significantly different levels of the transcript. And so it is proposed that the ovotestis contains both actively transcribing tissues and eggs containing already deposited transcripts and subsequently reveals a reduced effect to that seen in just the embryo tissue.

This hypothesis would be greatly supported by the inclusion of later embryonic stages in the qPCR experiments, holding the potential to reveal an increasing reduction of the sinistral *Ldia2* until the onset of zygotic transcription produces new copies of the transcript which may counteract the low levels of the gene or reveal alternative patterns of gene regulation.

Another explanation for the lack of DE between genotypes in the foot tissue, is that the transcript is in very low copy number and effectively not present in the foot, as is the case for the *dd* embryo tissue. This is supported by the similarly high Cq values observed in the foot tissue and *dd* embryo tissue for *Ldia2* (Table 22, Table 24).

Actin-related protein 2/3 complex

Significant expression differences of actin-related proteins between genotypes were only identified in the embryo tissue. This may reflect a downstream effect associated with of the different

quantities of *Ldia2* present to a greater extent in the embryo tissue. However, the differential expression observed only in the heterozygotes does not intuitively fit the linear reduction in *Ldia2* seen across the genotypes.

Larp2/3 1a showed increased level of transcript in the heterozygote whereas *Larp2/3 3* showed an increased level of transcript in the homozygote (Table 32). The alternate direction of differential expression observed in the two *arp2/3* targets here, support that the *arp2/3* subunits are differentially regulated (Gournier, Goley et al. 2001). However, the scale of the expression differences are very small (<1.5 fold change) and subsequently would require further analysis to justly infer any biological meaning of the relationship.

Differential gene expression between tissues

Ldia2 3' UTR and *Ldia2 ORF* were the only targets not to show significantly reduced expression in the wild-type (*DD*) embryo tissue compared to the somatic tissue. This highlights its functional importance in the early developmental stages in *L. stagnalis* providing strong support for *Ldia2* as the primary candidate for the chirality gene, above any candidate genes assessed here. The overall large reduction of relative quantity of the remaining GOI transcripts in the embryo tissue is assumed to reflect an economic strategy of providing sufficient, yet not excessive quantities of maternal transcripts per embryo.

Both of the arp2/3 transcripts were found in greater quantities in the ovotestis than in the foot and embryo tissue. This may reflect functional associations of the arp2/3 complex in cell motility and ultimately sperm motility (Lee, Kwon et al. 2015)

The increased relative expression of *Lunc93a* in the ovotestis compared to the foot tissue is in keeping with its previously identified association with the ovarian tissue (Liu, Dodds et al. 2002). Additionally, the increase of *Lcol11a 2/1* in the foot tissue echoes the contractile muscle functions associated with connective tissues abundant in collagen (Rigon, Manica et al. 2010). Similarly, the overexpression of *Lmhc* in the foot tissue demonstrates the expected increase of muscle proteins in the foot tissue compared to the ovotestis.

With the exception of the GOIs described above the expression differences between ovotestis and foot were less pronounced than comparisons to the embryo tissue. Over half of the comparisons showed no significant difference in expression. These results further highlight the overall greatly reduced relative expression level of the transcripts seen in the embryo tissue compared to both the ovotestis and foot tissue with the notable exception of *Ldia2*.

Differential expression of GOIs

Where NRQ values are calculated relative to the same calibrator sample, such as in the tissue analysis, the differences in expression level of transcripts can be compared to each other. For example, the relative quantities of the two copies of diaphanous formin in the embryo tissue can be compared by their expression values relative to the ovotestis tissue. In the wild-type, *DD*, embryo *Ldia2* transcripts were found more than 20 times higher than *Ldia1* (Table 42). This provides a quantitative value for the overexpression observed of *Ldia2* in the embryo relative to other GOIs in the embryo.

The relative expression of both the *Ldia2 3' UTR* and *Ldia2 ORF* in the *dd* embryos is lower than any other GOI assessed (Table 32, Table 42). Therefore, it is apparent that *Ldia2* is overrepresented in the *DD* embryo samples, yet also underrepresented in the *dd* embryo samples compared to the ovotestis and foot tissue.

Quality controls

Many of the issues surrounding the quality assessment of the samples have already been discussed in the previous chapter, due to the overlap of samples used. Discussed here are further comments relating specifically to the additional samples/GOIs used or their experimental application.

Extraction method

Differences between the extraction methods across the tissues, may lead to differences in sample quality and composition. The RNeasy extraction kits use columns to bind and extract total RNA from tissue. These columns are unable to retain small RNAs <200 bases long (Qiagen 2007), whereas the TRI Reagent method theoretically will retain all RNA. The TRI Reagent method also retains carryover genomic DNA as seen in the intronic PCRs from the foot and ovotestis samples. These differences should not pose a problem in the genotype comparisons as no systemic bias will have been introduced given that all samples within the tissue have been treated the same. However, when making comparisons between tissues, each tissue has been extracted using different method and therefore, differences in sample quality may create an undetected confounding variable, especially regarding the differences in genomic carryover.

Genomic carryover

Every embryo tissue sample used within this experiment failed to produce a PCR product from the intronic PCR and can therefore be assumed to be free of genomic contamination. This eliminates potential errors in quantification of expression of *Ldia1 3' UTR* and *Ldia2 3' UTR* due to amplification

of a genomic product at the same size. The foot and ovotestis samples however did amplify genomic products. The results of the genomic PCR tests are presented and further discussed in the SI (S5).

Importantly the positive control test PCRs showed that no cDNA sample amplified two size specific products (presented in the SI, S5, and in Figure 11 and Figure 12). Therefore, the genomic carryover observed in the foot and ovotestis samples was assumed to be outcompeted by the cDNA product and of little concern for the accurate quantification of the majority of GOIs in this experiment. However, it must be acknowledged that the *Ldia1 3' UTR* and *Ldia2 3' UTR* experiments will inevitably include quantification of both genomic and transcriptomic templates due to the lack of intron-spanning primers for these GOIs. Again this should not affect comparisons within tissues as all genotypes have been treated the same, yet due to the difference in level of genomic carryover between tissues, this may represent a systemic bias.

mRNA enrichment

The ongoing genomic analyses of *Ldia1* and *Ldia2* in the Davison research group, have revealed both to be large genes, more than 7 kilobases (kb) long (Davison et al, *awaiting publication*) and may be up to 9 kb in total (Northern Blot analysis presented in the SI, S1). When designing primers for the two genes, the 3' UTR was selected instead of the 5' UTR because it lies closer to the poly-A tail present at the end of the mRNA molecule. Any mRNA selection when generating cDNA may make it difficult to reverse-transcribe the 5' end of the long gene due to its increased distance from the selected poly-A tail (as described in Chapter 2, Introduction, Sample quality).

The method of cDNA synthesis employed here, utilised a combination of oligo dTs and random hexamers and as such should generate a more balanced coverage of the RNA molecules. Still the position of the qPCR primers at the 3' UTR was hoped to reduce the bias of the transcript being underrepresented as a result of insufficient reverse-transcription and obscuring true patterns of gene expression.

Primer specificity

In addition to verifying that the amplicons were of the expected size via gel electrophoresis (Figure 11, Figure 12), the amplicons of *Ldia2* and *Ldia1* were sequenced by Sanger sequencing to further verify specificity. This was important for *Ldia2* 3' UTR and ORF because of the multiple Tm peaks seen in some in the *dd* samples (Figure 15). Additionally, it was unknown whether the *Ldia2* ORF primer pair would amplify a product only from *Ldia2* or also produce products from the highly similar *Ldia1*. The Sanger sequences indicated specificity to *Ldia2* (sequences not presented). The spurious Tm peaks are assumed to represent various primer dimers resulting from the very low

concentration of the target gene present in the *dd* embryo samples. This is supported by the Sanger sequencing of non-quantitative PCR products and the general lack of multiple peaks in the other samples (Figure 15).

Due to the non-specific nature of the SYBR reporter dye, the amplification of primer dimer will contribute to the Cq values in these samples. Therefore the NRQs of the *dd* samples in *Ldia2 3' UTR* and *Ldia2 ORF* will contain a level of inaccuracy. To gain a more precise estimate of the fold-change difference in expression between genotypes of *Ldia2* the experiment could be repeated at a much higher cDNA concentration.

The Tm curves generated from the qPCR experiments for all other GOIs show sharp peaks, with the exception of the low levels peaks seen in some of the negative controls (Figure 13, Figure 14). Again, because these spurious peaks only occur in the negative controls and not in the experimental sample, they are assumed to represent various primer dimers amplified by the primers in the absence of the target sequence. Therefore the Cq values of all experimental GOIs (potentially excluding the *dd* samples in *Ldia2 3' UTR* & *Ldia2 ORF*) are not believed to be compromised by the amplification of any non-specific products.

Experimental Design

Amplification efficiencies and working dilution

The Cq values generated from the serial dilutions in the amplification efficiency experiments (data not shown) indicate the appropriate range of dilutions that can be used for the specific GOI. The acceptable range of Cq values (10-35 (AppliedBiosystems 2011)) was mostly included within the range of dilutions resulting in an acceptable amplification efficiency.

The primer efficiencies were calculated from a minimum of two standard curve experiments using a sample representing each genotype in the experiment. This is to ensure that the transcript will be present. However, for the most part primer efficiencies were only calculated within one tissue (with the exception being *Ldia2* 3' UTR, Table 20). Therefore, assumptions regarding acceptable sample dilutions are strictly only appropriate for the tissue, and even genotype, that the amplification efficiency was calculated in. This became apparent when performing the *Ldia2* 3' UTR experiment. An initial qPCR was performed using a 1:30 working dilution (data not presented). This dilution was regarded as acceptable according to the amplification efficiency experiment performed within embryo and foot samples (Table 20). However, due to the very low level of transcript observed in the *dd* samples, reliable quantification could not take place and the experiment was repeated at a

higher concentration (1:15). Similarly, having established an acceptable working dilution for *Lmhc* in the ovotestis tissue, it could not be assessed in the embryo tissue as the level of the transcript was too low for accurate quantification in any of the samples.

Ideally amplification efficiency experiments would be performed in all tissues they are to be assessed in to establish the appropriate working dilution. However this is often not possible due to limited sample and resources. The embryo sample in this experiment was very limited and therefore could only be used in a subset of amplification efficiency experiments.

Plate setup

The sample sizes of the embryo and the foot experiment especially, were small. It was therefore very important to minimise error and so technical replicates were performed in triplicate. The ovotestis experiment however, had far more samples included. Reducing the technical replicates from triplicate to duplicate saved 37 wells per GOI, and collectively 629 wells (7 experimental plates) in the whole study. The more informative data from the increased number of biological replicates was considered to outweigh the loss of information through reduction in technical replicates. However, reducing the technical replicates to two precluded the identification of outliers and therefore the cleaning of 'noisy' data. Due to the importance of the OvoRef sample, which functioned as the calibrator for the whole plate and tissue analysis, it was not reduced to duplicate repeat (Figure 10).

Targets/GOIs

Choice of GOIs

Only four of the fourteen GOIs assessed revealed expression differences between genotypes. Whilst this provides good support for the role of the candidate gene in LR asymmetry variation, there may be more informative target genes not included here. For example, no tubulins were included in the analyses, despite their potential involvement in LR determination (Vandenberg, Lemire et al. 2013). Yet the 2004 Shibazaki experiment (Shibazaki, Shimizu et al. 2004), concluded that tubulins were not essential for the establishment of asymmetry in *L. stagnalis* and therefore were not included as a priority in this experiment. Additionally, the nine 'functional targets' were originally identified through DE analysis of the eRAD dataset which has since been corrected and has not identified any of the targets here as DE.

Three additional genes in close proximity to the candidate locus have recently been identified, in the latest *L. stagnalis* genome annotation within the Davison research group (June 2015). These genes

will also be assessed for differential expression between genotypes in order to verify the current main candidate, *Ldia2*, however due to time limitations they could not be included in this analysis.

Two regions of the same GOI, *Ldia2* were included in this experiment, the 3' UTR and the ORF. Including both regions was hoped to elucidate any bias in patterns of gene regulation or directional degradation, such as nonsense mediated decay. Additionally, the inclusion of the two regions provided a form of repeat analysis and therefore greatly reduced the probability of the identified differential expression being identified as statistically significant due to chance.

Because of the reduced amount of embryo sample available, eight priority GOIs had to be selected for assessment within the embryo tissue. To support the role of the primary candidate *Ldia2* as the causal gene in LR determination, firstly *Ldia2* itself was a priority for assessment of differential expression between genotypes (both the 3' UTR and ORF were included for reasons above). Subsequently the other genes surrounding the chirality locus, *Ldia1*, *Lfat1* and *Lfry*, aimed to be excluded as candidates, through demonstrating no significant difference in expression. In addition to the proximal targets, both of the actin-related proteins were included in the embryo tissue. This was due to the previous studies specifically highlighting the importance of actins in early developmental stages of *L. stagnalis* (Shibazaki, Shimizu et al. 2004, Davison et al, awaiting publication), and to identify potential functional interactions with the diaphanous-related formin, *Ldia2*. *Lmhc* was included as the final GOI to be assessed in the embryo tissue. This decision was based on preliminary qPCR experiments in the ovotestis tissue (data not presented), which indicated potential differential expression between genotypes. The unforeseen low expression levels of *Lmhc* in the embryo tissue negated the inclusion of this GOI in the embryo study and further emphasises the importance of performing amplification efficiency experiments in representatives of all samples/tissues.

Choice of endogenous controls

The major caveat of the tissue comparison experiment is that the endogenous control genes were not tested for stability between tissues. Again due to finite timescales and resources not all controls can be performed. The stability assessment of the endogenous control genes revealed each of those used within this experiment to be stable across genotypes within each tissue and generally showed the same genes to be the most stable in all three tissues (Table 14). It seems unlikely therefore that there would be substantial variation in gene expression in the selected endogenous control genes between tissues. Furthermore, the tissue analysis did not appear to show substantially higher levels of within group variation than the genotypes analysis.

More importantly any confounding variation of the endogenous control genes has not resulted in an apparent systematic bias, in light of the bidirectional patterns of differential gene expression observed across tissues. For example, if both of the endogenous control genes used in the tissue comparison happened by chance, to be overrepresented in the embryo tissue compared to the foot tissue, this could create the observed reduced relative expression of the majority of the GOIs quantified in the embryo tissue (Figure 30-Figure 34). However, the increased relative expression of *Ldia2* in the embryo tissue compared to the foot tissue makes this highly unlikely (Figure 35, Figure 36).

The endogenous control genes included were selected based principally on the results of the geNorm stability assessment in order to ascertain the three best stability genes to include (Table 11, Figure 8). The foot and embryo tissue samples were quantified on the same experimental plate. This necessitated the use of the same endogenous control genes across the two tissues, or alternatively performing separate experiments for the endogenous control genes although this was not an economical option. Because all of the endogenous control genes were found to be highly stable in the embryo tissue the priority for gene selection was based largely on the foot tissue stability assessment. Therefore, *Lywhaz*, *Lube2* and *Lhis2a* were selected.

The endogenous controls selected for the ovotestis experiment employed the two most stable genes identified by geNorm *Lrpl14* and *Lube2*. However, *Lhis2a* was included instead of *Lywhaz* due to having the best individual stability score (Table 11). Although the genotype analysis allowed for different endogenous control genes to be employed between the tissues, the tissue analysis required all samples to be normalised to the same control genes. This reduced the number of endogenous controls used in the tissue analysis to two genes. With the exception of *Lef1a*, all of the endogenous control genes tested were considered to be stably expressed in all tissues. Therefore, employing the same three control genes in all tissues may have provided a better method.

Choice of Tissues

The choice of tissue in which to assess differential expression will be based on which target genes/functions are being explored. A priority in this experiment was supporting candidate genes as the maternal effect gene known to determine chirality in *L. stagnalis*. Therefore, unfertilised eggs (ovotestis) and single cell embryos were essential to isolate maternal transcripts.

Embryo tissue

The single cell embryo tissue provided an ideal sample tissue, in that it was very clean. Following decapsulation the chance of any somatic tissue carryover was minute. It has also been documented

that zygotic expression does not begin until the 24 cell stage (Morrill 1982) and therefore will only contain maternal transcripts transcribed prior to egg-laying. This was essential for looking at expression patterns of the causal gene, however there may be downstream effects of chirality on gene expression only apparent at later stages, for example after the third cell cleavage, which signposts a 'point of no return' for the development of LR asymmetry (Kuroda, Endo et al. 2009). It would be highly informative to include later developmental stages, especially the 4-8 cell stage, and possibly post-zygotic transcription stages. However, because the embryo collection process was very time and labour intensive these additional experiments were not possible within this project.

Foot tissue

The foot tissue appears to have provided a successful somatic tissue control due to the lack of any significant difference in gene expression found between genotypes. However, the small sample size and high levels of variation observed within the group may have obscured possible small scale differences in gene expression. It is unclear why the foot tissue presented such variable data. All samples within the genotype groups were from the same homozygote population and prepared within days of each other following the same protocols (Table 16). The foot tissue is also easily distinguishable from other tissues during extraction, yet there may be problems associated with the presence of mucous in the tissue (AppliedBiosystems 2010). It is probable that the high levels of variability relate to the TRI Reagent RNA extraction observed to produce variable levels of chemical and genomic material carryover. The 260/230 ratios recorded for the foot tissues do show a range of values generally lower than seen in the embryo samples yet not as low as those of the ovotestis samples (Table 15, Table 16, Table 17).

Ovotestis tissue

The ovotestis tissue was included to assess differences in gene expression of actively transcribing maternal RNAs. As previously discussed, the timing of ovotestis extraction was unable to be controlled beyond the individual snail being reproductively mature. Additionally, due to the internal organisation of *L. stagnalis*, this tissue is impossible to isolate without the inclusion of contaminating liver tissue (Figure 9). Both of these factors are likely contributors to the higher levels of within-group variation seen in the ovotestis samples.

Data Analysis

There are numerous ways of determining relative Cq values (Livak and Schmittgen 2001, Pfaffl 2001, Schefe, Lehmann et al. 2006, Hellemans, Mortier et al. 2007). The updated Pfaffl method (Hellemans, Mortier et al. 2007) was employed here because it is the only method which readily

incorporates primer amplification efficiency and enables normalisation to multiple endogenous control genes. This method can also include the SD of the average Cq throughout the calculations. Technical SD was not included in the creation of the NRQs here. Due to the generally low SD of the majority of raw average Cq values (<0.5), the effect of this technical variation is assumed to be negligible.

There are a number of software packages available to analyse relative qPCR data in addition to that which is included in the ABi 7500 qPCR fast Real-Time PCR system (Hellemans, Mortier et al. 2007, Pfaffl, Vandesompele et al. 2009), including the freely available REST software (Pfaffl, Horgan et al. 2002). Due to the limited flexibility of this software, often assuming that the requirements for parametric statistical tests are met, and the relative simplicity of the statistics undertaken to establish significance of differential expression, the statistical analyses here were performed using R software.

Statistical analyses were only performed on log transformed NRQ values. This is advised due to the non-linear distribution of relative Cq data (Rieu and Powers 2009). An NRQ value of 1 indicates a 1:1 expression ratio, therefore no difference in expression. Any gene that is down-regulated will be represented by a value between 0 and 1, whereas upregulated genes will be presented by any number greater than 1, which skews the distribution and often results in heterogeneity of variance (Rieu and Powers 2009).

Genotype Analysis

Because all samples of each tissue were included on the same experimental plate, any sample could be utilised as the calibrator for within tissue comparisons (Figure 10). An individual *DD* sample was used as a calibrator for the genotype analysis as this represents the wild-type condition. It is possible to include multiple samples as calibrators in order to minimise variation, similar to employing multiple endogenous controls (Vandesompele, De Preter et al. 2002, Hellemans, Mortier et al. 2007), although only one sample was utilised here. The single cell embryo pool was not included in any statistical analyses. It was originally included to function as a calibrator sample, although a *DD* sample provided a more intuitive relative value representing the homozygote dominant genotype. Because the single cell embryo pool was generated from the existing *Dd* embryo samples it would have presented a pseudo-replicate if included in the group means.

Non-parametric Wilcoxon-Mann-Whitney tests were performed for all group comparisons. The sample sizes of the embryo and foot experiments were too small to be considered normally distributed and so the more robust non-parametric test was performed to compare group means.

The ovotestis data however, had a greater sample size and potentially may have been appropriate for a parametric T-test. Histograms were generated of the NRQs of the ovotestis samples for each of the GOIs (presented in the SI, S10). Although for some GOIs a general bell-curve could be seen across all samples, this was not maintained within the genotype groups and so the more robust non-parametric test was employed.

The ovotestis data exhibited fairly high levels of variation shown in the generally wide boxplots in the majority of GOIs (Figure 16 -Figure 29). The ovotestis experiment included samples from different mating crosses, and a range of storage times and DNase treatments (Table 17). For a cleaner analysis the ovotestis samples could have been split according to the three main extraction/population groups (8515-9014; 10627-10640; 11347-11357). Furthermore, an analysis of variance (ANOVA) test could be performed to indicate whether there is an effect of extraction group on the group means. It should be noted however that the most variable data group was the foot tissue, which represented ten individual samples prepared within the same week and received equivalent extraction protocols, therefore the reduction of variance may not be improved through separating the dataset.

Tissue Analysis

The only sample able to function as a calibrator sample across tissues was the OvoRef sample, which comprised a pool of multiple ovotestis samples. Therefore, all NRQ values in the tissue analysis are relative to the ovotestis tissue and can loosely be considered as relative to a *Dd* sample.

Incorporating the two fixed factors within the tissue analysis, genotype and tissue, a two-way factorial ANOVA was considered in order to establish whether tissue or genotype was having a significant effect on the gene expression of each GOI. However, the data failed to meet the requirements of the model, and so more robust, pairwise comparisons were performed between the group means. In the knowledge that genotype has a significant effect on the expression of *Ldia2* and the Larp2/3 targets, the pairwise comparisons were performed between the means of genotype specific groups (Table 43 - Table 48). By separating the data according to genotype and tissue, the sample sizes again became too small for appropriate use of parametric statistical tests. The NRQ data from GOIs that were not found to be significantly different between genotypes (10 out of the 14 GOIs in this experiment) could have been pooled across genotypes and perhaps enabled the use of the more powerful parametric tests. Yet, the majority of significant differences in gene expression between tissues were found to be highly significant (<0.005) and therefore the risk of false negatives due to using a less sensitive statistical test is not a strong concern. Two of the expression differences

identified in *Ldia2 3' UTR* and *Ldia2 ORF* were very close to the significance boundary (Table 47). However, these targets were found to be highly significantly different between genotype and therefore could not have been pooled for a parametric test.

Performing pair-wise comparisons between each genotype within a tissue and between each tissue within a genotype, a total of 140 pairwise comparisons were performed. Therefore, inferring significance from a p value of <0.05 would lead to significant differences in expression being identified in nearly seven of these comparisons simply by chance. Significance corrections have not been formally applied to the results of pairwise comparisons, however inferences regarding the strength of the relationship can be made from the p value itself. The majority of significant differences identified in the analyses were <0.01. Obtaining a p value of <0.01 would occur less than twice simply by chance within the 140 comparisons. Therefore, the statistical significance of these results are not believed to represent false positives. There were in fact seven occurrences of statistical significance only <0.05 and >0.01. Therefore, caution may be advised when inferring biological meaning from these differences, especially the tissue comparison of *Ldia2* 3' UTR expression between *Dd* ovotestis and embryo tissues bearing a p value of 0.045 (Table 47).

Conclusion

The qPCR experiments have succeeded in highlighting significant differential gene expression within and between embryo, ovotestis and foot tissues in chiral variants of *L. stagnalis*. The only GOIs found to be significantly DE between chiral genotypes were the primary candidate gene, *Ldia2* and arp2/3 complex genes, which are directly linked to the function of the primary candidate. Furthermore significant differences between chiral genotypes were substantially more pronounced in the embryo tissue than the ovotestis tissue, and absent in the somatic foot tissue.

The frameshift mutation in the sinistral copy of *Ldia2* appears to have resulted in a pronounced reduction in the quantity of the transcript, although due to the relative nature of the quantitative comparison it is difficult to infer direction of regulation. Interestingly the *Dd* samples showed an expression level of *Ldia2* almost exactly halfway between that of the homozygote groups in both the embryo and the ovotestis tissue, indicating equal regulation of the two gene copies.

The lack of DE observed in the remaining GOIs and especially the alternate diaphanous formin, *Ldia1*, strongly supports the frameshift mutation being responsible for the tight association of *Ldia2* with chirality determination in *L. stagnalis*. The patterns of gene expression identified in the arp2/3 transcripts however, require further analysis in order to propose reliable biological interpretations.

All GOIs, with the exception of *Ldia2*, were found to be greatly underrepresented in the single-cell embryo compared to the ovotestis and foot tissue. The comparative relationships of gene expression indicated that the sinistral copy of *Ldia2* is significantly under-represented in the embryo compared to the somatic tissue, whereas the functional *Ldia2* in the wild-type embryo is over-represented compared to the somatic tissues. Thus regulation is apparently occurring in both versions of the transcript.

Chapter 4: Expression-RAD: a reduced representation next-generation sequencing method for transcriptomic analysis

Introduction

When embarking on the search for the chirality determining gene (or closely linked group of genes) in *Lymnaea stagnalis*, the mechanism of chirality determination was largely unknown, although the genetic mode of inheritance was well described and observed (reviewed in Chapter 1, *Lymnaea* as a model system). Therefore, the Davison research group adopted a purely genetic approach to initially identifying candidate genes. *L. stagnalis* does not currently have a reference genome; however recent advances in high-throughput DNA sequencing facilitated genome-wide comparative analyses in *L. stagnalis*.

Next generation sequencing

Over the last 10 years there has been a surge of so called 'next-generation sequencing' (NGS) technologies, representing their advancement from traditional Sanger sequencing (Sanger, Air et al. 1977, Sanger, Nicklen et al. 1977). Many different methods have been employed to perform massively parallel sequencing of whole genomes, including pyrosequencing: Roche 454 (Margulies, Egholm et al. 2005)), sequencing by ligation: SOLiD (McKernan, Peckham et al. 2009), ion semiconductor sequencing: Ion Torrent, Life Technologies (Rothberg, Hinz et al. 2011), and the current leading market method, sequencing-by-synthesis: Illumina (Bentley, Balasubramanian et al. 2008). These technologies each have their own variable capabilities and limitations discussed in many previous reviews (Mardis 2011, Loman, Misra et al. 2012, Quail, Smith et al. 2012, Reuter, Spacek et al. 2015) and most recent capabilities quoted on their respective websites. Yet they each represent a platform for sequencing a massive number of DNA molecules simultaneously, from (almost) random start sites across the genome, within an automated instrument. They also share the major limitation of short reads. 454 sequencing produces longer reads of ~400 base pairs (bp) (although new systems to be released quote ~1 kilobases (kb)) compared to Illumina, SOLiD and Ion Torrent reads, all of which are generally less than 150 bp (Mardis 2011, Loman, Misra et al. 2012, Quail, Smith et al. 2012). This has previously been the highlighted benefit of 454 sequencing,

allowing for better genome assembly due to the longer sequence reads. However, each of these sequencers still requires the assembly of short reads to generate sufficient descriptive sequence data. Furthermore, each of these technologies requires an amplification step of the library to enable detection by the sequencer introducing a substantial source of sequencing error.

New techniques are emerging able to generate substantially longer sequencing reads from single molecules in real-time negating the process of library amplification (Eid, Fehr et al. 2009, Pushkarev, Neff et al. 2009, Mikheyev and Tin 2014). The currently available PacBio (Pacific Biosystems) sequencer can already produce average read lengths of 1,500 bp far exceeding that of its competitors, however demonstrates a much higher error rate (Quail, Smith et al. 2012). The MinION (Oxford Nanopore Technologies) sequencer promotes a 'handheld' sequencer able to sequence DNA molecules up to a giga-base in read length in real-time for a fraction of the running costs of other sequencers (Eisenstein 2012). Ultimately the maximum read-length can only meet the length/quality of the molecule input to the system. These 'third generation sequencing' methods are expected to revolutionise, yet again, the current applications of whole genome sequencing (WGS). The first publications of the MinION early access programme are emerging and reveal that the MinION can achieve read lengths comparable to the input DNA, yet the sequencing error rate needs improving (Mikheyev and Tin 2014, Quick, Quinlan et al. 2014, Ashton, Nair et al. 2015). Currently Illumina sequencing is still considered to attain the best balance of read lengths, error rates and costs (Loman, Misra et al. 2012, Mikheyev and Tin 2014). The term NGS is used throughout this thesis to refer to all high-throughput sequencing, although strictly most of the technologies discussed here are now largely considered to represent 'second-generation' sequencing.

The data output of the currently available technologies has dramatically increased since their introduction (Mardis 2011). Currently a single Illumina HiSeq 2000 run can generate up to 600 gigabases (Gb) of sequence data, the equivalent of >5 human genomes at 30 times coverage (Illumina 2011) (note that Illumina's latest sequencer HiSeq 4000 has increased this to 12 human genomes (Illumina 2015)). The capacity to gain such a wealth of sequence data *de novo* has enabled a new scale of genome-wide screening for genetic variation within organisms with or without a reference genome. The continuing fall in costs of NGS technologies coupled with the increasing sequencing coverage is bridging the gap between model and non-model organisms in genetic research.

Applications of NGS data

The advent of NGS technologies has generated a shift of genomic studies from controlled laboratory based model-organisms to ecologically well characterised species, enabling exploration of the genetic basis of ecologically and evolutionarily important questions. The high-throughput of NGS data facilitates the identification of hundreds of microsatellite loci and thousands of single-nucleotide polymorphisms (SNPs) in a fraction of a single run, permitting the inclusion of many individuals. Thus dramatically increasing the capabilities of large-scale comparisons of genetic variation and marker discovery in relation to phenotypic traits, such as quantitative trait loci (QTL) mapping and genome wide association studies (GWAS) in addition to population analyses evaluating inter-relatedness and genetic structure of groups of individuals (reviewed in Ekblom and Galindo 2011). The high throughput of NGS data can be spread further to perform metagenomics, able to rapidly analyse species diversity via pooled environmental samples through species specific 'barcodes' (eg. Fonseca, Carvalho et al. 2010). NGS data is also being applied to reveal epigenetic modifications characterised by methylation patterns (eg. Taylor, Kramer et al. 2007, Cokus, Feng et al. 2008) and the structure of DNA and chromatin packing (eg. Johnson, Tan et al. 2006, Barski, Cuddapah et al. 2007), which both play significant roles in the regulation of gene expression. The analytical possibilities of NGS data are immeasurable, such that deep sequencing projects represent a hypothesis-free experimental design, generating results available for a vast number of subsequent queries.

The accessibility of WGS through NGS has promoted re-sequencing of some already available reference genomes and the generation of new reference genomes assembled *de novo*. A growing number of species now have a reference genome available, and some have multiple reference genomes, most notably the human genome, which is preparing to describe 1000 genomes (www.1000genomes.org). However, even with the increasing availability of NGS sequence data, assembling the genome without a reference, *de novo*, is still a significant challenge beyond the capabilities of most molecular ecology research groups.

Reduced representation sequencing methods

For the majority of applications of NGS data, the whole genome is not required. Shorter sequence contigs can provide the sequence information necessary to identify genetic variation and subsequent linkage markers. SNPs represent the most abundant genetic markers within the genome, and subsequently facilitates rapid linkage-mapping across individuals and populations (Baird, Etter et al. 2008). However, single base changes in genetic sequence are common errors within the current NGS platforms (Mardis 2011). Therefore, it is essential to have sufficient sequence data to verify the

occurence of SNPs in the presence of occasional sequencing error. In organisms lacking a reference genome this will necessitate acquiring a good depth of sequencing at the same genetic loci in multiple individuals.

Although NGS provides a cost-effective method to obtain substantial amounts of sequence data, it still represents a significant financial investment for a research group. As such many experiments are limited to one or two lanes of sequencing. Therefore, prioritising sequencing effort is an important part of experimental design. Reducing the complexity of the genome will result in an increased depth of sequencing at a smaller number of loci. These 'reduced-representation' methods also increase the likelihood of sequencing the same loci in multiple samples. A number of approaches have been appropriated, most of which involve high-throughput sequencing of restriction enzyme fragmented DNA (Miller, Dunham et al. 2007, Baird, Etter et al. 2008, Huang, Feng et al. 2009, Andolfatto, Davison et al. 2011, Elshire, Glaubitz et al. 2011, Etter, Preston et al. 2011).

Restriction-site associated DNA Sequencing (RAD Seq)

The restriction-site associated DNA Sequencing (RADSeq) method as described by Baird *et al.* (2008), combines Illumina paired-end sequencing and the disruption of restriction endonuclease recognition sites to initiate sequencing only at genomic regions which flank a particular restriction enzyme recognition site. In this manner the same specific regions across the genomic are 'over-sequenced' in multiple individuals, providing sufficient depth of sequencing to permit the identification, verification and scoring of a large number of SNPs simultaneously within the experiment enabling genetic mapping (Baird, Etter et al. 2008), population genomics (Hohenlohe, Bassham et al. 2010) and phylogeography (Emerson, Merz et al. 2010) analyses, amongst others (Rowe, Renaut et al. 2011).

RADSeq is at present the most popular reduced-representation sequencing technique (Henri, Cariou et al. 2015) with 185 associated publications currently listed in the Web of Science since the first experimental papers were published in 2010 (search results recorded 11th November 2015). However, for clarity, an overview of the method is described here and summarised in Figure 44.

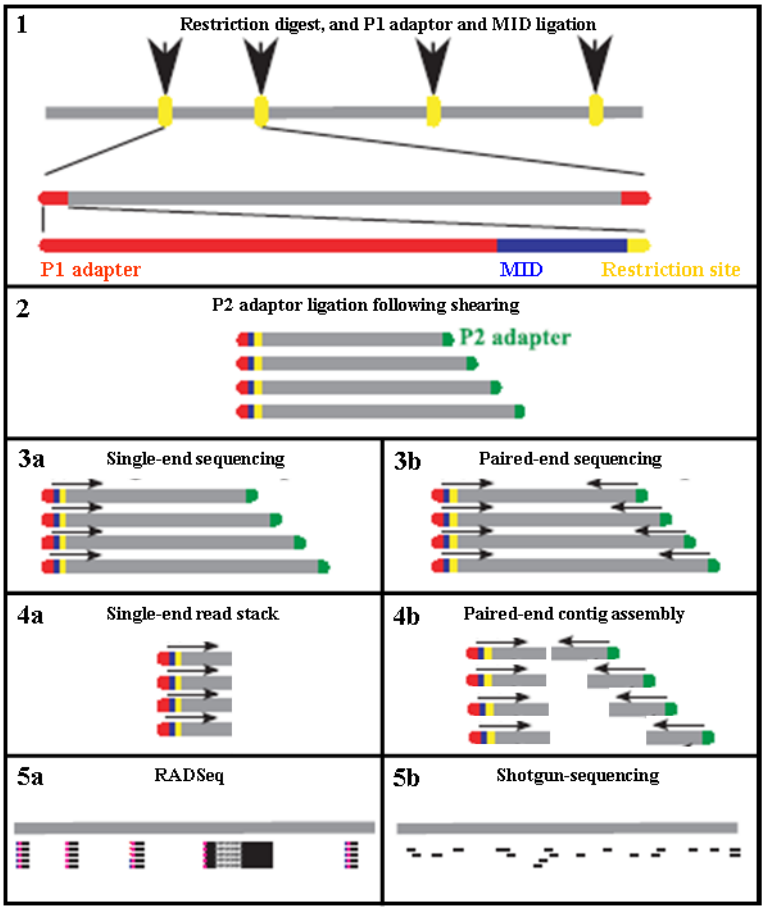


Figure 44. Overview of RADSeq method. 1. Restriction enzyme fragments DNA at the specific recognition sites (indicated in yellow). P1 Illumina sequencing adaptor (indicated in red) and molecular identifier (MID) (indicated in blue) are ligated to the cut site overhang. 2. Ligation of P2 Illumina sequencing adaptor (indicated in green) following random shearing of fragments. 3a. Arrows indicate sequencing direction originating from P1 adaptor in single-end sequencing. 3b. Arrows indicate sequencing direction originating from P1 and P2 adaptors in paired-end sequencing. 4a. Over-sequencing originating from the same P1 adaptor flanking a specific restriction site. 4b. Overlapping paired-end contigs assembled to the same P1 adaptor sequence. 5. Visual representation of the focused sequencing power of RADSeq (a) compared to shot-gun whole genome sequencing (b). Partially redrawn and adapted from Rowe *et al.* 2011.

Firstly, the individual sample DNA is digested with the restriction enzyme of choice, resulting in DNA fragments with a 'sticky-end' overhang. The fragments are then ligated to a P1 Illumina sequencing adaptor (both P1 and P2 adaptors are required for sequencing within the Illumina platform) and a unique molecular identifier (MID) sequence (which specifies the fragments to the sample), specifically tailored to complement the restrict fragment overhang. The DNA fragments now contain a sequencing adaptor and MID and so multiple samples can be pooled according to the requirements of the sequencing experiment. Provided that MIDs are not reused within the same sequencing lane, each sequenced read can be ascribed to the sample it was generated from. The pooled library is then sheared randomly to generate fragments within an average length of a few hundred base pairs (to fit the input requirements of the Illumina sequencing platform). P2 Illumina

adaptors are then ligated to the sheared fragments. The library of DNA fragments is then amplified via a PCR utilising primers specific to both of the Illumina sequencing adaptors. Due to a modification on the end of the P2 adaptor, only fragments containing both a P1 and P2 adaptor are amplified. This ensures that every sequence within the sequencing library contains both Illumina sequencing adaptors and the essential MID. Because the Illumina P1 sequencing adaptors are only associated with the restriction site overhangs, sequencing will only originate from these regions. Following paired-end sequencing, overlapping P2 sequence reads, which originate from the random sheared ends of the DNA fragments, can be assembled to form mini-contigs associated with the specific RAD tag P1 sequence.

The methodology of RADSeq is suitable for any organism containing restriction recognition sites and therefore an accessible method for genome-wide marker discovery in organisms lacking any prior sequence information. Furthermore, the methodology is highly versatile. The depth of sequencing can be prioritised over the overall span of the genome sequenced by using a restriction enzyme with a less frequently occurring recognition sequence, or vice versa. Additionally, double-digest RADSeq (*dd*-RADSeq) has since been developed to increase the number of sequenced loci (Peterson, Weber et al. 2012). Ultimately however, the capacity to accurately design the RADSeq genome coverage is dependent on prior knowledge of predicted restriction recognition sites within the genome to estimate frequency of sequenced loci.

RNA Seq

NGS technologies have since been applied to transcriptome analyses, both descriptive and quantitative. RNA sequencing (RNA Seq), the most common method of NGS transcriptome analysis, involves sequencing complementary DNA (cDNA) reverse-transcribed from RNA and subsequently fragmented into suitably-sized inserts for the specified NGS platform. Alternatively, the high-throughput power of NGS can be applied to tag based approaches such as expression sequence tags (ESTs) (Bouck and Vision 2007) and serial analysis of gene expression (SAGE) (Velculescu, Zhang et al. 1995, Harbers and Carninci 2005, Nielsen, Hogh et al. 2006). A less frequented transcriptomic sequencing approach is that of 'exome-capture' to preferentially sequence the coding regions from DNA, as a result this method cannot provide quantitative gene expression data (Choi, Scholl et al. 2009, Teer and Mullikin 2010).

The availability of RNA Seq has greatly increased the possibilities of gene expression analysis in both model and non-model organisms. One of the major breakthroughs of RNA Seq is the ability to perform experimental comparisons of differential gene expression without prior knowledge of the

target sequence, one of the principle limitations of tag-based approaches (Bouck and Vision 2007). Relative gene expression levels can subsequently be inferred from the sequencing depth of specific transcripts (Marioni, Mason et al. 2008). Thus RNA Seq offers incredibly high-throughput descriptive and quantitative transcriptome experiments, simultaneously able to identify sequence variation, such as SNPs and splice-variants, and differential gene expression between conditions. Furthermore, variant marker discovery within mRNA sequences are expected to be of greater interest when exploring adaptive differences due to their functional association with protein-coding genes. As a consequence, RNA Seq can be considered a form of reduced representation sequencing.

RNA Seq does not suffer the same high background noise and cross hybridisation as is common in microarray experiments, and shows a greater dynamic range of transcript detection, better able to detect rare transcripts and differences in gene expression with increased resolution (Ekblom and Galindo 2011). The ability to detect differences in expression of previously unknown transcripts across the whole range of the transcriptome allows for a more complete comparison of gene expression patterns between species, conditions or tissues being compared. Additionally, known candidate genes can be extracted from the RNA Seq data and examined after the experiment has taken place. However, the ability to do this will depend on whether the sequencing coverage of the transcriptome was sufficient to capture the candidate gene in question.

The sequencing depth will be divided across the number of samples included within the sequencing reaction. Therefore, RNA Seq can become an expensive option if intending to compare a large numbers of samples at a sufficient level of coverage to confidently describe transcripts in organisms lacking a reference genome. Although RNA Seq appears to have surpassed its predecessors as the method of choice for identification of transcriptomic variation (McGettigan 2013), further analyses using more 'in-house' techniques such as microarrays or quantitative real-time PCR (qPCR) provide valuable tools to increase power of predicted differential gene expression in an increased number of replicates or conditions.

RNA Seq generates a tremendous amount of sequence data, however there are still significant challenges in assembling the reads *de novo* (Haas and Zody 2010). In addition to the computational issues in assembling short sequence reads common to all NGS data, transcriptome data suffers unbalanced coverage across both the whole transcriptome due to quantitative gene expression patterns, and the individual transcripts due to sequencing biases such as Illumina's GC bias (Minoche, Dohm et al. 2011) and the 3' sequencing bias known to occur from mRNA selection and the cDNA fragmenting process (Wang, Gerstein et al. 2009). Additionally, functionally conserved

sequences between genes may create ambiguity in the assembly. The assembler must also be able to generate multiple transcripts per locus to accommodate splice variants. Subsequently errors may be incorporated due transcripts encoded by adjacent loci which may overlap and be misrepresented as chimeric loci. Alternatively, a highly expressed transcript containing a sequencing error may dominate the assembly over a correct but lowly expressed sequence (summarised in Grabherr, Haas et al. 2011).

A number of transcriptome assemblies have already been successfully generated completely *de novo* (eg. Vera, Wheat et al. 2008). The increase in sequencing depth and read length of NGS platforms now available, in addition to the growing capabilities of assembly programs, are continuously improving *de novo* transcriptome analyses (Grabherr, Haas et al. 2011), however for the reasons outlined above, RNA Seq without a reference genome remains a challenge.

Expression RADSeq (eRAD)

Similarly, to genomic analyses, informative gene expression analyses can be performed that do not require a complete assembly of the transcriptome. Therefore, it is of interest to explore the possibilities of reduced-representation sequencing methods within the transcriptomic analyses. The novel technique of expression RAD sequencing (eRAD) was presented by members of the Cresko laboratory (University of Oregon) at a conference in 2010 (National Center for Research Resources 2010). No analyses have yet been published regarding eRAD from this group.

A comparative analysis was proposed by the Davison research group in collaboration with the Genome Analysis Centre (TGAC) and contributions from Dr Susan Bassham of the Cresko Laboratory (University of Oregon) to assess whether the RADSeq method could be applied to double-stranded cDNA to perform reduced-representation transcriptome sequencing and assess the output in contrast to traditional RNA Seq in an organism lacking reference genome: *L. stagnalis*.

RADSeq was developed and is primarily employed to identify large numbers of small scale sequence variations, such as SNPs by reducing the complexity of the genome in order to focus sequencing efforts. eRAD would also provide this capability, yet reducing further the complexity of the genome by only examining protein coding regions. This may limit the potential to find sequence variation, due to functional restraints on genetic variation within protein coding sequences. However, the identified sequence variation is also more likely to be indicative of adaptive processes due to such functional constraints. Consequently, eRAD may facilitate rapid genetic linkage mapping in organisms with complex or polyploid genomes by further reducing the search area.

It was hoped that the eRAD method may provide a quantitative as well as descriptive, transcriptomic sweep, to assess gene expression across the transcriptome. The expected increased depth of sequencing of a smaller number of loci via the eRAD method was intended to improve the reliability of, sequence and count data. A number of features of RNA Seq data question its suitability of use for quantitative DE analysis (discussed in Bullard, Purdom et al. 2010, Robinson and Oshlack 2010, and Tarazona, Garcia-Alcalde et al. 2011 amongst others). Some of the key features are introduced here and how eRAD may influence such effects.

The count data from RNA Seq results from sub-sampling of the 'population' of total transcripts within individuals and inferring differences in counts as relating to the grouping factors. In light of reasonably high level of reproducibility of RNA Seq DE analysis (Marioni, Mason et al. 2008) this does not appear to be a large cause for concern, however the extent of the error will ultimately depend on the sequencing effort and variability of the transcriptome assessed. The focussed sequencing effort in eRAD may reduce unevenness and the erroneous effects of subsampling, due to the increased likelihood of sequencing the same loci across individuals.

More reliable comparisons of read counts between individuals are expected with higher sequencing depth, as it becomes more likely that the transcript has received sufficient sequencing effort across all individuals, consequently reducing variability. However, expression estimates can still be problematic when coverage is high (McIntyre, Lopiano et al. 2011, Tarazona, Garcia-Alcalde et al. 2011). Higher depth of sequencing also provides increased statistical power due to the greater sampling size (Oshlack and Wakefield 2009). It has been previously estimated that to gain accurate quantification of >95% of transcripts in a mammalian transcriptome would require ~700 million reads, although this is reduced to <10 million for 80% of reads (Blencowe, Ahmad et al. 2009). Therefore, the increased depth of sequencing provided for fewer transcripts by eRAD was predicted to improve reliability of quantification via read counts. Additionally, many RNA Seq DE analyses suffer from reduced statistical power by the inclusion of a minimal number of replicates. The increased sequencing depth per locus gained from eRAD would also allow for the inclusion of more individuals, thus improving statistical power.

One of the most common biases in short read RNA Seq count data is that longer transcripts have a much higher probability of being sequenced than shorter ones, simply due to the increased range available to initiate sequencing (Oshlack and Wakefield 2009). Increased sequencing depth results in increased statistical power due to reasons described above. Therefore, longer transcripts are more likely to be recognised as DE than shorter transcripts. It is unknown how the eRAD method will affect

this bias. Although all transcripts will initiate from the same restriction recognition site, there is a greater probability of a recognition site being present in a longer transcript. Therefore this bias will likely affect eRAD count data in a similar fashion.

There are however, capabilities of RNA Seq, which eRAD will not be able to fulfil such as strand specific analyses (Levin, Yassour et al. 2010). Furthermore, although unable to generate a fully assembled transcriptome due to the reduced representation method of sequencing, eRAD intended to improve the quality of *de novo* assembled contigs compared to RNA Seq through the increased depth and therefore reliability of the sequenced loci. As described earlier, many of the functional uses of sequence data do not require a fully assembled transcriptome/genome, yet sequence accuracy is of great importance.

NGS in *Lymnaea stagnalis*

The Davison research group has generated both genomic and transcriptomic resources for the inbred chirally dimorphic lines of *L. stagnalis*. As introduced in Chapter 1, successful RADSeq has been performed, leading to the identification of tightly linked chirality markers and ultimately the chirality locus (Liu, Davey et al. 2013). Additionally, transcriptomic data has been acquired from both one cell and 32 cell *L. stagnalis* embryos, via RNA Seq on the 454 sequencing platform, although quantitative analyses were not performed on this data (Liu, Davey et al. 2014). Accordingly, there is a considerable amount of both genomic and transcriptomic sequence data available for *L. stagnalis* within the Davison research group, though still in the process of assembly. Collaborative efforts are continuing on the *L. stagnalis* genome sequencing project based in France (Genoscope-CEA, de la recherche à l'industrie). As of this year 3/4 of the genome (946 Mb) has been captured at 20 times depth (project update, January 2015 *pers. comm.*).

As described previously, the chiral determinant is known to be a maternal transcript and so in order to assess causative transcripts as opposed to downstream gene expression associated with chiral variation the ovotestis was selected for sequencing. The eRAD data was hoped to provide improved *de novo* quantitative gene expression data across the transcriptome in addition to a descriptive transcriptomic resource for the ovotestis tissue. When the eRAD/RNA Seq project was commenced, the Davison research group was yet to identify the primary candidate, *Ldia2*. In light of the recent progress in the genomic data analysis, which has finely mapped the chirality locus and associated sequence variation, the transcriptome data analysis was prioritised for quantitative gene expression differences. The quantitative abilities of the transcriptomic data would allow variation to be characterised in genes that do not exhibit sequence variation and potentially reveal functional

processes associated with chiral dimorphism unobtainable from the genomic data analysis ongoing within the Davison research group.

The overall aim of this project however, was not only to generate sequence information contributing to the characterisation of the chirality determinant in *L. stagnalis* but also to develop a novel sequencing method to improve transcriptomic analyses in non-model organisms.

Methods

Samples

Each individual used in the eRAD experiment were offspring from a single, self-fertilised *Lymnaea stagnalis* (5995) bearing the sinistral phenotype with heterozygous (*Dd*) genotype. It is important to note that because *L. stagnalis* is hermaphroditic, genetic recombination still occurs when self-fertilising.

The entire ovotestis (ranging from 10 - 25 mg, data not shown) was dissected from a fresh individual adult snail and snap-frozen within a microcentrifuge tube using a dry ice/ethanol slurry to minimise RNA degradation whilst the remaining samples were extracted. Total RNA was extracted within an hour of freezing the fresh tissue, using TRI Reagent® solution (Applied Biosystems) according to the standard protocol and eluted into 100 µl RNase free water.

The genotypes of the snails were inferred firstly from the coiling direction of their offspring. The individuals producing sinistral offspring were scored as homozygote recessive, *dd*. Those producing dextral offspring however, required additional DNA extractions and PCRs using genetic linkage markers previously established for this population to specify chiral genotype. The specific PCR used was 'b3g FP1 F8R8'. The PCR reaction followed the protocol outlined in Box 2, and the primer sequences are described in the SI (S2). Twenty-two samples were selected for the eRAD sequencing experiment. These comprised four *DD*, ten *Dd* and eight *dd* samples.

Total RNA samples were enriched for mRNA using the Poly(A) Purist™ kit (Ambion/Applied Biosystems) according to the kit specified protocol. Due to the relatively low starting quantity of total RNA, ~10 µg (Table 49) (note that the Poly(A) kit allows for 2 - 400 µg), the final mRNA was reprecipitated in the minimum appropriate volume: 10 µl of RNA storage solution. The mRNA samples were quantified using spectrophotometer (Nanodrop 2000, ThermoFisher Scientific). Due to the low quantity of sample, extensive quality assessments of the RNA could not be performed, however the 260/230 and 260/280 ratios from the Nanodrop data provide an indication of sample quality (ThermoScientific 2010).

Double-stranded complimentary DNA (ds cDNA) was synthesised from approximately 500 ng of mRNA using the protocol provided by Susan Bassham, Oregon University (2011, *pers. comm.*) described in Box 4. The ds cDNA samples were then purified and concentrated using the MinElute® reaction clean-up kit (Qiagen) and eluted into 12 µl. The samples were then quantified via fluorometer (Qubit®, ds DNA broad-range BR assay, Invitrogen). The remaining mRNA not used in

cDNA synthesis was stored at -80°C and subsequently sent to TGAC to be prepared for RNA Sequencing (August 2012).

eRAD Library Preparation

The eRAD library preparation protocol was adapted from the original RAD sequencing method (Baird, Etter et al. 2008) in accordance with advice from Simon Baxter (University of Adelaide), Natalie Lowe (University of Edinburgh), Darren Heavens (TGAC) and Mengning (Maureen) Liu (previous Davison research group member). The libraries were prepared by both Mengning (Maureen) Liu and me. Two eRAD libraries were prepared independently. The 22 samples were split into two groups of 11 allocating 2 *DD*, 5 *Dd* & 4 *dd* individuals in each library. The two libraries are referred to here as Library 3 and Library 4. The generalised library preparation protocol is described in Box 5 - Box 8. The initial sample preparation and digest steps of the protocol (Box 5) were adapted to correct for the reduced amount starting material. Additionally, the amount of P1 adaptor was adjusted to the variable input quantity of samples within Library 4. The specific volumes are presented in Table 49.

Throughout the eRAD library preparations, a number of steps require the product to be purified and concentrated to a smaller volume. Throughout the protocols this is often abbreviated to 'purify and concentrate'. This was performed using the MinElute® kit, which employs DNA binding spin columns (Qiagen), using either the reaction clean-up kit, gel extraction kit or the PCR clean-up kit. In all cases the protocol was followed as specified by the supplier, although an additional waiting time of 1 minute was included prior to each centrifugation to allow for maximum binding to the column membrane. Where possible, the final elution was performed in two steps, each using half of the desired final volume, provided that the reduced elution volume did not contradict the minimum required volume for elution. For example, if the final elution volume was 20 µl, two elutions of 10 µl would have been performed. The final gel extraction of the entire prepared library was performed twice on Library 4, in attempt to further clean the high level of primer dimer (gel images of each size selection step are presented in the SI, S11.3).

Each library preparation was completed over three consecutive days. The two libraries were sent to the Genome Analysis Centre (TGAC) and sequenced by paired-end Illumina sequencing on two HiSeq lanes. Library 3 was sequenced on lane 6 and Library 4 was sequenced on lane 7, the sequencing data is hereafter referred to as 'L006' and 'L007' respectively. The library samples were spiked with 20% PhiX. This is a much higher concentration than the standard 1%, recommended due to expected highly repetitive nature of the dataset (Matthew Clark, TGAC, *pers. comm*. December 2012).

Double stranded cDNA synthesis protocol

Combine:

2 µl Random Primer Mix (NEB)

0.8 µl 10 mM dNTP mix

X µl (up to 0.5 µg) RNA

X µl (if necessary) RNase free water to bring total to 13 µl **Total: 13 µl**

Heat to 65°C for 5 minutes, then ice

Collect contents at bottom of tube by brief centrifugation.

Add:

4 µl 5x First-Strand Buffer (Invitrogen)

1 µl 0.1 mM DTT (Invitrogen)

1 µl RNase inhibitor (RNAseOUT™ 40 U/µl, Invitrogen)

1 µl Superscript III reverse transcriptase (200 U/µl, Invitrogen) **Total: 20 µl**

Mix by gentle aspiration

Incubate at 25°C for 10 min.

First strand synthesis: Incubate at 50°C for 50 minutes.

Inactivation: 85°C for 5 minutes. Chill on ice, collect contents to bottom by short spin.

Add (on ice):

106.6 µl water

15 µl 10x Second-Strand Synthesis Reaction Buffer (B6117S, NEB)

3 µl dNTP mix (10 mM)

1 µl *E. coli* ligase (10 U/µl)

4 µl *E. coli* DNA polymerase I (10 U/µl)

0.4 µl *E. coli* RNase H (5 U/µl) **Total: 150 µl**

Mix by aspiration on ice

Second strand synthesis: incubate at 16˚C for 2 hours (not allowing to warm above 16˚C)

[Reaction can be scaled up to accommodate more starting RNA]

Box 4 Generalised protocol for generating double-stranded complementary DNA (cDNA) from RNA.

eRAD library protocol. Part 1: Sample preparation

1. Sample preparation

Purify and concentrate double-stranded cDNA using the MinElute® reaction cleanup kit (Qiagen) and quantify using a fluorometer (Qubit, Invitrogen).

2. Restriction Digest

Digest 100 ng double-stranded cDNA (or maximum possible within 10 µl volume) with PstI.

Add [No mastermix]:

8.7 µl	ds cDNA
1 µl	NEB3 (10x)
0.1 µl	BSA (100x)
0.2 µl	PstI (20 U/ µl)
X µl	water (if necessary to bring total to 10 µl)

Total: 10 µl

Incubate: 37 °C for 2 hrs, 80 °C for 20 min, 0.1 °C / sec ramp down to 20 °C

3. Ligation of P1 adapter

Ligate the specific barcode including P1 adaptor to individual samples. Adjust the amount of adapter to account for reduced input quantity of cDNA.

	1 reaction	Master Mix (14)	Per sample
NEB2 (x10)	0.2 µl	2.8 µl	
rATP (100 mM)	0.2 µl	2.8 µl	
H_20	1.2 µl	16.8 µl	**1.6** µl
Mix by aspiration, then add:			
Library 3: specific barcode (100 nM)	0.15 µl		
Library 4: specific barcode (20 nM)	0.1-0.3 µl		
Mix by aspiration, 15 minute incubation at room temp, then add:			
T4 Ligase (2 MU/mL)	0.1 µl		
Total	**2 µl**		

Incubate: 22 °C for 1 hr, 65 °C for 20 min, 0.1 °C / sec ramp down to 20 °C

Box 5 Generalised protocol for the generation of an expression RAD sequencing library, part one of four.

eRAD library protocol. Part 2: Library shearing & size selection

Pool samples, now ligated to a unique sequence identifier, prior to shearing.

4. Shear fragments via sonication

Fragment approximately 125 µl of the library (capacity of Covaris® tube) using the Covaris® ultrasonicator.

Settings: Duty cycle: 5%; Intensity: 3; Cycles/Burst: 200; Mode: Freq sweeping; Duration: 30 sec.

These settings should result in a fragment size of approximately 400 bp

Purify & concentrate: Elution volume: 20 µl EB

5. Gel electrophoresis and size extraction

Load 20 µl of sheared product with 5 µl of loading dye (Qiagen) into a 0.5x TBE and 2% agarose gel with ethidium bromide for UV visualisation. Run at 100V for 90 minutes.

Excise sheared DNA sized between 300-700 bp from the gel. Extract DNA from the gel fragment using Qiagen Gel Extraction Kit (dissolve gel at room temperature). Elution volume: 20 µl EB

Box 6 Generalised protocol for the generation of an expression RAD sequencing library, part two of four.

Figure 45 Specific sequence and cut sites (indicated by red triangles) recognised by restriction digest enzymes SbfI and PstI. Resulting overhang ('sticky ends') are indicated by coloured text.

eRAD library protocol. Part 3: Ligate P2 adapter

6. Quick Blunt

'Polish' the ends of the DNA using the Quick Blunting Kit (NEB). Note: it is assumed 1 µl is lost through elution through column.

Add:

19 µl	DNA
2.5 µl	Buffer
2.5 µl	dNTP
1.0 µl	Enzyme

Total: 25 µl

Incubate: room temperature, 30 minutes.

Purify & concentrate: Elution volume: 24 µl EB

7. Add dATP

Add:

23 µl	DNA
3 µl	Buffer NEB2
1 µl	dATP (10 mM)
3 µl	Klenow exo- (15 Units)

Total: 30 µl

Incubate: 37 °C for 30 minutes. Allow reaction to cool slowly (within the inactivated heat block) to room temperature (approximately 15 minutes).

Purify & concentrate: Elution volume: 26 µl EB

8. Ligate P2 Adapter

Add:

25 µl	DNA
3 µl	Buffer NEB2
0.5 µl	rATP (100 mM)
1 µl	P2 Adapter (10 µM)
0.5 µl	T4 Ligase

Total: 30 µl

Incubate: room temperature, 30 minutes.

Purify & concentrate: Elution volume: 20 µl EB

Box 7 Generalised protocol for the generation of an expression RAD sequencing library, part three of four.

eRAD library protocol. Part 4: PCR amplification

9. Trial PCR

Trial PCR: 20 µl volume. Positive control: nRAD7. Negative control: water.

Add:

1 µl	DNA library
10 µl	Phusion® High fidelity PCR master mix (NEB)
1 µl	P1-PCR primer (10 µM)
1 µl	P2-PCR primer (10 µM)
7ul	H_2O

Total: 20 µl

Thermocycling parameters:

1. 98°C 30 secs
2. 98°C 10 secs
3. 65°C 30 secs
4. 72°C 30 secs
5. Cycle from step 2, 29 more times
6. 72°C 5 mins

10. Visualise PCR products on agarose gel:

Load the entire PCR product with 5 µl of 6x loading dye (Qiagen), into a 0.5x TBE and 2% agarose gel with ethidium bromide for UV visualisation. Run at 100V for 1 hour.

11. Bulk PCR

Following successful test PCR perform the bulk PCR, made in 120 µl volume and split into 6 x 20 µl reactions. Negative control: water

Add:

63 µl	H_2O
24 µl	Phusion® buffer (5x) (NEB)
4.8 µl	dNTP (5 µM)
6 µl	P1-PCR primer (10 µM)
6 µl	P2-PCR primer (10 µM)
15 µl	DNA library
1.2 µl	Phusion® polymerase

Total: 120 µl → 6 x 20 µl reactions

Thermocycling parameters: as above but reduced to 18 cycles.

Purify & concentrate: Elution volume: 20 µl EB

12. Visualise PCR products on agarose gel:

Load the entire PCR product with 5 µl of 6x loading dye (Qiagen), into a 0.5x TBE and 2% agarose gel with ethidium bromide for UV visualisation. Run at 100V for 1 hour.

Excise the PCR product sized between 300-700 bp from the gel. Extract DNA from the gel fragment using Qiagen Gel Extraction Kit (gel dissolved at room temperature). Elution volume: 20 µl EB

Box 8 Generalised protocol for the generation of an expression RAD sequencing library, part four of four.

Data Analysis

The raw RNA Seq data of the mRNA samples was received in October 2015. Therefore, due to time limitations, the RNA Seq data analysis was unable to be included within this thesis. The following methods and results are only in relation to the eRAD sequencing data.

Raw data preparation

The sequence data was received from TGAC as multiple fasta format sequence files labelled according to sequencing lane and either R1 (forward/primary sequence read) or R2 (paired-end sequence read). Quality control (QC) summary reports were generated using FastQC (Andrews 2010).

The multiple fasta files were then concatenated into 4 large fasta files according to library and read direction; L006 R1 reads, L006 R2 reads, L007 R1 reads and L007 R2 reads. The resulting collated raw sequence data fit the format required for use in the stacks analysis program (Catchen, Amores et al. 2011, Catchen, Hohenlohe et al. 2013).

Stacks analysis

The stacks program builds a catalogue of 'radtags' using only the R1 reads. The R2 reads are used to enable the clone filter program and to assemble the transcriptomic contigs, however do not directly contribute to the radtag count data.

Process Radtags

The program '*process_radtags*' identifies raw sequence reads containing the specified restriction cut-site sequence and unique barcode at the start of the sequence and converts them into 'radtags' per individual. Any read without the specific cut site or barcode identifier is removed from the dataset. R2 reads are matched to their R1 read using the information provided in the raw sequence fasta file. Both libraries were processed independently, using the same parameters, which were as follows. The function -c was enabled to remove any read with an uncalled base. The quality scores were specified as encoded using the Illumina 'Phred33' cut off and the function -q was enabled to discard reads with low q scores. The function -r was enabled to 'rescue barcodes' where possible, the minimum sequence difference between all barcodes was specified as 3 (calculation presented in the SI, S11.4). Finally, all reads were truncated to 70 bases using the -t function, specified at 70.

Once completed the program output a new fasta format file containing the radtags specific within each individual barcode. The 11 files generated were then renamed according to the individual snail ID they were derived from. Ultimately 22 individual files containing the cleaned radtags sequenced

within each sample were generated. A log data file was also produced providing summaries of read counts per library.

Clone filter

The initial data analysis employed the '*clone_filter*' program, which removes any R1 and R2 read pair which has an identical sequence to another R1 and R2 pair within an individual sequence file. The original R1 and R2 reads remain and only the duplicate pair is removed. These datasets are referred to with the prefix 'SUPER'. In order to assess the effects of using the clone filter, a second method of data analysis, which employed the same *process_radtags* parameters but did not utilise the clone filter step, was performed. These datasets are referred to with the prefix 'FULLFAT'.

The results from process radtags and the clone filter were used to calculate descriptive information for each library. Proportional representation of individuals and libraries were calculated as percentages within Microsoft Excel 2010.

Catalogue assembly

The '*denovo_map*' program was employed, which computes the three core Stacks programs, ustacks, cstacks and sstacks sequentially. All datasets were analysed as a mapping cross, using a 'superparent' generated by concatenating the processed individual RAD tag files of all samples used within the dataset. The catalogue loci were generated only from stacks created from the superparent. Stacks created from each of the individual samples were then mapped to the catalogue loci. This then generates a form of count data, providing tag counts corresponding to the sequencing depth of specific stacks at each locus within an individual.

The only stacks parameter that was altered from the default settings was the minimum sequencing depth required to create a stack in the parent (-m). A number of varied settings, including the default value of 3, were trialled to best suit the novel eRAD data and are presented in the SI (S14). The minimum sequencing depth to create a stack in the progeny (-P) was kept at the default value of 3. Six different stacks catalogues were chosen for further analysis, which are described briefly here and summarised in Table 52.

Firstly, three separate datasets were created from the sample files which employed the clone filter. One catalogue was generated using all 22 samples from both libraries, using a superparent created from collating all 22 samples as the 'parent sample'. This dataset is referred to as 'SUPER'. Another was generated using only the samples from L006 and using a superparent created by concatenating all 11 samples from that library. This dataset is referred to as 'SUPER6'. The third dataset was

generated using only the samples from L007 and using a superparent created by concatenating all 11 samples from that library. This dataset is referred to as 'SUPER7'.

By the same rationality, three datasets were created from the samples which did not employ the clone filter. These are referred to as 'FULLFAT', 'FULLFAT6' & 'FULLFAT7' and are also summarised in Table 52.

The datasets generated with a superparent created from 22 individuals (SUPER and FULLFAT) were analysed using an -m value of 15. The datasets generated with a superparent created from 11 individuals (SUPER6, SUPER7, FULLFAT6, FULLFAT7) were analysed with an -m value of 12.

Paired-end contig assembly

The paired-end reads were first allocated to individuals and stacks using the Stacks program 'sort_read_pairs.pl'. The paired sequence reads were then assembled into paired-end contigs (pe contigs) *de novo* within the program 'Velvet' (Zerbino and Birney 2008). Contig assemblies were generated for all three of the clone-filtered datasets (SUPER, SUPER6 and SUPER7). The superparent was not included in the contig assembly. The minimum contig length was specified to be -150 bases.

Differential expression analysis

Count data from the superparent and from any multi-allelic loci were excluded from the differential gene expression analysis.

The 'raw' count data exported from Stacks of each of the six catalogues (SUPER, SUPER6, SUPER7, FULLFAT, FULLFAT6 & FULLFAT7) were analysed in the program 'edgeR' (version 3.12.0, a freely available package within R) (Robinson and Smyth 2007, Robinson and Smyth 2008, Robinson, McCarthy et al. 2010, Robinson and Oshlack 2010). A range of parameters can be altered within the analysis. The analyses were performed here using the advice provided in the edgeR user's guide (Robinson, McCarthy et al. 2012) to identify differentially expressed (DE) RAD tags between the three genotypic groups: *DD*, *Dd*, *dd*. A number of variations of the datasets were assessed, which are described below and summarised in Table 55. The specific individuals included within each dataset are presented within the summary statistics in the SI (S15).

full dataset

Each sample sequenced within the eRAD library was included in the initial DE analysis. This resulted in two datasets of 22 individuals (SUPER & FULLFAT) and 4 datasets of 11 individuals (SUPER6, SUPER7, FULLFAT6 & FULLFAT7).

3Q

With the aim to reduce high variation within groups caused by predicted large numbers of zero counts, only those samples with the upper third quartile of tag counts >0 were included in the '3Q' analysis. This resulted in one dataset of 17 individuals (SUPER_3Q), one dataset of 16 individuals (FULLFAT_3Q) and four datasets of nine individuals (SUPER6_3Q, SUPER7_3Q, FULLFAT6_3Q & FULLFAT7_3Q). The SUPER7_3Q and FULLFAT7_3Q datasets necessitated the removal of one of the two *DD* samples. Thus statistical comparisons of group means were not included for this group due to the lack of available replicates.

Bd

With the aim to omit sample group number bias, balanced 'Bd' datasets were generated which included equal numbers of each genotype within the DE analysis. This resulted in two datasets of 12 individuals (SUPER_Bd & FULLFAT_Bd) and four datasets of six individuals (SUPER6_Bd, SUPER7_Bd, FULLFAT6_Bd & FULLFAT7_Bd). Where required, the first samples omitted were those with third quartile count data of 0. The remaining samples were removed without known bias, although in the case of SUPER_Bd and FULLFAT_Bd equal numbers were removed from each library.

Rm

To test the biological inferences of findings and the capabilities of the FDR corrections, a final dataset was created from the SUPER catalogue to test a random 'Rm' DE analysis. Three groups were created; A, B & C, each containing one *DD*, one *Dd* and one *dd* individual sample.

EdgeR analysis

EdgeR is based on a negative binomial model. It first automatically calculates library sizes from the total of counts within the sample. The tags were filtered to a minimum of 1 count per million (cpm) for the maximum number of samples representative of a genotypic group, therefore this filter ranged from 2-4 (Table 55). Each sample's count data was made relative, 'normalised', using the trimmed mean of M values (TMM) method via the inbuilt algorithm 'calcNormFactors' (Robinson and Oshlack 2010).

EdgeR calculated the dispersion using the quantile-adjusted maximum likelihood (qCML) method across the entire dataset (common dispersion) and for each tag within the dataset (tagwise dispersion). Subsequently the biological coefficient of variation (BCV) was calculated, which represents total variation between replicates. Mean variance plots were also generated to assess whether the dataset appropriately fits the model. Multi-dimensional scaling (MDS) plots were generated to explore clustering of overall gene expression patterns across samples.

Pairwise comparisons of normalised RAD tag counts were performed between each pair of genotypes: *dd* - *DD*; *Dd* - *DD* & *dd* - *Dd*. EdgeR identified DE of RAD tags between genotype groups using the 'exact test'. Based on the qCML methods the exact test incorporated the tagwise dispersion for each comparison. Statistical probability values were automatically adjusted using the Benjamini and Hochberg algorithm (Benjamini and Hochberg 1995) to account for multiple comparisons and control the false discovery rate (FDR).

The tags identified as differentially expressed within the 'SUPER', 'SUPER6' and 'SUPER7' datasets, were assessed for associated function by performing an NCBI Blastx online (http://blast.ncbi.nlm.nih.gov/Blast.cgi) on their associated paired-end contigs. If a Blast hit was not acquired for the pe contig or if the consensus sequence lacked a pe contig, the consensus sequence was Blasted firstly against the SUPER assembly and then the local *L. stagnalis* genomic assembly (version 10) to identify a related contig. The top contig was then assessed for function via Blastx online.

Gene Ontology

Blast2GO® was used to assess the gene ontology of the pe contigs (Conesa, Gotz et al. 2005). The functional annotation was performed by Dr. Teri Evans within the Blast2GO Pro software. Each of the three assemblies; SUPER, SUPER6 and SUPER7 were analysed separately. Firstly, each contig was characterised via Blastx, using an e value limit of 0.01, to a non-redundant protein database prepared by Dr Evans, including all vertebrate and invertebrate sequences downloaded from NCBI (20th June 2011) and subsequently run through Cd-hit (Li and Godzik 2006) to ensure it was non-redundant. The final protein database contained 407,788 sequences (the protein database was originally developed in Evans, Wade et al. 2014). The contigs were then annotated with likely associated Gene Ontology (GO) terms within Blast2GO®. Following the generation of the Blast2GO annotated assemblies, the remaining analyses were performed by me.

A number of the RAD tags generated multiple paired-end contigs. The annotations for each contig were pooled according to the RAD tag ID they were generated from. Therefore, each consensus RAD tag is represented only once within the functional annotation of the assembly. Descriptive summaries of the annotation of each of the transcriptomic assemblies were created, however further analyses within Blast2GO were limited by user licence access.

Results

Library preparation

Once enriched for mRNA, the RNA samples retained on average, 10% of their original concentration (data presented in the SI, S11.1). The ds cDNA samples, although all generated from approximately 500 ng of mRNA as advised, were of a generally low concentration, such that for the majority of individuals, the entire sample had to be used in the restriction digest reaction and did not reach the desired 100 ng per sample (Table 49). This was more pronounced in library 4 (L007), where the cDNA concentrations were more variable.

The eRAD libraries showed a reasonably high level of primer dimer, although this was removed through gel extraction. The second gel extraction of Library 4 showed no obvious primer dimer remaining in the library (gel images are presented in the SI, S11.3).

eRAD sequence data

Fast QC

The Fast QC output showed generally low quality sequence data was generated from the two eRAD libraries. L006 was considerably worse quality the L007. This is reflected in the per base quality scores along the sequence read shown in Figure 46.

Each library failed the QC check for 'overrepresented sequences'. These sequences were often classed by the program as the Illumina sequencing primers. Second to the Illumina sequence primers, the most common overexpressed sequence was identified as relating to soma ferritin or ferritin.

Table 49 Sample information for the 22 *L. stagnalis* ovotestis samples used in eRAD library 3 and 4. Information includes sample identifier (ID) and genotype (Geno), the mRNA sample quantity (ng/µl) with associated 260/280 and 260/230 ratios and the resulting total yield (µg) according to the sample volume (10 µl). The ds cDNA synthesis information includes the amount of mRNA used in the reaction (µl mRNA and ng mRNA), and the resulting cDNA concentration (ng/µl) and total yield (ng) in resulting the 12 µl sample volume. The library sample preparation shows in which library the sample was included, the amount of cDNA used per sample (µl & ng), the specific five nucleotide sequence identifier (Barcode) and the amount of P1 adaptor (µl P1) added to each sample. Note: P1 was used at 100nM concentration in Library 3, and 20nM in Library 4.

Sample Info		mRNA, sample volume 10 µl				ds cDNA synthesis, sample volume 12 µl				Library sample preparation				
ID	Geno	ng/µl	260/280	260/230	Total yield (µg)	µl mRNA	ng mRNA	ng/µl	Total yield (ng)	Library	µl cDNA	ng cDNA	Barcode	µl P1
9014	*DD*	77.00	1.66	1.84	0.77	6.5	499.7	8.99	107.88	3	10.0	89.9	ATGCT	0.15
8515	*DD*	94.04	1.71	1.89	0.94	5.3	500.3	11.75	141.00	3	8.5	100.0	CCAAC	0.15
8869	*Dd*	142.16	1.84	2.23	1.42	3.5	500.4	23.85	286.20	3	4.2	99.9	AGCTG	0.15
9013	*Dd*	111.74	1.69	2.03	1.12	4.5	500.6	9.62	115.38	3	10.0	96.2	CATGA	0.15
8562	*Dd*	87.88	1.71	1.85	0.88	5.7	500.0	10.80	129.60	3	10.0	108.0	GCCGG	0.15
8559	*Dd*	103.57	1.70	1.92	1.04	4.8	500.2	8.85	106.20	3	10.0	88.5	GAGAT	0.15
8544	*Dd*	150.08	1.68	1.96	1.50	3.3	499.8	9.67	115.98	3	10.0	96.7	TGCAA	0.15
8862	*dd*	96.23	1.69	1.98	0.96	5.2	499.4	9.32	111.78	3	10.0	93.2	ACGTA	0.15
8808	*dd*	116.94	1.69	1.98	1.17	4.3	500.5	10.40	124.80	3	10.0	104.0	CGTAT	0.15
9007	*dd*	83.43	1.66	1.93	0.83	6.0	500.6	8.83	105.90	3	10.0	88.3	GTACA	0.15
9009	*dd*	91.31	1.69	1.62	0.91	5.5	500.4	14.50	174.00	3	6.9	99.8	TAATG	0.15
8582	*DD*	103.35	1.59	1.43	1.03	4.8	496.1	2.47	29.64	4	10.0	24.7	TACGT	0.1
8502	*DD*	100.02	1.60	1.51	1.00	5.0	500.1	7.74	92.88	4	10.0	77.4	GTTGT	0.2
9001	*Dd*	115.58	1.64	1.63	1.16	4.3	497.0	9.30	111.60	4	10.0	93.0	TGACC	0.3
8522	*Dd*	125.53	1.63	1.47	1.26	4.0	502.1	10.70	128.40	4	10.0	107.0	GGTTC	0.3
8500	*Dd*	118.02	1.67	1.73	1.18	4.2	495.7	10.30	123.60	4	10.0	103.0	CAGTC	0.3
8530	*Dd*	115.69	1.63	1.42	1.16	4.3	497.5	7.09	85.08	4	10.0	70.9	CCTTG	0.2
8560	*Dd*	109.62	1.59	1.50	1.10	4.6	504.3	7.28	87.36	4	10.0	72.8	ATTAG	0.2
8531	*dd*	88.99	1.67	1.66	0.89	5.6	498.3	9.10	109.20	4	10.0	91.0	TCTCT	0.3
9000	*dd*	107.10	1.59	1.44	1.07	4.7	503.4	4.67	56.04	4	10.0	46.7	ATCGA	0.15
8867	*dd*	86.72	1.59	1.83	0.87	5.7	494.3	3.83	45.96	4	10.0	38.3	CTTCC	0.15
8587	*dd*	55.25	1.74	1.91	0.55	9.0	497.3	5.06	60.72	4	10.0	50.6	AGAGT	0.15

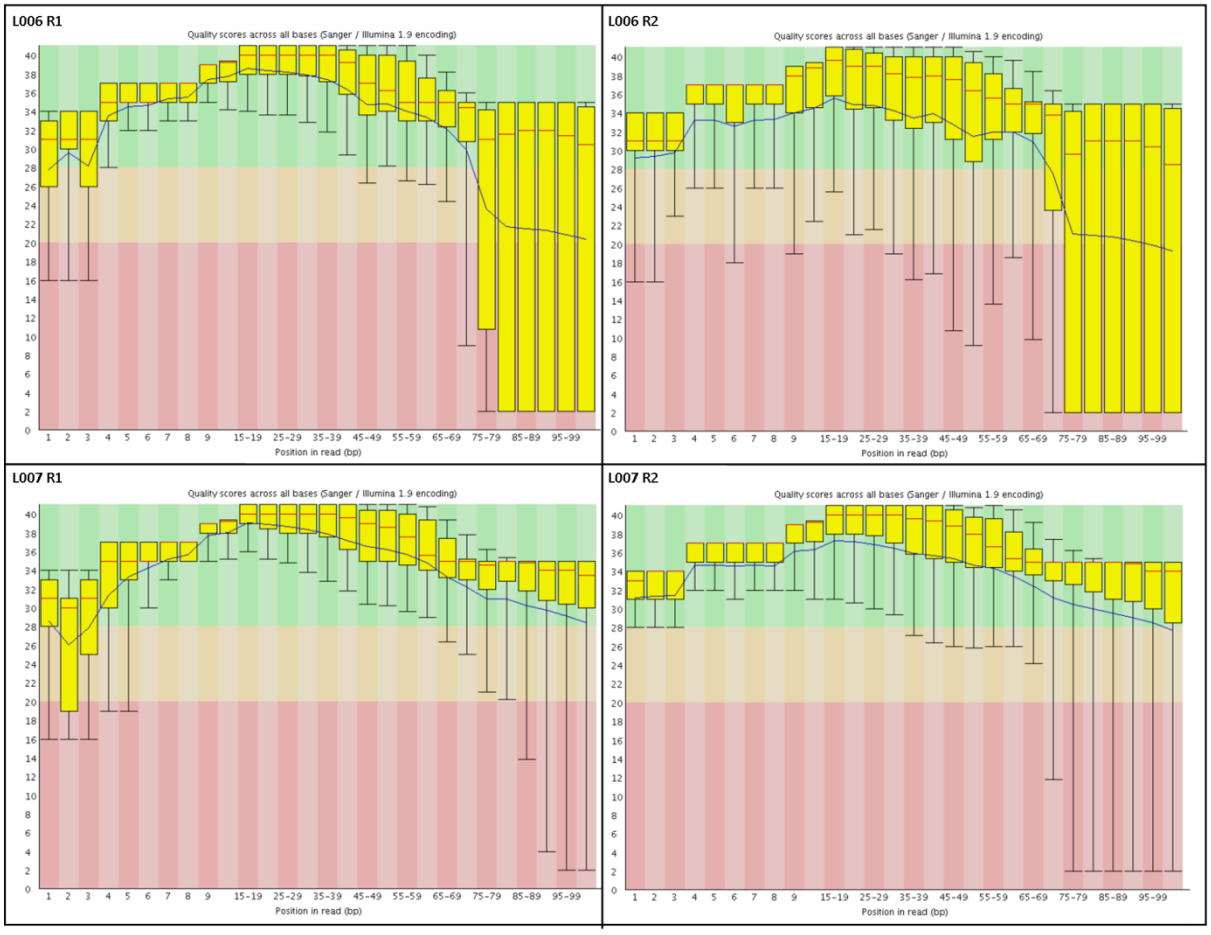


Figure 46 Quality scores per base along the length of Illumina sequencing reads (1-101) for L006 and L007 primary sequence reads (R1) and paired-end reads (R2). Q40 represents the best currently available quality score. The green, top zone indicates good quality reads, the bottom dark pink zone indicates poor quality reads. The error bars indicate the variability of sequence quality at this base position across all reads.

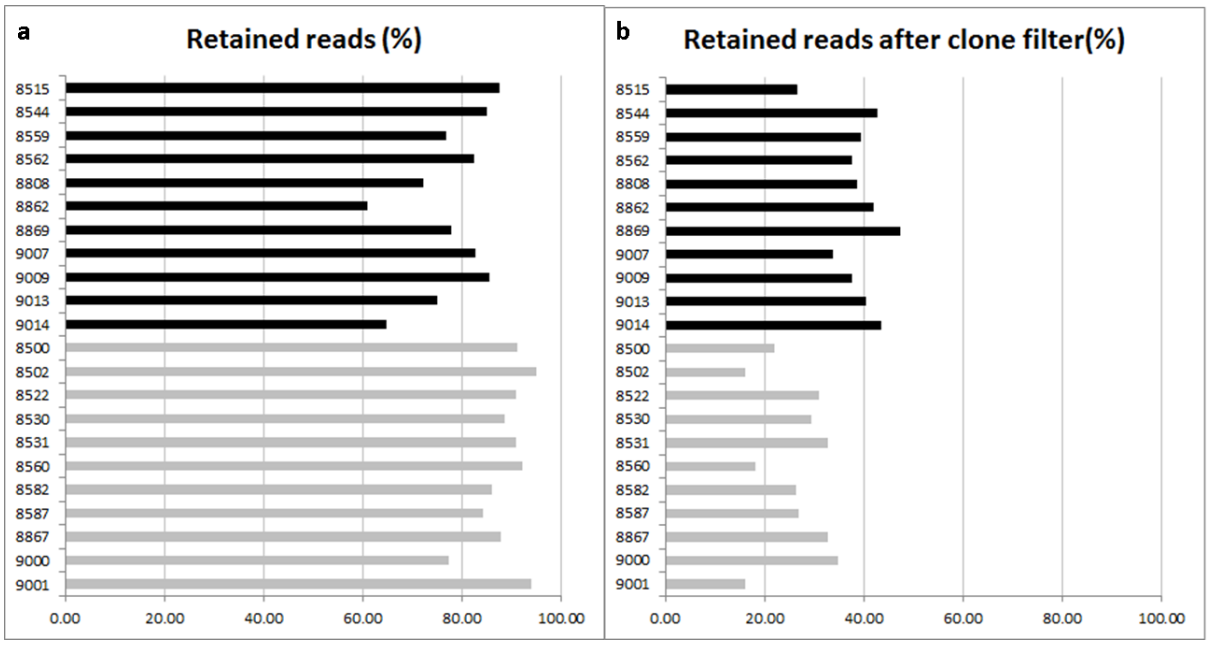


Figure 47 Bar chart shows percentage of reads retained of the original total reads per individual. L006 is shown in black. L007 is shown in grey.

Process radtags & clone filter

The summarised input and output data from the process radtags program is presented in Table 50. Each library produced a similar amount of retained reads. L006 and L007 represented 48% and 52% of the total combined dataset respectively. Yet the individual sample sequencing coverage within libraries was highly variable, with a number of individuals representing less than 1% of the dataset and notable one sample, *Dd* 9001, represented nearly half of the total sequences of L007 (Table 50, Figure 48).

The proportion of retained sequence reads prior to the clone filter reflected the overall quality of the library as indicated by the FastQC report. The samples in L007 retained more reads than those in L006. Averages are presented in Table 50 of the total retained reads across libraries. Both libraries, on average, retained over 77% of total reads, although substantial variation was observed between individuals, ranging from 60.99% - 95.01% retained (Figure 47).

The input and output of the clone filter program is summarised in Table 51. Firstly, it shows that of the retained reads, on average, over 75% had a paired-end (pe) read. Again variation between individuals was high, ranging from 46.10% - 95.69%. L007 generally showed higher levels of pe reads in individuals, and also higher levels of PCR clones.

The clone-filtered data resulted in greater differences between the sequencing effort of each library (Table 51). The clone filter identified, on average, over 65% of read pairs to be PCR clones. This was more pronounced in L007, such that following the clone filter, L007 represented only 32% of the combined dataset and retained only a quarter of its original sequencing reads (Table 51, Figure 47). However, the proportional representation of individuals within each dataset following the clone filter was somewhat more balanced, with dominance of high proportion individuals reduced (Figure 48, Table 51).

The sample sequencing bias which resulted in the highly varied representation of samples within the datasets was not found to be correlated with percentage GC composition of the five nucleotide sequence barcode, nor the cDNA starting concentration, quantity or quality, nor the amount of P1 sequence adaptor added (correlations are presented in the SI, S13).

Table 50 Summarised output from the process_radtags data preparation. Sample descriptions include sample ID (ID) and genotype (Geno) and library. Total reads input for each sample are presented (Total reads) with those retained after process_radtags (Ret. Reads). The amount of reads retained from the total input for each individual are presented as a percentage (% Ret./total). Finally the proportional representation of each individual within the library (% lib) and across the combined dataset (% dt) is presented as a percentage of the total reads. Totals and averages (Av) are presented for each library and across the whole dataset.

Sample description			process_radtags summary				
Library	**ID**	**Geno**	**Total reads**	**Ret. reads**	**% Ret./total**	**% lib**	**% dt**
L006	9014	*DD*	41,527,216	26,837,862	64.63	11.79	5.69
L006	8515	*DD*	55,050,824	48,120,113	87.41	21.15	10.20
L006	8869	*Dd*	11,326,668	8,824,349	77.91	3.88	1.87
L006	9013	*Dd*	16,606,626	12,434,572	74.88	5.46	2.64
L006	8559	*Dd*	16,173,978	12,402,370	76.68	5.45	2.63
L006	8562	*Dd*	10,247,336	8,453,365	82.49	3.71	1.79
L006	8544	*Dd*	3,047,000	2,588,442	84.95	1.14	0.55
L006	8862	*dd*	73,417,832	44,775,198	60.99	19.68	9.49
L006	8808	*dd*	4,797,174	3,460,814	72.14	1.52	0.73
L006	9007	*dd*	37,313,460	30,796,771	82.54	13.53	6.53
L006	9009	*dd*	33,755,044	28,878,053	85.55	12.69	6.12
Total per library	**11**	**2 *DD*, 5 *Dd*, 4 *dd***	**303,263,158**	**227,571,909**	**Av: 77.29**	**100.00**	**48.23**
L007	8502	*DD*	30,564,502	29,037,979	95.01	11.89	6.15
L007	8582	*DD*	3,210,450	2,759,419	85.95	1.13	0.58
L007	8560	*Dd*	24,488,372	22,579,501	92.20	9.24	4.79
L007	8500	*Dd*	34,298,456	31,215,375	91.01	12.78	6.62
L007	8530	*Dd*	5,573,926	4,938,501	88.60	2.02	1.05
L007	8522	*Dd*	11,695,918	10,625,576	90.85	4.35	2.25
L007	9001	*Dd*	111,770,998	105,076,155	94.01	43.01	22.27
L007	8587	*dd*	29,755,458	25,049,071	84.18	10.25	5.31
L007	9000	*dd*	2,166,416	1,672,050	77.18	0.68	0.35
L007	8867	*dd*	5,752,952	5,042,771	87.66	2.06	1.07
L007	8531	*dd*	6,922,482	6,283,512	90.77	2.57	1.33
Total per library	**11**	**2 *DD*, 5 *Dd*, 4 *dd***	**266,199,930**	**244,279,910**	**Av: 82.82**	**100.00**	**51.77**
Total in dataset	**22**	**4 *DD*, 10 *Dd*, 8 *dd***	**569,463,088**	**471,851,819**	**Av: 83.07**	**-**	**100.00**

Table 51 Summarised output from the clone filter program. Sample descriptions include sample ID (ID) and genotype (Geno) and library. The proportion of retained reads per individual including a paired end sequence are presented as a percentage (% pe/Ret.) in addition to the number of pairs of reads input and output. Also shown are; the proportion of input pairs identified as clones presented as a percentage (% clones/pe); the total number of reads retained per individual after the clone filter (Ret. After cf); the proportional representation of each individual within the library (% lib after cf) and across the combined dataset (% dt after cf) presented as a percentage of the original total reads prior to process_radtags; totals and averages (Av) for each library and across the whole dataset.

Sample description			clone filter summary							
Library	**ID**	**Geno**	**% pe/Ret.**	**input (pairs)**	**output (pairs)**	**% Clones/pe**	**Ret. after cf**	**% Lib. after cf**	**% dt after cf**	**% Ret after cf/total reads**
L006	9014	*DD*	54.01	7,247,847	2,852,694	60.64	18,047,556	15.82	10.74	43.46
L006	8515	*DD*	88.29	21,241,905	4,449,997	79.05	14,536,297	12.74	8.65	26.41
L006	8869	*Dd*	76.81	3,389,104	1,650,905	51.29	5,347,951	4.69	3.18	47.22
L006	9013	*Dd*	72.15	4,485,636	1,622,757	63.82	6,708,814	5.88	3.99	40.40
L006	8559	*Dd*	75.06	4,654,751	1,636,189	64.85	6,365,246	5.58	3.79	39.35
L006	8562	*Dd*	83.23	3,518,067	1,219,602	65.33	3,856,435	3.38	2.29	37.63
L006	8544	*Dd*	85.81	1,110,594	468,252	57.84	1,303,758	1.14	0.78	42.79
L006	8862	*dd*	46.10	10,320,373	3,325,375	67.78	30,785,202	26.99	18.32	41.93
L006	8808	*dd*	77.56	1,342,118	535,407	60.11	1,847,392	1.62	1.10	38.51
L006	9007	*dd*	82.93	12,770,129	3,650,671	71.41	12,557,855	11.01	7.47	33.66
L006	9009	*dd*	86.55	12,496,571	4,420,583	64.63	12,726,077	11.16	7.57	37.70
Total per library	**11**	**2 *DD*, 5 *Dd*, 4 *dd***	**Av: 75.32**	**82,577,095**	**25,832,432**	**Av: 64.25**	**114,082,583**	**100.00**	**67.88**	**Av: 39.00**
L007	8502	*DD*	95.69	13,893,495	1,823,011	86.88	4,897,011	9.07	2.91	16.02
L007	8582	*DD*	85.69	1,182,275	225,825	80.90	846,519	1.57	0.50	26.37
L007	8560	*Dd*	92.84	10,481,626	1,413,584	86.51	4,443,417	8.23	2.64	18.15
L007	8500	*Dd*	92.62	14,456,511	2,622,990	81.86	7,548,333	13.98	4.49	22.01
L007	8530	*Dd*	89.72	2,215,354	563,807	74.55	1,635,407	3.03	0.97	29.34
L007	8522	*Dd*	92.61	4,920,152	1,418,285	71.17	3,621,842	6.71	2.15	30.97
L007	9001	*Dd*	95.02	49,923,502	6,446,509	87.09	18,122,169	33.57	10.78	16.21
L007	8587	*dd*	83.80	10,495,084	1,948,894	81.43	7,956,691	14.74	4.73	26.74
L007	9000	*dd*	74.09	619,415	160,432	74.10	754,084	1.40	0.45	34.81
L007	8867	*dd*	88.73	2,237,114	659,425	70.52	1,887,393	3.50	1.12	32.81
L007	8531	*dd*	91.87	2,886,324	880,379	69.50	2,271,622	4.21	1.35	32.82
Total per library	**11**	**2 *DD*, 5 *Dd*, 4 *dd***	**Av: 89.33**	**113,310,852**	**18,163,141**	**Av: 78.59**	**53,984,488**	**100.00**	**32.12**	**Av: 26.02**
Grand Total	**22**	**4 *DD*, 10 *Dd*, 8 *dd***	**Av: 82.33**	**195,887,947**	**43,995,573**	**Av: 71.42**	**168,067,071**	**~**	**100.00**	**Av: 32.23**

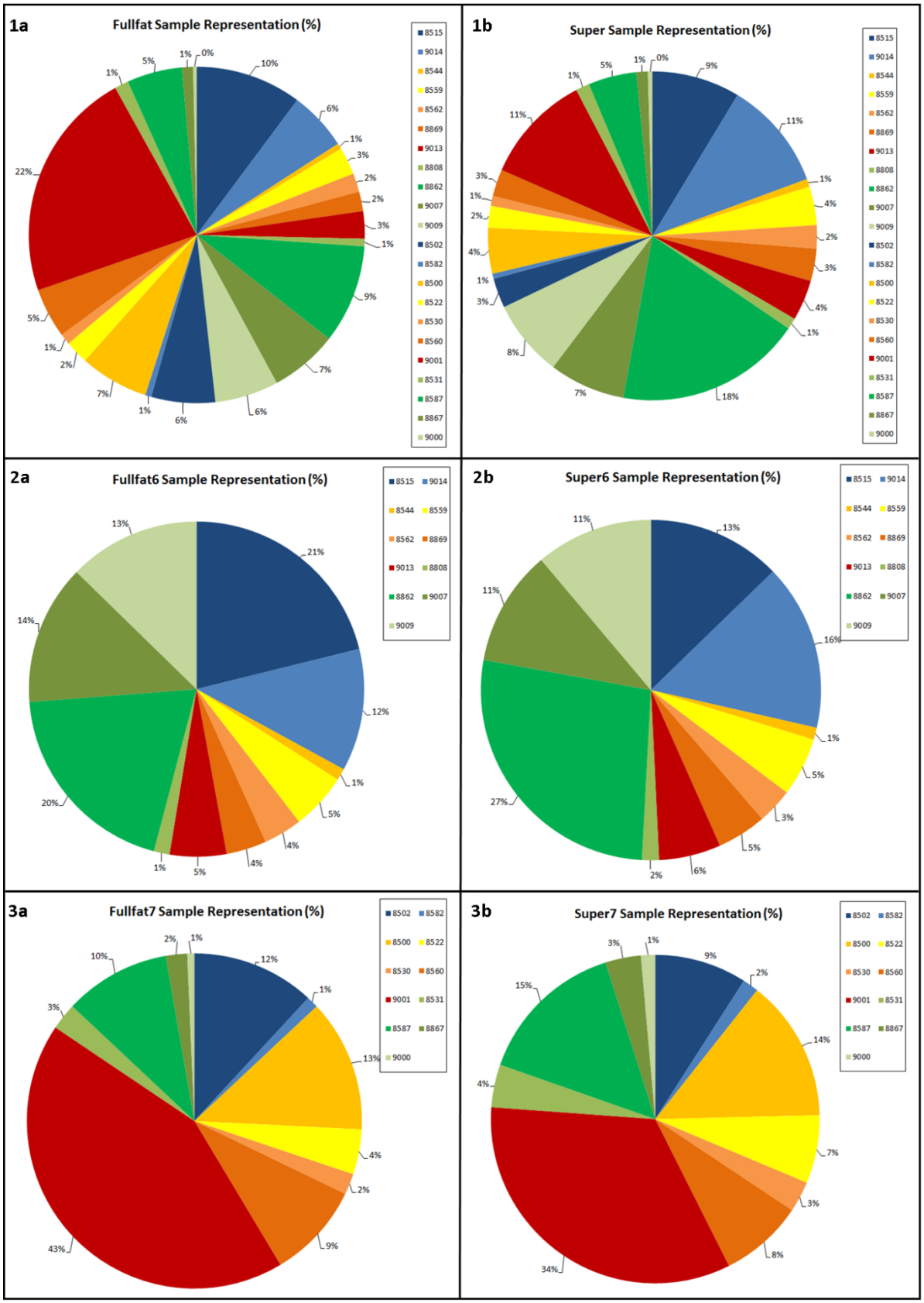


Figure 48 Pie charts show proportional representation of each individual sample's retained reads within each library and dataset, presented as a percentage of the total reads within each dataset. FULLFAT (1a) includes all samples prior to the clone filter. FULLFAT6 (2a) and FULLFAT7 (3a) include each sample from L006 and L007 respectively, prior to the clone filter. SUPER (1b) includes all samples after the clone filter. SUPER6 (2b) and SUPER7 (3b) include each sample from L006 and L007 respectively, after the clone filter. Colour identifiers are reused in chart 1a and 1b. Samples run clockwise from top.

Stacks analysis

A summary of the six Stacks catalogues; SUPER, SUPER6; SUPER7; FULLFAT; FULLFAT6 & FULLFAT7 is presented in Table 52. A similar summary of the catalogues generated through the alternative parameters trialled is presented in the SI (S14).

Each of the non-clone-filtered catalogues produced more unique stacks/loci than their clone-filtered equivalent (Table 52). The extent of this difference varied between individuals. Because the catalogue is generated only from the superparent, the stacks generated within each individual/progeny sample did not vary as the minimum stack depth for progeny was consistently '3'. Therefore, the total unique stacks generated within each individual is presented only once for the clone-filtered (SUPER) and non-clone-filtered (FULLFAT) data (Table 53, Table 54). Some individuals such as 9001 and 8587 showed almost twice as many unique stacks in the FULLFAT analysis compared to the SUPER analysis, whereas others such as 8808 and 8869 showed only a couple thousand tags different (Table 53, Table 54).

There was a discrepancy between the total number of stacks in the SUPER catalogue (56,899) and the number present in the parent (56, 8967). The 'missing' two stacks also appeared in the SUPER6 catalogue, yet not in any others (Table 52). The sum of the number of unique loci identified in the library specific datasets was approximately 55% higher than that of the combined dataset (for both clone-filtered and non-clone-filtered data, Table 52). It is therefore assumed that L006 and L007 have a reasonable level of overlap of transcriptome coverage.

Table 52 Summary of the six Stacks eRAD catalogues. The total number of individuals included per catalogue (n) is shown in addition to the number of each genotype (n, *DD*; n, *Dd* & n, *dd*). The minimum sequencing depth to create a stack is quoted per catalogue (m). The total number of loci/stacks created in the catalogue is shown (Unique stacks), with the number single-nucleotide polymorphisms (SNPs) found. Also shown are the number of 'blacklisted' stacks and subsequent final number of stacks within the catalogue (Total). The radtags corresponding to a catalogue locus within the Superparent (Parent) and increasing numbers of progeny (prog1 – prog20). Finally, the numbers of unique radtags, which contain a minimum of one SNP in one to five progeny are listed (prog1-5).

	Catalogue	**SUPER**	**SUPER6**	**SUPER7**	**FULLFAT**	**FULLFAT6**	**FULLFAT7**
Dataset Description	**Library**	L006 & L007	L006	L007	L006 & L007	L006	L007
	Clone filter	Yes	Yes	Yes	No	No	No
	Total n	22	11	11	22	11	11
	n, *DD*	4	2	2	4	2	2
	n, *Dd*	10	5	5	10	5	5
	n, *dd*	8	4	4	8	4	4
	m	15	12	12	15	12	12
Catalogue summary	**Unique stacks**	59,259	51,557	40,275	82,475	60,930	67,109
	SNPs found	4,148,130	3,608,990	2,819,250	5,773,250	4,265,100	4,697,630
	Blacklisted	2,360	1,649	1,945	2,765	1,906	2,341
	Total	56,899	49,908	38,330	79,710	59,024	64,768
Unique radtags	**Parent**	56,897	49,906	38,330	79,710	59,024	64,766
	prog 1	56,393	49,485	38,205	79,368	58,632	64,544
	prog 2	53,241	45,958	35,547	73,088	54,636	53,018
	prog 3	46,557	38,651	29,192	62,320	45,143	39,966
	prog 4	39,987	32,227	23,809	51,592	36,637	30,691
	prog 5	34,988	27,458	19,922	43,758	30,602	24,555
	prog 6	31,341	23,519	16,839	38,014	25,665	20,137
	prog 7	28,291	19,598	14,039	33,598	21,178	16,552
	prog 8	25,638	16,102	11,388	29,969	17,162	13,279
	prog 9	23,395	12,623	8,614	27,105	13,390	9,951
	prog 10	21,362	8,744	4,351	24,601	9,263	5,588
	prog 20	4,915	-	-	5,573	-	-
Unique radtags, SNP = 1+	**prog 1**	3,172	2,731	1,577	4,968	3,526	3,247
	prog 2	2,819	2,335	1,381	4,449	3,019	2,611
	prog 3	2,298	1,789	1,047	3,563	2,292	1,762
	prog 4	1,761	1,284	768	2,703	1,632	1,156
	prog 5	1,389	963	547	2,046	1,192	772

Table 53 Total unique stacks and single-nucleotide polymorphisms (SNPs) found in each of the clone-filtered samples, used in catalogue SUPER; SUPER6 & SUPER7.

SUPER		
ID	Unique Stacks	SNPs Found
8500	32,044	2,243,080
8560	23,089	1,616,230
9001	54,079	3,785,530
8522	23,390	1,637,300
8530	14,705	1,029,350
8582	7,298	510,860
8502	25,154	1,760,780
8531	19,002	1,330,140
9000	6,324	442,680
8867	17,129	1,199,030
8587	36,721	2,570,470
9013	24,085	1,685,950
8559	28,674	2,007,180
8869	28,188	1,973,160
8562	21,616	1,513,120
8544	12,538	877,660
9014	44,362	3,105,340
8515	40,417	2,829,190
8862	41,704	2,919,280
8808	13,022	911,540
9007	39,071	2,734,970
9009	46,996	3,289,720

Table 54 Total unique stacks and single-nucleotide polymorphisms (SNPs) found in each of the non-clone-filtered samples, used in catalogue FULLFAT; FULLFAT6 & FULLFAT7.

FULLFAT		
ID	Unique Stacks	SNPs Found
8500	50,483	3,533,810
8560	37,221	2,605,470
9001	97,395	6,817,650
8522	27,168	1,901,760
8530	18,491	1,294,370
8582	11,586	811,020
8502	39,171	2,741,970
8531	21,738	1,521,660
9000	10,170	711,900
8867	22,719	1,590,330
8587	70,412	4,928,840
9013	28,673	2,007,110
8559	36,206	2,534,420
8869	30,511	2,135,770
8562	25,421	1,779,470
8544	15,020	1,051,400
9014	60,497	4,234,790
8515	50,516	3,536,120
8862	52,177	3,652,390
8808	14,893	1,042,510
9007	47,787	3,345,090
9009	59,788	4,185,160

Differential expression analysis

A summary of the multiple edgeR analyses of the eRAD tag count data is presented in Table 55. Summaries of the descriptive statistics of the count data included in each of the analyses are presented in the SI (S15). Overall variation was high and count data sets included 27,000 - 57,000 tags for pairwise comparisons. An overview of the data distribution of the clone-filtered (SUPER) and non-clone-filtered (FULLFAT) datasets is presented in Figure 49. Both count datasets exhibited similar patterns of data distribution, yet the SUPER dataset presented lower average levels of variation compared to FULLFAT.

No clear clustering was seen in the multi-dimensional scaling (MDS) plots of the expression patterns of individuals (Figure 49.1), although there was a potential bias of the library specific samples to one side of the plot. L006 samples appear generally on the left hand side of the SUPER dataset and L007

samples appear more on the right side. This pattern was reversed in the FULLFAT data (Figure 49.1). Furthermore, sample 9001 represented a potential outlier in both datasets. The biological coefficient of variation (BCV) plots revealed very high levels of variation in the majority of tags (Figure 49.2). This was more pronounced in those tags with lower average expression levels. The Mean-Variance plots clearly show that the variance of the data does not fit the Poisson distribution of variance (Figure 49.3). The tagwise dispersion was used within all exact tests to compare group means. Due to the similarities of the data distribution of the clone-filtered and non-clone-filtered data, only the clone-filtered results are presented in detail here.

The MDS plots of SUPER6 and SUPER7 also did not show any apparent clustering, although sample 9009 and 9001 represented potential outliers in SUPER6 and SUPER7 respectively (Figure 50.1). The BCV plots revealed a similar pattern of tagwise variation within SUPER6 and SUPER7 compared to the combined SUPER dataset, yet the overall BCV was reduced in SUPER7 (Figure 49.b, Figure 50.b). The Mean-Variance plots again failed to fit the Poisson distribution, although were generally closer to the fit-line than those of SUPER (Figure 49.c, Figure 50.c).

Very few tags were identified as significantly differentially expressed (DE) between genotypes following correction for false discovery rate (FDR). Furthermore, two tags were identified as significantly DE between randomised groups (containing an equal number of individuals of each genotype) (Table 55). Due to the lack of any substantial differences in the data variance or DE results in the reduced datasets, only the full SUPER, SUPER6 and SUPER7 datasets are described in detail here. Further details of the reduced datasets can be viewed in the SI (S15, S16).

The SUPER dataset identified only one tag as significantly DE between genotypes, catalogue tag 38405. This tag was found to be significantly underrepresented in the *DD* samples compared to both *Dd* and *dd* samples (Table 55, Figure 51). The top NCBI Blastx hit of the associated pe contig was to an uncharacterised protein, the second hit, however of significantly reduced quality was to a von Willebrand factor-like protein (Table 57).

The SUPER6 dataset identified four tags as significantly DE between genotypes, catalogue tags 14295, 49725, 34376 and 49600. Each of these tags was found to be significantly underrepresented in the *DD* samples compared to both *Dd* and *dd* samples (Table 55, Table 56, Figure 52). The top NCBI Blastx hits of the associated pe contigs were as follows. Tag 14295 initially blasted to an uncharacterised protein in *Biomphalaria glabrata*, another freshwater snail. An alternative Blast hit with a good E value identified a involucrin-like protein in the sea slug, *Aplysia californica* (Table 57). Tag 49725 top hit described a leucine-rich repeat serine/threonine-protein kinase in *B. glabrata*

(Table 57). Tag 34376 had two separate contigs within the SUPER6 assembly. Subsequently both were included in the Blastx. The top hit of the shorter contig was an elongation factor 2 - like protein described in *B. glabrata* (Table 57). The longer contig did not generate a significant Blast hit. Finally, tag 49600 identified a hypothetical protein described in the bacteria *Escherichia coli* (Table 57).

The SUPER7 dataset identified two tags as significantly DE only between genotypes *DD* and *dd* catalogue tags 11332 and 14621. Contrary to the findings in the other datasets, each of these tags were found to be significantly underrepresented in the *dd* samples compared to the *DD* (Table 55, Table 56, Figure 52). Tag 11332 did have an associated pe contig within the SUPER7 assembly however no significant Blast hit was generated. The following top Blastx hits of the genomic contig associated with SUPER7 pe contig 11332, were a repetitive proline-rich cell wall-like protein in *B. glabrata* and a number of keratin-related proteins (only top hit shown, from the bird *Opisthocomus hoazin*) (Table 57). Tag 14621 did not have an associated pe contig. The consensus sequence was found to be related to a pe contig in the SUPER assembly and in the *L. stagnalis* genomic assembly. However, both contigs were relatively short and failed to acquire a significant Blastx hit (Table 57).

All bar one of the top hits within the NCBI Blastx analysis of the DE contigs was to a sequence described in another gastropod (Table 57).

De novo transcriptome assembly and Gene Ontology

The SUPER6 assembly produced a total of 30,438 contigs, whereas the SUPER7 assembly produced 25624 contigs, although the average length of contigs were slightly longer in SUPER7 (310, compared to 306). The highest number of contigs at the greatest length however came from the combined SUPER assembly producing 35,696 contigs with an average length of 313. The majority of contigs in the assemblies did not get a significant Blast hit (Figure 53, Figure 54, Figure 55). Approximately 30% of the contigs in each assembly were able to be annotated. The average length of contigs with annotation in the SUPER, SUPER6 and SUPER7 assembly were 352, 341 and 351 respectively (Figure 53.a, Figure 54.a, Figure 55.a). A quantitative summary is presented in the SI (S17).

The top tags species allocations presented in Figure 53.d, Figure 54.d and Figure 55.d, all include '*Lymnaea* stagnalis' twice. Listed firstly as '*Lymnaea* stagnalis' and second as '*Lymnaea stagnalis*,'. Combining the counts for both of the listings puts *L. stagnalis* in 11th position, after *Gallus gallus* in the SUPER assembly (Figure 53.d) and 12th position, after *Salmo salar* in both SUPER6 and SUPER7 (Figure 54.d, Figure 55.d).

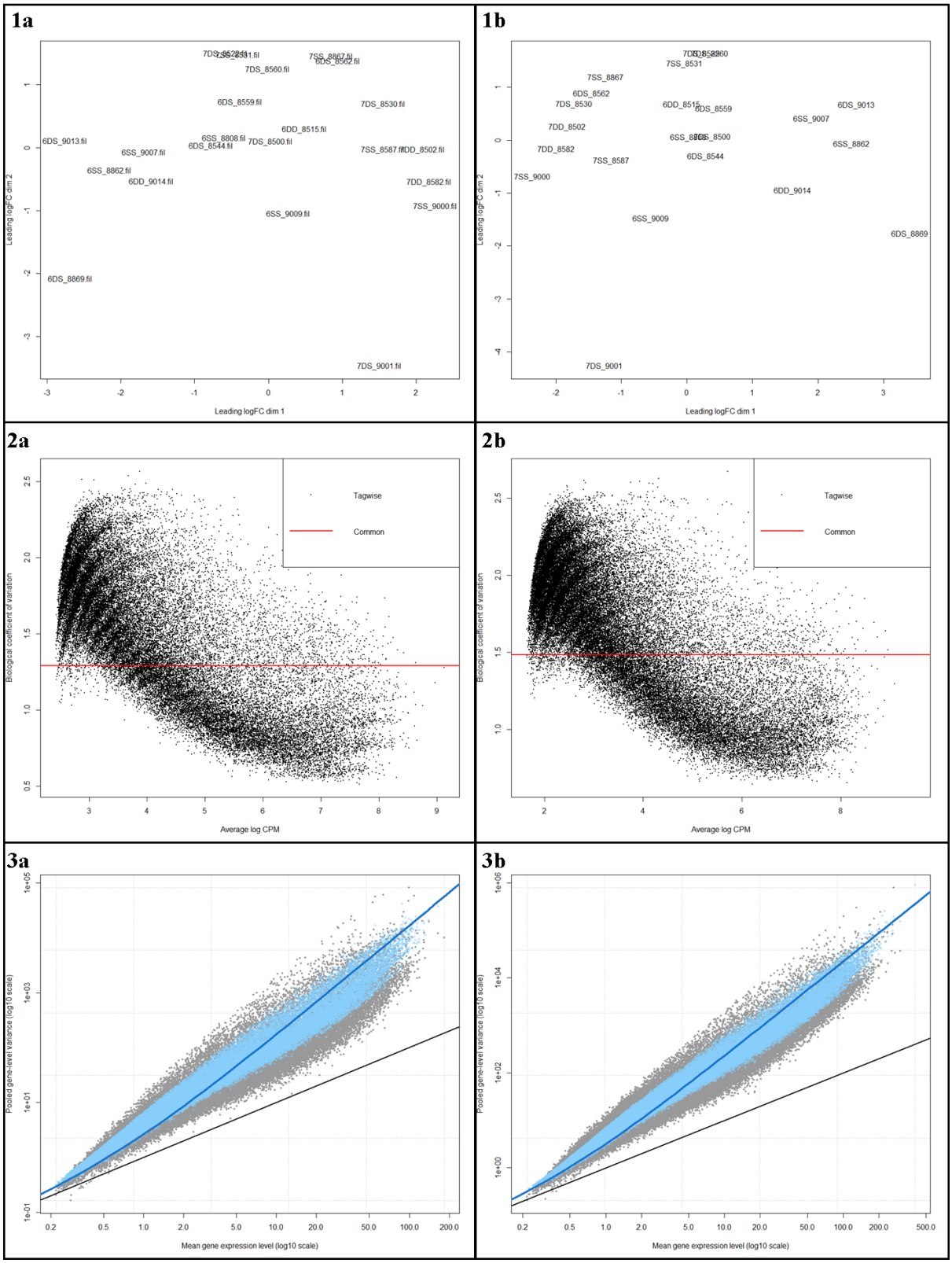


Figure 49 Data distribution of SUPER (a) and FULLFAT (b) count datasets visualised in edgeR. MDS plots (1) indicate the relatedness of overall expression pattern of individual samples within the dataset. Samples are labelled with a prefix of their sequencing library (6 or 7) and their genotype represented here as (*DD* = *DD*, DS = *Dd* or SS = *dd*). BCV plots (2) show the dispersion of each tag (represented by a black dot at the average log transformed level of expression in counts per million (cpm)) compared to the common dispersion of the whole dataset (red line). Mean-Variance plots (3) show the log transformed variance per 'gene' (eRAD tag), including raw variance of counts (grey dots) and variance using the tagwise dispersion (blue dots), the common dispersion (solid blue line) and *Poisson* variance (solid black line), all plotted against log transformed average gene expression level.

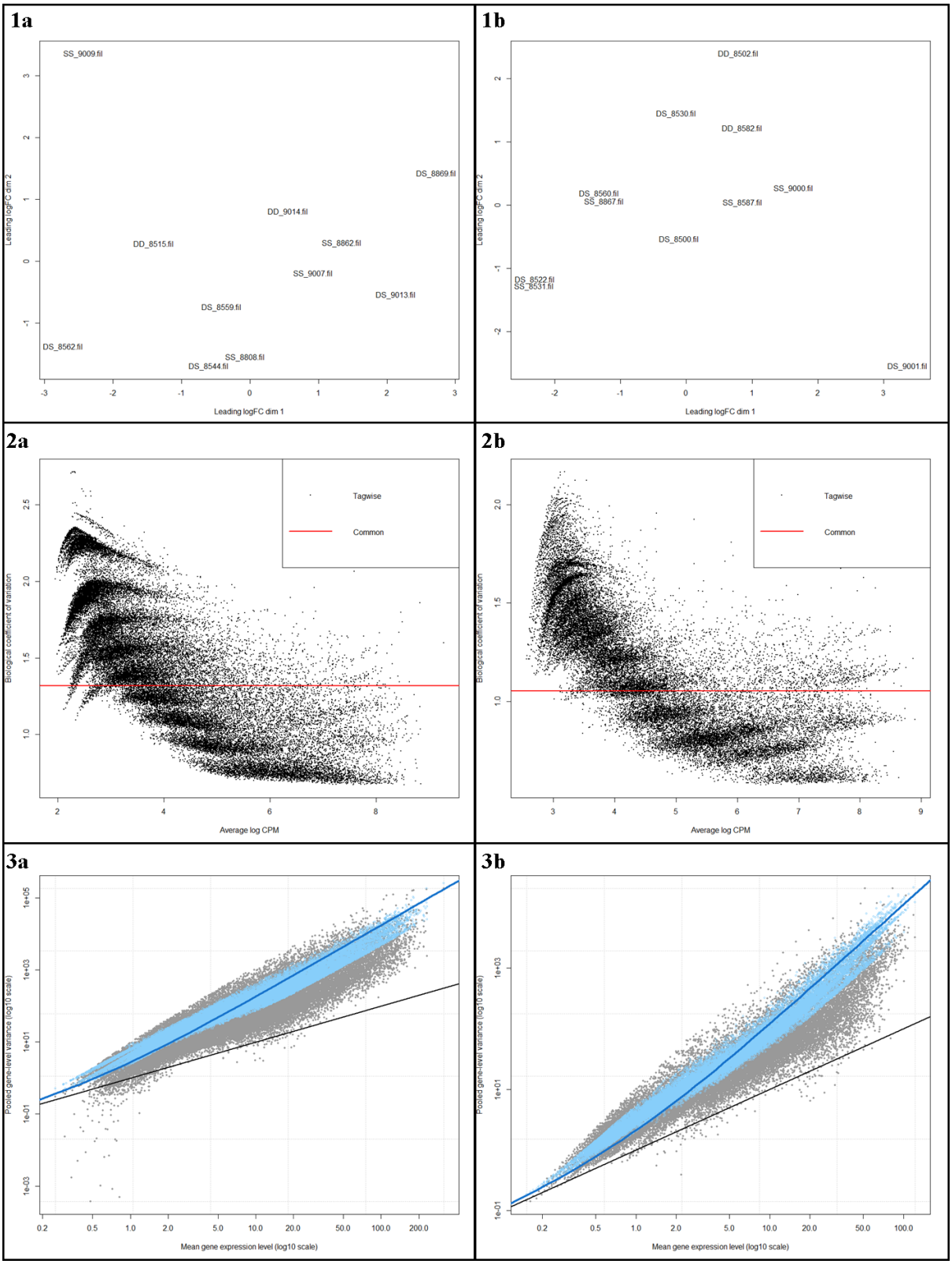


Figure 50 Data distribution of SUPER6 (a) and SUPER7 (b) count datasets visualised in edgeR. MDS plots (1) indicate the relatedness of overall expression pattern of individual samples within the dataset. Samples are labelled with a prefix of their genotype represented here as (*DD* = *DD*, DS = *Dd* or SS = *dd*). BCV plots (2) show the dispersion of each tag (represented by a black dot at the average log transformed level of expression in counts per million (cpm)) compared to the common dispersion of the whole dataset (red line). Mean-Variance plots (3) show the log transformed variance per 'gene' (eRAD tag), including raw variance of counts (grey dots) and variance using the tagwise dispersion (blue dots), the common dispersion (solid blue line) and *Poisson* variance (solid black line), all plotted against log transformed average gene expression level.

Table 55 Summary of multiple analyses within edgeR, A description of the individuals included within the dataset (Description) is presented in addition to the total number of individuals included (n) and the number specific to each genotype (DDn, Ddn, ddn). The total number of RAD tags within each count dataset (Tags), the number of tags with a sum total of zero (Zero Tags), the number of retained tags following filtering (Ret. Tags) and the proportion of tags filtered (% filt.) is shown, plus the number specified in the cpm filter (Filt.). The common dispersion (Disp) and biological coefficient of variation (BCV) is presented. Finally the number of tags identified as significantly up (+1) or down (-1) regulated according to the genotype group comparison specified and adjusted for false discovery rate (FDR) are summarised.

EdgeR analysis		n				Tags	Zero Tags	Filt.	Ret. Tags	% filt.	Variance		deTags (FDR <0.05)					
													dd-DD		*Dd-DD*		*dd-Dd*	
Dataset	**Description**	**Total n**	**DDn**	**Ddn**	**ddn**						**Disp**	**BCV**	**+1**	**-1**	**+1**	**-1**	**+1**	**-1**
SUPER	All samples in catalogue included. Clone-filtered	22	4	10	8	53,589	0	4	38,226	28.67	1.663	1.290	1	0	1	0	0	0
SUPER6		11	2	5	4	46,995	0	2	43,623	7.18	1.736	1.317	4	0	4	0	0	0
SUPER7		11	2	5	4	36,706	0	2	34,166	6.92	1.111	1.054	0	2	0	0	0	0
FULLFAT	All samples in catalogue included. Non-clone-filtered	22	4	10	8	74,496	0	4	48,332	35.12	2.206	1.485	0	0	0	0	1	0
FULLFAT6		11	2	5	4	55,222	0	2	51,617	6.53	1.942	1.394	3	0	2	0	0	0
FULLFAT7		11	2	5	4	61,372	0	2	49,894	18.70	1.690	1.300	0	0	0	1	1	0
SUPER_3Q	3rd quartile of count data > 0. Clone-filtered	17	3	8	6	53,589	417	3	43,985	17.92	1.752	1.324	1	0	1	0	0	0
SUPER6_3Q		9	2	4	3	46,995	321	2	43,387	7.68	1.624	1.274	4	4	4	0	0	6
SUPER7_3Q		9	1	5	3	36,706	90	3	28,029	23.64	0.948	0.974	n/a	n/a	n/a	n/a	1	0
FULLFAT_3Q	3rd quartile of count data > 0. Non-clone-filtered	16	3	8	5	74,496	376	3	56,872	23.66	2.187	1.479	0	1	0	0	0	1
FULLFAT6_3Q		9	2	4	3	55,222	227	2	51,282	7.13	1.947	1.395	3	1	1	0	0	0
FULLFAT7_3Q		8	1	4	3	61,372	281	3	36,791	40.05	1.355	1.164	n/a	n/a	n/a	n/a	1	0
SUPER_Bd	Equal n per genotype. Clone-filtered	12	4	4	4	53,589	2,115	4	30,882	42.37	1.307	1.143	5	0	4	0	0	0
SUPER6_Bd		6	2	2	2	46,995	1,142	2	39,753	15.41	1.420	1.192	2	0	0	0	0	0
SUPER7_Bd		6	2	2	2	36,706	2,639	2	27,026	26.37	1.103	1.050	0	0	0	0	0	0
FULLFAT_Bd	Equal n per genotype. Non-clone-filtered	12	4	4	4	74,496	4,078	4	38,344	48.53	1.756	1.325	0	0	1	0	0	0
FULLFAT6_Bd		6	2	2	2	55,222	1,160	2	47,106	14.70	1.768	1.330	0	0	0	0	0	0
FULLFAT7_Bd		6	2	2	2	61,372	7,085	2	37,809	38.39	1.507	1.227	0	0	0	0	0	0
Dataset	**Description**	**Total n**	**A**	**B**	**C**	**Tags**	**Tags = 0**	**Filt.**	**Ret. Tags**	**% filt.**	**Disp**	**BCV**	**AvsB**	**AvsB**	**AvsC**	**AvsC**	**BvsC**	**BvsC**
SUPER_Rm	Randomised groups. Clone-filtered	9	1 *DD*, 1 *Dd*, 1 *dd* in each group			53,589	3,306	3	33,827	36.88	1.258	1.122	0	0	2	0	1	0

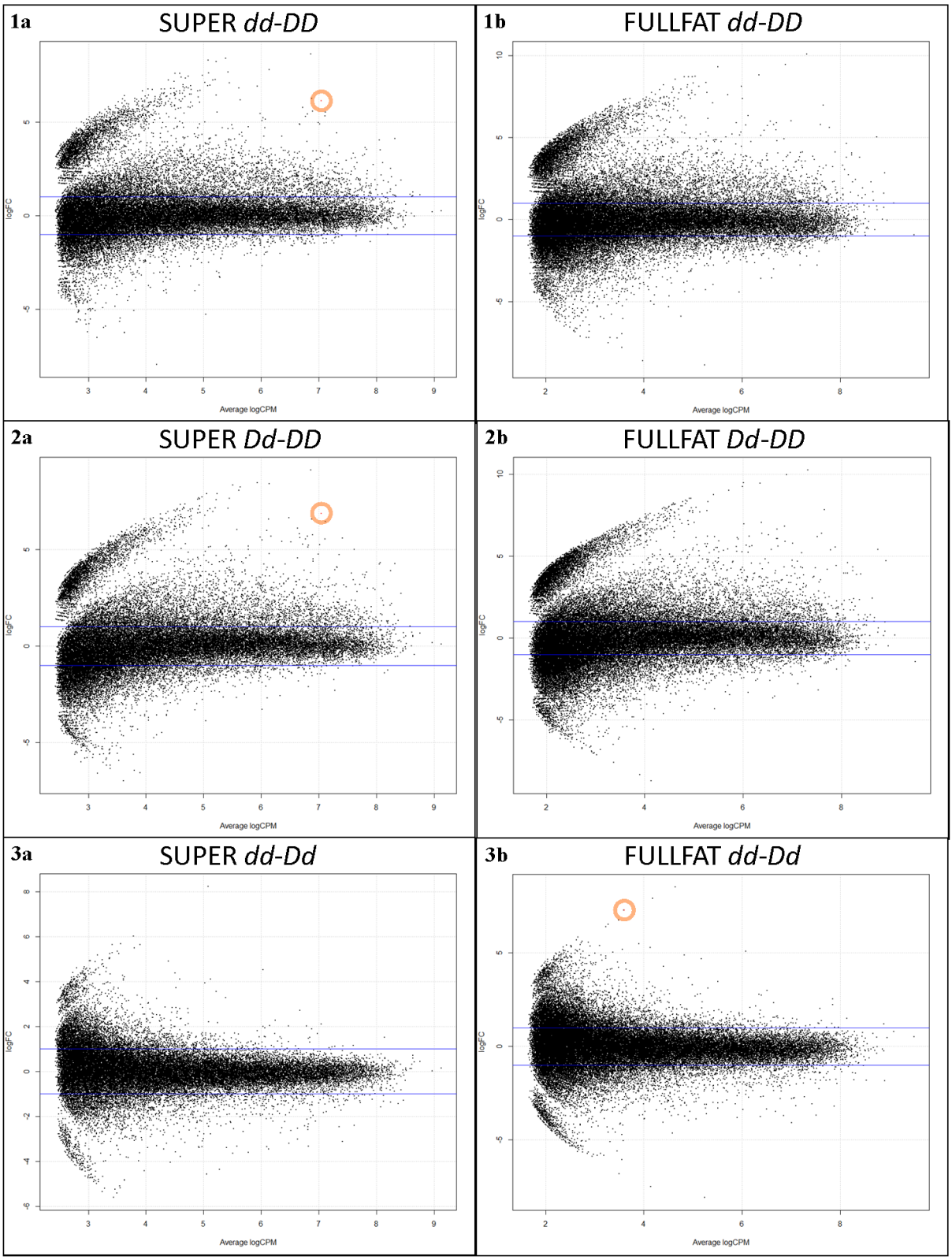


Figure 51 Log fold change in eRAD tag sequence counts between genotypes *DD* & *dd* (1); *DD* & *Dd* (2) and *Dd* & *dd* (3) in datasets SUPER (a) and FULLFAT (b).The direction of relative expression is indicated in the title of each plot. Each data point (shown in black) represents an eRAD tag. Significantly differentially expressed tags are shown in red and emphasized by a circle. The blue lines indicate a two-fold difference in expression.

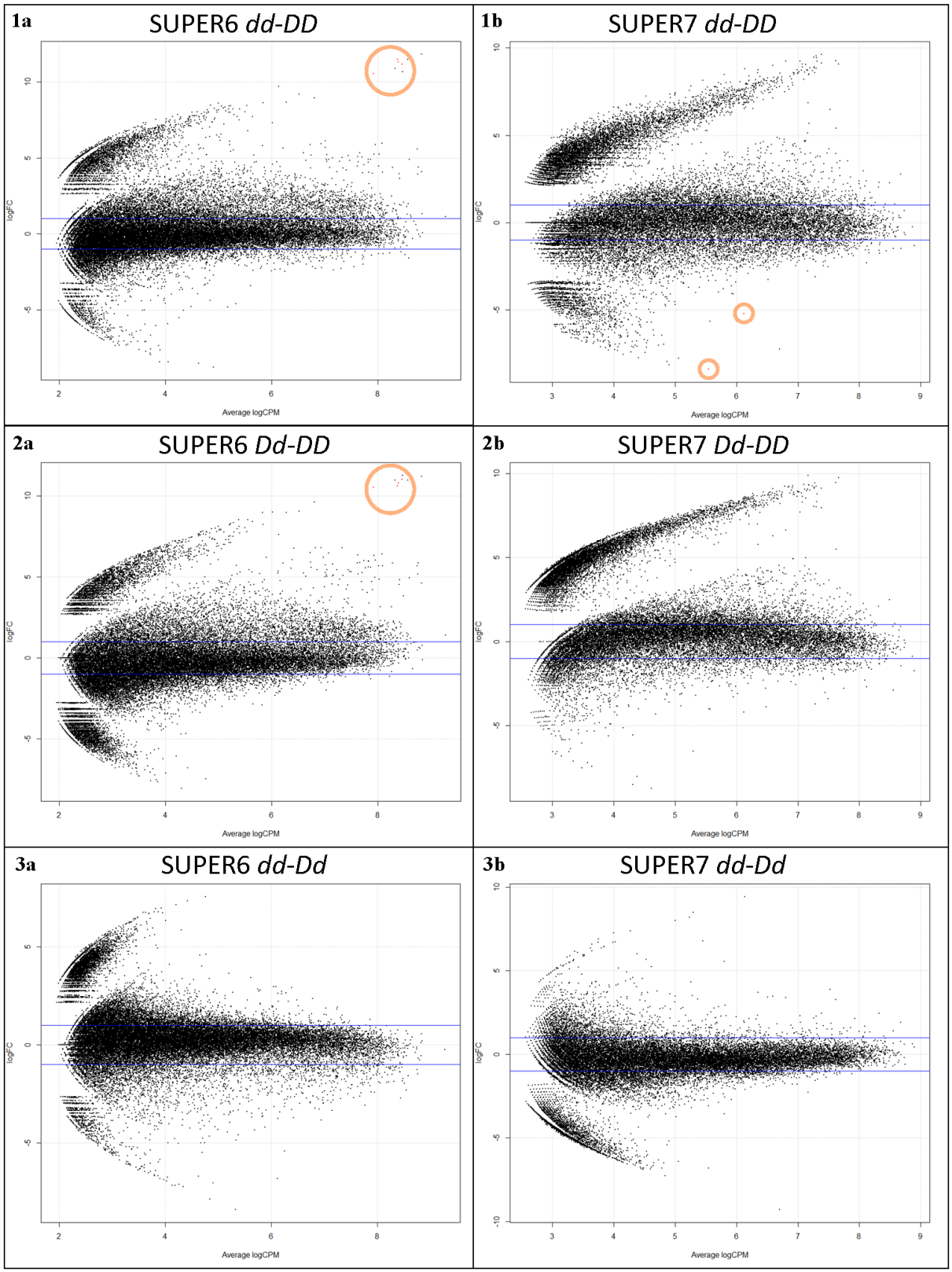


Figure 52 Log fold change in eRAD tag sequence counts between genotypes *DD* & *dd* (1); *DD* & *Dd* (2) and *Dd* & *dd* (3) in datasets SUPER6 (a) and SUPER7 (b).The direction of relative expression is indicated in the title of each plot. Each data point (shown in black) represents an eRAD tag. Significantly differentially expressed tags are shown in red and emphasized by a circle. The blue lines indicate a two-fold difference in expression.

Table 56 Statistically significant differentially expressed Radtags between genotypic groups. The *dd-DD*, 1+ shows the cat ID of the tag counts which were higher in the *dd* genotype compared the *DD* genotype, whereas 1- indicates the opposite expression pattern. In the *Dd-DD* comparison 1+ shows the cat ID of the tag counts which were higher in the *Dd* genotype compared to the *DD* genotype. Dt denotes which dataset the tags were identified in. The average log transformed counts per million (logCPM) and log transformed fold-change in tag counts (logFC) is shown for each tag. The probability value (p) is shown for each comparison with the probability value corrected for false discovery rate (FDR). An FDR of 0.000 refers to a value <0.001.

	dd-DD						*Dd-DD*					
Dt	1+	1-	logFC	logCPM	p	FDR	1+	1-	logFC	logCPM	p	FDR
SUPER	38405	-	6.13	7.04	1.12E-06	0.043	38405	-	6.85	7.04	4.01E-08	0.002
SUPER6	14295	-	11.49	8.37	5.71E-09	0.000	14295	-	10.62	8.37	5.66E-08	0.002
	49725	-	11.32	8.39	2.22E-08	0.000	49725	-	10.80	8.39	7.74E-08	0.002
	34376	-	11.17	8.46	6.80E-07	0.010	34376	-	11.04	8.46	8.47E-07	0.012
	49600	-	10.55	7.92	1.91E-06	0.021	49600	-	10.53	7.92	1.85E-06	0.020
SUPER7	-	11332	-8.41	5.54	9.55E-07	0.033	-	-	-	-	-	-
	-	14621	-5.21	6.12	2.68E-06	0.046	-	-	-	-	-	-

Table 57 Annotation of the differentially expressed Radtags (Cat ID) identified within each dataset (Dt) and the length of their associated paired-end contig (Ln), * indicates contig is a genomic contig. Annotation information shows the top Blastx hits, with percentage query sequence cover (Q), and associated E value and gene accession number (Acc.).

Tag description			Annotation			
Dt	Cat ID	Ln	Blastx	Q	E value	Acc.
SUPER	38405	159	PREDICTED: uncharacterized protein LOC106071482 [*Biomphalaria glabrata*]	96%	0.002	XP_013087059.1
			PREDICTED: von Willebrand factor A domain-containing protein 2-like [*Biomphalaria glabrata*]	43%	3.2	XP_013087030.1
SUPER6	14295	345	PREDICTED: uncharacterized protein LOC106053297 isoform X2 [*Biomphalaria glabrata*]	82%	6.00E-43	XP_013064285.1
			PREDICTED: involucrin-like [*Aplysia californica*]	81%	3.00E-24	XP_005091491.2
	49725	355	PREDICTED: leucine-rich repeat serine/threonine-protein kinase 1-like isoform X1 [*Biomphalaria glabrata*]	99%	5.00E-58	XP_013066303.1
	34376	154	PREDICTED: elongation factor 2-like [*Biomphalaria glabrata*]	64%	9.00E-10	XP_013091168.1
		185	No sig.	-	-	-
	49600	188	hypothetical protein UC40_25285 [*Escherichia coli*]	68%	1.00E-20	KJG94335.1
SUPER7	11332	367	No sig.	-	-	-
		*1964	PREDICTED: repetitive proline-rich cell wall protein 2-like [*Biomphalaria glabrata*]	29%	1.00E-06	XP_013074751.1
			PREDICTED: keratin-associated protein 4-3-like [*Opisthocomus hoazin*]	25%	8.00E-05	XP_009942640.1
	14621	x	-	-	-	-
		177	No sig.			
		*212	No sig.			

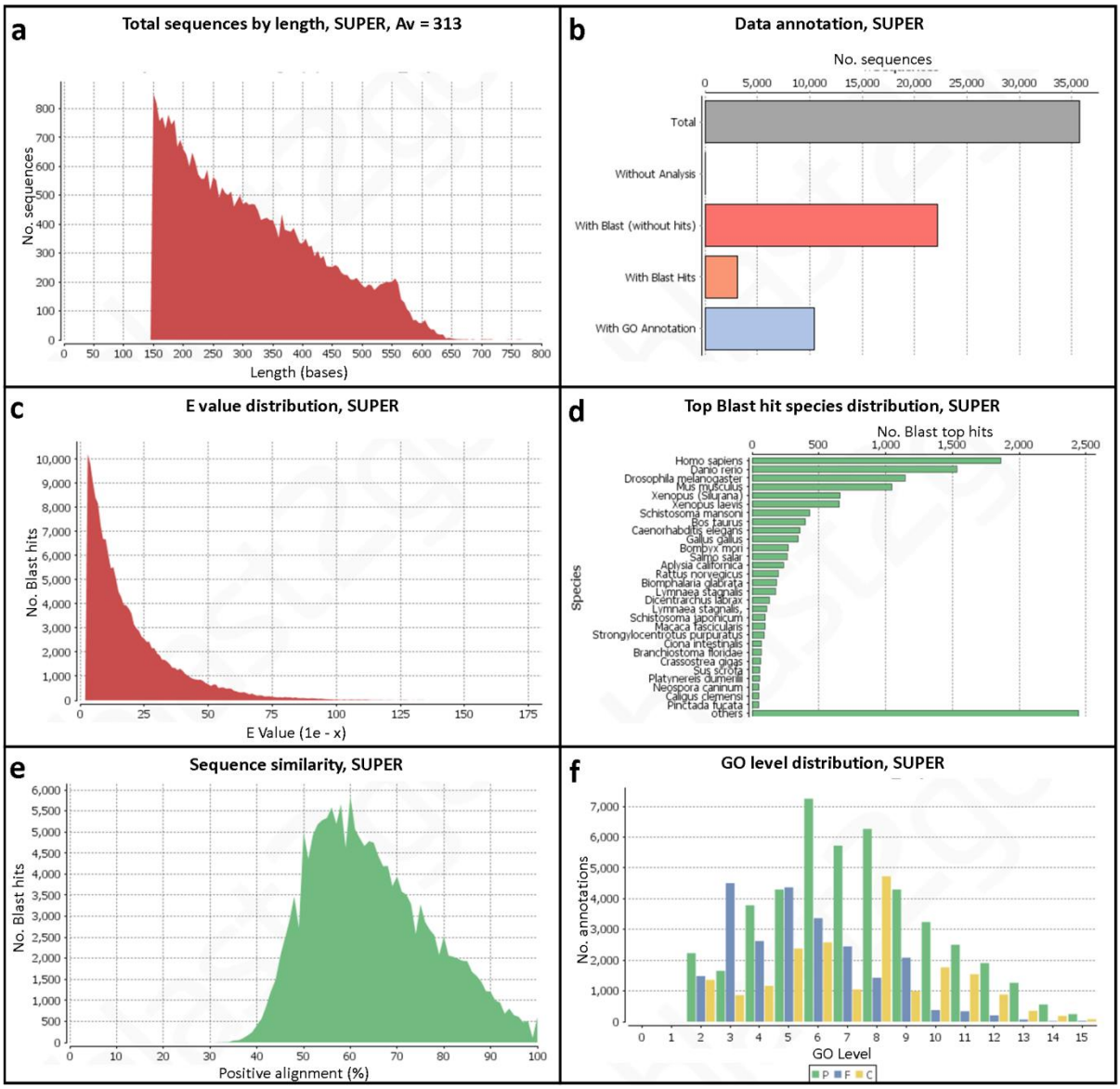


Figure 53 Descriptive summary of the 'SUPER' assembly and annotation. (a) shows the distribution of lengths of the 313 contigs, including average (Av) length of contigs. (b) provides a summary of the annotation analysis of the assembly, including the 'total' number of contigs, the number of contigs 'without analysis', the number of contigs without a blast hit following analysis 'with Blast (without hits)', the number of contigs 'with blast hits' without annotation, and the number of contigs 'with GO annotation' in addition to the Blast hit. (c) shows the distribution of E Values of the Blast hits within the assembly, starting from 0.1. (d) shows the number top Blast hits associated with a species. (f) shows the distribution of contig sequence similarity within the alignment. This is represented as a percentage, calculated from number of positive hits divided by the length of alignment. (f) shows the distribution of GO terms within the annotated dataset. 'Parent' (P) terms are shown in green, 'Fake' (F) terms are shown in blue and 'child' (C) terms are shown in yellow.

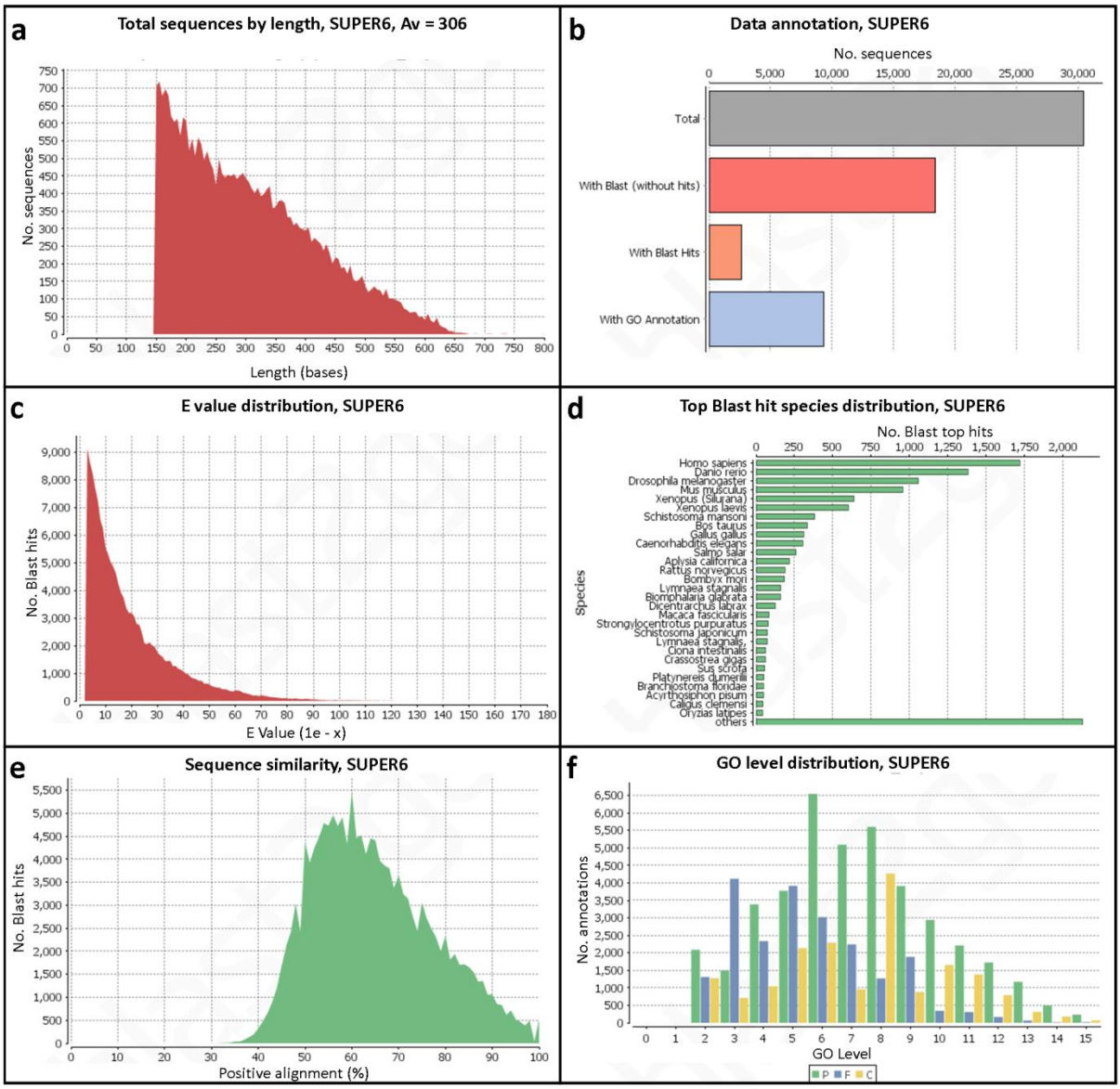


Figure 54 Descriptive summary of the 'SUPER6' assembly and annotation. (a) shows the distribution of lengths of the 306 contigs, including average (Av) length of contigs. (b) provides a summary of the annotation analysis of the assembly, including the 'total' number of contigs, the number of contigs without a blast hit following analysis 'with Blast (without hits)', the number of contigs 'with blast hits' without annotation, and the number of contigs 'with GO annotation' in addition to the Blast hit. (c) shows the distribution of E Values of the Blast hits within the assembly, starting from 0.1. (d) shows the number top Blast hits associated with a species. (f) shows the distribution of contig sequence similarity within the alignment. This is represented as a percentage, calculated from number of positive hits divided by the length of alignment. (f) shows the distribution of GO terms within the annotated dataset. 'Parent' (P) terms are shown in green, 'Fake' (F) terms are shown in blue and 'child' (C) terms are shown in yellow.

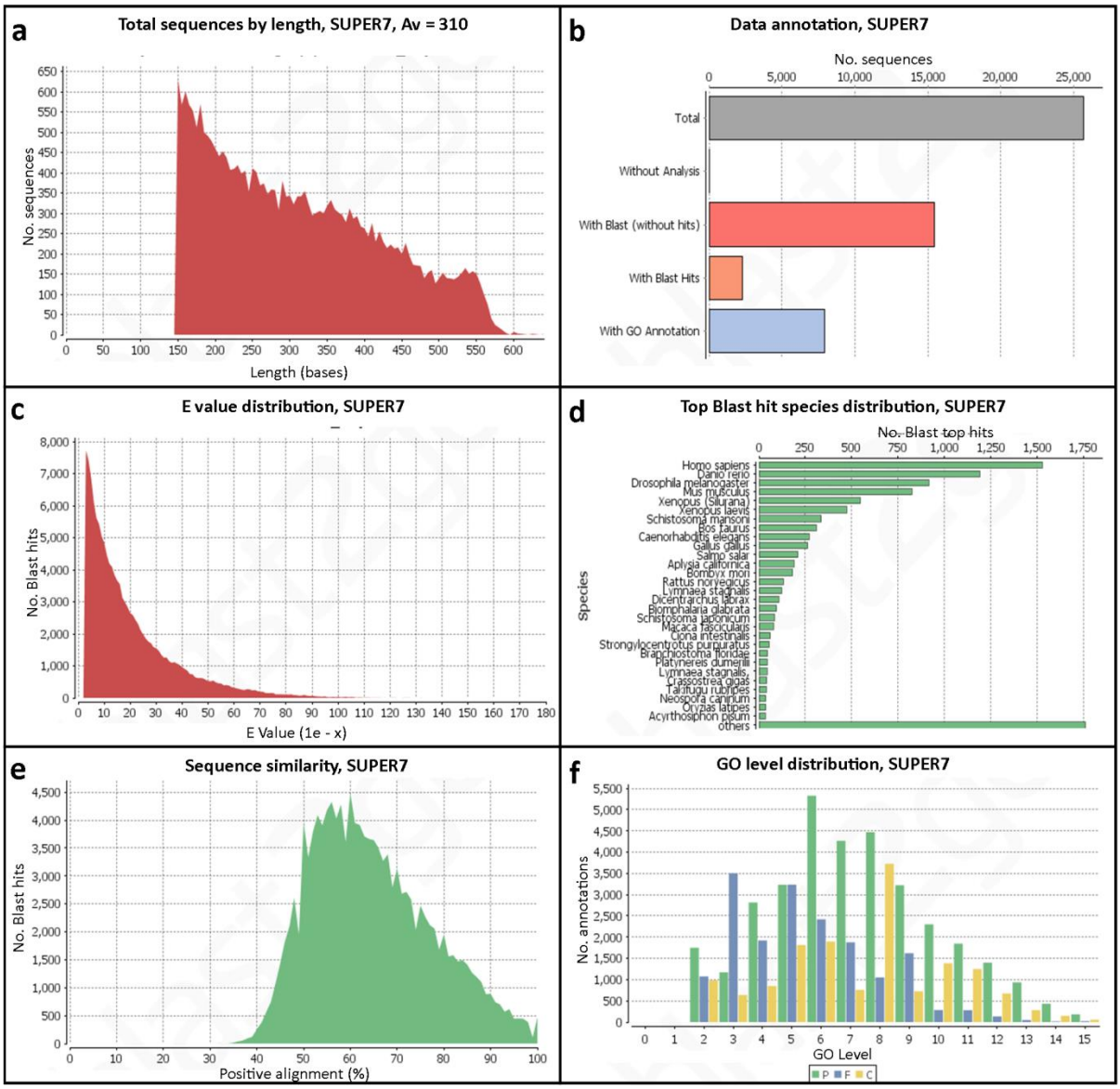


Figure 55 Descriptive summary of the 'SUPER7' assembly and annotation. (a) shows the distribution of lengths of the 310 contigs, including average (Av) length of contigs. (b) provides a summary of the annotation analysis of the assembly, including the 'total' number of contigs, the number of contigs without a blast hit following analysis 'with Blast (without hits)', the number of contigs 'with blast hits' without annotation, and the number of contigs 'with GO annotation' in addition to the Blast hit. (c) shows the distribution of E Values of the Blast hits within the assembly, starting from 0.1. (d) shows the number top Blast hits associated with a species. (f) shows the distribution of contig sequence similarity within the alignment. This is represented as a percentage, calculated from number of positive hits divided by the length of alignment. (f) shows the distribution of GO terms within the annotated dataset. 'Parent' (P) terms are shown in green, 'Fake' (F) terms are shown in blue and 'child' (C) terms are shown in yellow.

Discussion

Experimental design and library preparation

This experiment was intended to be a comparative analysis of the capabilities of eRAD and traditional RNA Seq. However, the RNA Seq data was not received until October 2015 and therefore could not be included within the analyses here or discussion. This has limited the ability to distinguish failings in the data as a result of the samples generated for sequencing or the eRAD method itself.

Sample quality

The samples were of very low starting concentration, subsequently the eRAD libraries were intended to be generated from 100 ng ds cDNA, one tenth of the 1 µg reaction in the RADSeq library preparation instructed by Etter et al. (2011) (100 ng was the minimum starting quantity utilsed in Baird, Etter et al. 2008). However, some samples were still unable to meet the requirements of the reduced setup (Table 49). To account for the more variable starting quantities of individual cDNA samples in library 4/L007, the amount of P1 adaptor was altered. This was with the aim of reducing the level of excess primer present within the reaction, which can result in primer dimer within the library. In light of the increased primer dimer visible in Library 4/L007 this is unlikely to of had much influence (gel images of libraries are presented in the SI, S11.3).

The starting concentration was firstly limited by the finite and relatively small amount of starting material, the ovotestis organ (<30 mg). Further mRNA selection reduced the total RNA samples to approximately 10% of their original concentration (data presented in the SI, S11.1). Initial trials of the library preparation using total RNA instead of mRNA (Library 1 and Library 2, *data not presented*), showed general failure to produce a good library concentration when visualised via gel electrophoresis. Although total RNA holds the potential to show informative regulatory mechanisms in gene expression (as introduced in Chapter 1, Gene expression analysis), the overabundance of ribosomal RNA (rRNA) was predicted to dominate the sequence library and obscure information from more informative protein-coding mRNAs. Poly-adenylation is not entirely limited to protein-coding genes, and is present on both long and short ncRNAs (Carninci, Kasukawa et al. 2005, Grzechnik and Kufel 2008, plus additional examples summarised in Tarazona, Garcia-Alcalde et al. 2011), although the size-selection step should omit the majority of ncRNAs (<300 bases). Furthermore, rRNA represents approximately 80% of the total RNA sample therefore it is likely there will be some carryover following the mRNA enrichment (AppliedBiosystems 2008). As such it is assumed that some non-coding RNAs will still be present within the eRAD sequence data, provided that they contain a PSTI recognition site.

Although all ds cDNA samples were generated from approximately 500 ng mRNA (494.3 ng - 504.3 ng), the ds cDNA sample yields were substantially varied (29.64 ng - 286.20 ng) (Table 49). This may be a reflection of sample quality differences. No correlation was apparent between mRNA yield or 260/230 ratios and the final ds cDNA yield, yet a potential positive relationship was observed between the 260/280 ratios and the total ds cDNA yield (R^2 = 0.6041, plots shown in the SI, S11.2).

Choice of samples

The ovotestis tissue was selected for this experiment, with the aim to limit sequencing to unfertilised eggs, thereby capturing maternal RNAs known to contain the heritable chiral determinant. However, this tissue has produced a number of confounding variables as discussed in previous chapters. Firstly, due to the internal organisation of the *L. stagnalis*, it is impossible to extract the ovotestis without some level of carryover of liver tissue (Figure 9). No subsequent treatment or washing steps were performed on the extracted ovotestis tissue in order to minimise handling of the tissue prior to the RNA extraction. Therefore, it is assumed that some sequences within the eRAD libraries are in fact specific to the liver tissue, not the ovotestis. The extent of this carryover is discussed further in section: 'eRAD Sequencing data, QC'. Secondly, there was no way to control the biological timing of the tissue extraction.

As seen from the qPCR data in previous chapters, the ovotestis exhibited more variable gene expression than the embryo tissue. This is likely due to variation in both the probable liver tissue carryover and in the sampling time. In hindsight the single-cell embryo tissue would have provided a cleaner and better controlled tissue to use within the eRAD experiment. However, at the time of this experiment, I did not yet have the skills to prepare embryo RNA samples. Additionally, the samples would have required a much larger number of embryos than used in the qPCR experiments to generate an mRNA enriched sample of sufficient quantity.

A total of 22 individuals were included in the eRAD libraries. It was initially thought that both library preparations would be sequenced on one lane, however each library was allocated an entire lane. Therefore, it may have been beneficial to include more individuals due to the increased amount of sequencing capacity. Moreover, the reduced number of *DD* individuals inhibited downstream DE analyses. If repeated, it would be recommended to include equal numbers of individuals from the comparable groups. Additionally, the unbalanced representation of genotypes within the sequenced libraries may have resulted in a biased transcriptome assembly, due to potential sequence differentiation in the underrepresented samples being dismissed as error. This was not believed to be a large issue in this dataset due to the relatively low amount of sequence variation observed,

however to combat this bias, genotype specific contig assemblies could have been generated including only individuals of that genotype.

Choice of enzyme

PstI was chosen as the restriction enzyme for the eRAD sequence library. This was due in large part to the availability of specific barcode sequence adaptors. The end of each adaptor was complementary to the PstI cut site overhang (5'- TGCA -3') and therefore could only be used with the restriction enzyme PstI or SbfI, which generates the same sequence overhang. The frequency of PstI recognition sites within the *L. stagnalis* transcriptome was unknown prior to this experiment, yet PstI was selected for the digest instead of SbfI as it is predicted to be a more frequent cutter due to its recognition sequence being two bases shorter than that of SbfI (Figure 45). Alternatively, both enzymes could have been used together in a double digest reaction to further increase the number of sequencing start sites within the libraries. The choice of restriction digest within the RAD library preparation depends on whether capturing an increased number of sequence loci or an increased depth of sequencing of a smaller number of loci is a priority. Again because the number of PstI sites across the *L. stagnalis* transcriptome was unknown, the single PstI digest represented a mid-range of the three possible digests.

eRAD Sequencing data

QC

L007 generally showed higher sequence quality but also higher sample sequence bias and PCR clones, than L006. There were two key differences in the library preparations of L006 and L007, which could have resulted in these differences. Firstly, L007 was gel extracted twice to reduce the level of primer dimer within the library, whereas L006 was only extracted once. This may account for the increased sequence quality in L007.

Secondly, the starting quantity of input cDNA (and subsequent volume of P1 adaptor added) was much more varied in L007. The sample representation bias did not however, show any strong correlation with the input cDNA quantity or concentration or the amount of P1 adaptor added to the reaction (correlation plots are presented in the SI, S13). The Illumina sequencing platform is observed to show a GC sequencing bias resulting in increased coverage of GC rich sequences (Minoche, Dohm et al. 2011). It is possible that the sequencing method exhibited variable preferences for the five base unique identifier sequence at the start of all individual transcripts. However, no strong correlation was observed between the total retained sequences of each individual and the percentage GC of their specific barcode adaptor (S13). This phenomenon of

sample sequencing bias has been observed in other RADSeq libraries within the Davison research group and other publications, and remains largely unexplained. Degradation of DNA samples has recently been implicated as a cause for reduced sequence depth of samples in RADSeq (Graham, Glenn et al. 2015). Here, no strong correlation was seen between 260/230 or 260/280 ratios of mRNA samples and total retained reads (S13). However, no quality values are available for the ds cDNA; therefore, there may have been differences in the ds cDNA quality not detectable in the quantification which could explain the difference in sample sequencing effort.

Each of the four raw sequence files failed the Fast QC check for overexpressed sequences. The majority of these in L006 were described as the Illumina sequencing primers, whereas in L007 the primer sequences were reduced (*data not presented*). This may reflect the effectiveness of the second gel extraction performed in the preparation of L007. The overexpressed reads were flagged in the raw reads and therefore still contain the 5 base unique barcode. As such many of the overexpressed sequences are in fact the same sequence but from a different individual. The Fast QC failed to provide a descriptive hit for any sequence other than the Illumina primers, due to the short 50 base sequence presented. Local Blasts to the SUPER assembly revealed the pe contigs associated with the highest overexpressed sequences were described as 'ferritin-like' or 'soma ferritin'.

Ferritin represents the primary yolk storage protein in *L. stagnalis* (Bottke 1986). Therefore identifying ferritins within the overexpressed sequences provides support that the sequences were generated from the ovotestis. However, *L. stagnalis* contains two types of the iron storage protein ferritin: soma-ferritin and yolk- or vitellogenic - ferritin (Bottke and Crichton 1984, Vondarl, Harrison et al. 1994). Identifying soma ferritin specifically, would indicate there were high levels of somatic tissue contributing to the dataset. Due to the short reads provided in the Fast QC summary, the appropriate level of sequence specificity is lacking in order to determine which of the ferritins is overexpressed. There is no reason to my knowledge why soma ferritin would be present in high levels in the probable carryover liver tissue and even so the level of carryover tissue relative to the extracted ovotestis tissue is minimal. It therefore seems likely that the overrepresented sequences reflect the *L. stagnalis* yolk protein.

Overall the sequencing data was not of high quality. The low starting quantity of the samples within the library preparation is believed to have contributed in large part to this. The RNA Seq experiment used the remaining mRNA of each of the individuals within the eRAD libraries. These samples were of even lower starting quantity, therefore if the sequence quality of the RNA Seq data is improved, the generally poor quality sequence data here can be more confidently attributed the eRAD library preparation and methodology as opposed to the sample quality.

Descriptive analysis

Throughout the data analysis accommodations had to be made to account for the fact that the sequence data neither represents RADSeq or RNA Seq data.

Process radtags

In addition to the basic commands within the *process_radtags* program to allocate sequences to the individual samples, such as specifying the paired-end data and restriction enzyme, a number of flexible parameters were included, which are discussed here.

All sequence reads were truncated to a length of 70 bases using the function '-t'. This was performed to accommodate the substantial reduction in quality of sequence reads greater than 70 bases in L006 (Figure 46). Due to the improved quality of sequence reads in L007, the individual reads in this library could have been truncated to 90 bases, or not at all. However, the individuals of each library were intended to be compared to each other in downstream analyses and as such were treated using the same parameters.

Sequencing reads containing a barcode which did not match those specified in the experiment were able to be 'rescued' and included in the dataset if the program could confidently match the barcode using the function '-r'. This required the distance between barcodes to be specified. Of all the barcodes used in the experiment the maximum aligned sequence similarity was calculated as two bases (calculation shown in SI, S11.4). Therefore, the barcode distance was specified as three. This function may increase the inclusion of sequences from the multiple individuals through coincidental sequencing error, although avoids the discarding large amounts of the sequence data. For a more rigorous analysis this function could have not been included. Due to the relatively low number of individuals included on each sequencing lane, each individual likely received enough coverage to withstand some loss of sequencing reads. However, in light of the high level of sample sequencing bias this was not the case and therefore rescue barcodes function was incorporated to retain as many reads as possible per individual.

The function '-q' discards reads with low quality scores (as specified within the raw data file). Similarly, the function '-c' cleans the data by removing any read containing an uncalled base. These filters could result in low quality individuals containing lower sequence counts for tags, however due to the short sequence length of the reads, sequence specificity is highly important. Additionally, the poor quality sequence reads should be equally distributed across the individual's total reads and therefore should not result in a systemic bias in the resulting count data.

Process_radtags was performed independently for each of the sequenced libraries. When performed on all individuals from both libraries simultaneously (data not shown), the output was varied indicating that a number of individual were allocated reads from the other sequencing lane, likely through coincidental sequencing errors or the rescue barcodes function.

Clone filter

The *clone_filter* program removes any read which contains a R1 and R2 sequence pair identical to one already identified in the individual. In traditional RADSeq data analysis this aims to reduce sequencing errors amplified through PCR. There is some debate however as to the appropriate use of the clone filter in quantitative sequence data and the removal of PCR duplicates has been advised against in RNA Seq data analysis (RNA Seq data analysis short course, University of Leicester, *pers. comm.* Nov 2014).

Because the lengths of mRNA transcripts are substantially shorter than those of genomic DNA sequences, there is a greater probability of duplicated sequence reads by chance. The presence of PCR duplicates should affect all individuals equally and therefore their inclusion would not create a systemic bias. Shorter genes would be more likely to contain duplicates by chance, again all individuals should experience this bias equally and therefore it will not systemically bias count data comparisons between genotypic groups. Thus the clone filter could potentially remove genuine counts from individual tags and obscure DE patterns. Subsequently the analyses were performed on both clone-filtered and non-clone-filtered datasets to compare the effect of the clone filter on quantitative gene expression analysis

The eRAD libraries were size selected to include only sequences of 300-700 base-pairs long. Therefore 1 in every 400 reads per tag per individual could hold an identical paired-end (pe) read by chance alone. Thus the clone filter would have identified on average 0.25% of pe reads to be clones by chance, however the average percentage pe reads identified as clones was >70% (Table 51).

Although the eRAD libraries may contain a higher proportion of shorter fragments, resulting in more than 1 in every 400 pe contigs representing a clone, it is unlikely that the high number of removed duplicate counts were present due to chance. Furthermore, the individuals did not show an equal proportion of clones within their pe reads, which ranged from 51.29% clones to 87.09% (Table 51). Therefore, it is assumed that the FULLFAT dataset contains a large proportion of counts generated from PCR clones with substantial variation between individuals.

Overall, following the sequence data preparation only ~30% of the original total reads was retained for analysis. Thus the eRAD sequencing has not provided a very efficient use of NGS power. The

majority of RAD sequencing publications only quote total read counts following *process_radtags* and *clone filter*, and so it is difficult to ascertain whether this is common.

Stacks parameters

The individual samples within the eRAD sequencing datasets were all offspring from a self-fertilised *Dd* snail. Subsequently the data was treated as a genetic cross experiment, although there was no sequence data for the parent. This was combated by creating a 'superparent' sample. In order to produce two independent datasets for downstream analyses, L006 and L007 were analysed separately. Statistical power of the genotypic group comparisons was expected to be hindered due to the low number of *DD* individuals present within each library, and so a combined catalogue including every individual within the experiment was also generated.

Whilst the only parameter altered from the default settings in the denovo_map stacks program was the minimum depth required to create a stack in the catalogue, '-m'. The central parameter settings are briefly discussed here.

Firstly, the number of sequence differences (SNPs) allowed within a stack was kept at the default value of two. Any stack containing more than two base differences was classified as a new stack. It was not expected that there would be more than two naturally occurring SNPs present within the 70 base consensus RAD tag sequence. Therefore, these sequences represent either sequence error or a new stack. Following this, the maximum number of stacks allowed per locus was kept at the default value of three. Theoretically, due to the self-fertilisation of the diploid heterozygote individual snail, there was a maximum of two possible different stacks (alleles) per loci. However, this was kept at three to allow some room for error. The number of differences allowed *between* loci was kept at the default value of zero. Because the catalogue was generated by the superparent, in the event of differentially fixed loci being present between individuals, they will have already been included within the superparent as one locus and classified as polymorphic.

The default minimum sequencing depth required to create a stack, '-m', is three. This default value was maintained for the progeny, however due to the extremely large number of reads within the superparent, the minimum stacks depth of three was expected to include a high number of erroneous tags. Of course it is impossible to know the 'correct' parameters to use to assess the eRAD sequencing data, however summaries of the total catalogue loci generated from varying the -m value provided inferences into what may represent sequencing error and what is true variation (Stacks summaries are presented in the SI, S14).

When -m was small, this generated far higher numbers of unique stacks. However, a large number of these stacks were only present in the parent and not the progeny. This demonstrates that these sequences were not present in any one individual to the minimum stacks depth of 3 and were therefore likely generated from the combined sum of few erroneous reads across multiple individuals. When -m was increased, this generally increased the number of stacks present in a larger number of progeny, providing far more informative tags. Additionally, due to the level of inbreeding of the samples included, it is unlikely that there will be substantial biological variation in the tags present within individuals.

Another indication of the increased error produced when -m was small, was demonstrated by the discrepancy between total number of stacks in the catalogue and those present within the superparent. The catalogue was generated from the superparent consequently every stack should be identified within the parent. The chosen -m value was selected to maximise the number of stacks generated, whilst minimising large differences in the number of stacks present within the parent and increasing numbers of progeny (S14). This was believed to signify a more stable and biologically feasible dataset.

The large differences shown in the number of unique stacks generated in the SUPER and FULLFAT datasets raise questions about the use of the clone-filter. The clone filter removes only the duplicate sequence and leaves the original copy (Julian Catchen, *pers. comm*. April 2015, Stacks Google group). Therefore, the SUPER and FULLFAT catalogues should show the same unique stacks, but hold different counts. This is clearly not the case as the FULLFAT dataset identified many more unique stacks than the SUPER dataset (Table 52). It is assumed that following removal via the clone filter, these additional unique sequences fall below the minimum stack depth threshold to generate a new stack and are therefore not present in the clone-filtered catalogue. This suggests that they are in low unique/true quantity and likely the product of PCR generated sequencing errors.

Paired-end contig assemblies

The pe contig assemblies were generated only from the clone-filtered data. Although losing a large number of sequence reads, this data was chosen to generate a more robust assembly in light of the likely increased sequencing error within the non-clone-filtered data. Additionally, the superparents were not included in the assemblies to avoid pseudo replication of the input sequences.

The *de novo* assembler Velvet was used as advised in the stacks tutorial ('building mini-contigs from paired-end sequences', Available at: http://catchenlab.life.illinois.edu/stacks/pe_tut.php). Importantly the Velvet aligner allows for multiple contigs per locus and is therefore suitable for the transcriptomic sequence data here. The standard Velvet aligner has however been shown to not

represent the best *de novo* contig assembler (Davey, Cezard et al. 2013). Assemblers are available, which better incorporate the variable nature of the coverage and heterozygosity of RAD data, such as 'VelvetOptimiser' (Victorian Bioinformatics Consortium, Copyright 2009 - Simon Gladman). This is a wrapper script which can be added to the Velvet aligner and has been identified to produce the best de novo assemblies, although requires a substantially greater computing effort, quoted as 200 times longer than other assemblers (Davey, Cezard et al. 2013).

The contigs generated through the Velvet assembly of the eRAD sequence data have provided successful templates for the qPCR genes of interest described in Chapter 3 and identified intron-exon junctions successfully (Figure 12). Thus the quality of assembly is assumed to be of acceptable quality for use within this project. However, one final reference assembly generated from the combined SUPER dataset may be repeated using a higher quality assembler for publication.

Because eRAD is a reduced representation method, the gene ontology (GO) does not provide a true summary of the functional processes within the ovotestis. DE tags were intended to be tested for enrichment of cell function compared to the overall GO of the dataset however due to the overall lack of DE, this was not performed.

Differential expression analysis

In performing DE analysis, the sequencing depth of eRAD tag data was treated as the equivalent of RNA Seq raw count data lacking a reference genome. As previously introduced, a number of important facets will affect the appropriateness of DE comparisons derived from sequence read depth. This subject has been addressed in many publications (including Bullard, Purdom et al. 2010, Robinson and Oshlack 2010, Tarazona, Garcia-Alcalde et al. 2011). The main considerations are discussed here in relation to the current eRAD DE analysis within edgeR.

EdgeR was selected to test the count data for DE between genotypic groups because it has been shown to be one of the most stable software packages available for DE analysis and the most appropriate for RNA Seq data with low numbers of replicates (Kvam, Lu et al. 2012, Schurch, Schofield et al. 2015). DE results were originally intended to be supported by repeated analysis using DESeq (Anders and Huber 2010), another top ranked DE analysis software (Kvam, Lu et al. 2012, Schurch, Schofield et al. 2015). DESeq has been shown to be better suited to datasets with more than 12 replicates and has a tendency for false positives, therefore was only intended to support results from the more appropriate edgeR analysis (Schurch, Schofield et al. 2015). Due to the lack of significant results from the edgeR, the DESeq analysis was not considered informative and therefore is not presented here.

One of the primary factors when comparing count data is normalisation. As in the qPCR comparisons of gene expression, comparisons will have to standardise for starting quantity, yet NGS data must also account for varied sequencing effort. Normalisation strategy continues to be debated (eg. Li, Piao et al. 2015), however it is generally agreed that more sophisticated methods of normalisation are need than simply dividing by total read counts per individual. For example, if sample A expresses all genes in equal quantity to sample B yet expresses a selection of genes that are absent in sample B, this will result in sample A having higher overall read counts and inaccurate 'normalisation' of the expression of genes present in both samples (Robinson and Oshlack 2010). The TMM (trimmed mean of M values) normalisation method employed in edgeR generates a global fold-change gene expression estimate between each individual for all genes/tags with the assumption that the majority of genes are not DE. Thus generating relative library sizes for each sample (Robinson and Oshlack 2010). Comparisons of normalisation strategies have demonstrated the method of normalisation to be critical to identifying DE. In light of edgeR ranking as one of the most reliable methods to identify DE (Schurch, Schofield et al. 2015) supports this strategy of normalisation.

Accurate assumptions about the distribution of the count data are essential for appropriate use of statistical modelling. EdgeR assumes a negative binomial distribution of the data to account for the expected high number of zero counts present. A recent study performing RNA Seq with 48 replicates has enabled testing of assumptions regarding the distribution RNA Seq count data. The observed read counts were consistent with both log-normal and negative binomial distributions providing further support for the use of edgeR (Gierlinski, Cole et al. 2015). EdgeR employs parametric tests when comparing count data for DE. However, the small number of replicates in NGS experiments usually prohibits testing the fit of the data for use in parametric tests. The eRAD data here had extremely high levels of variance which did not follow a Poisson distribution (Figure 49.3, Figure 50.3). Although tagwise dispersion was employed rather than assuming equal variance, the appropriateness of parametric testing in this dataset is questionable. Non-parametric RNA Seq DE analysis methods are available such as NOISeq (Tarazona, Garcia-Alcalde et al. 2011). However, this method has not been recommended for use due its lack of consistency with other methods (Schurch, Schofield et al. 2015).

The removal of individual samples did not result in a large reduction in variance of the overall dataset and in some cases increased variance (represented by the BCV values, Table 55), although this is believed to be partly due to the counts per million (CPM) filter being relaxed to account for the reduced number of individuals within the smallest group (Table 55). Each dataset had one potential outlier individual, namely 9001 and 9009 (Figure 49, Figure 50). The analyses could have

been repeated following the removal of these individuals however, one outlier in one genotypic group does not provide an explanation for the high levels of variation and lack of significant DE in group comparisons in which it is not included.

Only monoallelic RAD tags were included in the DE analysis. This was due to the inability to export allele-specific count data from stacks. Pooling of the counts was considered, as a number of the multi-allelic tags likely represent sequencing errors of an invariable sequence tag. Yet this would not be appropriate for all tags due to the risk of obscuring true allele specific DE (Wood, Nones et al. 2015). Those tags which likely represented sequencing error could be inferred from the distribution of counts across the alleles (i.e. very low reads counts of the erroneous transcript); however, the number of multi-allelic tags would require an automated method to perform this which was not developed.

As previously stated, higher count transcripts have greater statistical power. The log-fold change plots reveal clusters of tags with clear fold-change differences, likely representing presence/absence relationships between the genotypic groups. The few tags identified as DE are generally found to be of the highest counts per million (CPM) in the comparison (Figure 51, Figure 52). In light of the massive reduction in the probability value required for significance through corrections for FDR, no transcript with a relatively small fold-change will be identified as DE. Furthermore, the low number individuals within the *DD* sample group, and unbalanced sequencing efforts between groups further inhibited statistical power of comparisons.

Differential expression results

Of the complete SUPER datasets, only seven tags were identified as DE, five of which identified a characterised function (Table 56, Table 57). The associated functions of the tags identified as DE are briefly discussed here, although it is unlikely that these relationships reflect a true biological meaning given that the random group dataset (Rm) identified a comparable number of DE tags as the experimental group data (Table 55). Furthermore, the majority of DE tags were identified as absent in the *DD* individuals, or present in very low count in only one *DD* sample (normalised count data not presented). The *DD* group contained the lowest number of replicates and therefore absence in *DD* group is the most likely occurrence by chance.

The von Willebrand factor A domain is involved in cell adhesion, extracellular matrix proteins, and in integrin receptors essential for cell-cell interactions (Whittaker and Hynes 2002). These functions hold the potential for important functional associations with chirality. The von Willebrand factor C, also referred to as 'brorin' has been characterised as an antagonist of bone-morphogenetic proteins

(BMP) which are essential for morphogenesis and development (Miwa, Miyake et al. 2009). If the transcript functions as an inhibiting factor, overexpression may be associated with the sinistral developmental vulnerability and provide an explanation for the absence in the *DD* group.

Involucrin is a soluble protein and provides a marker for keratinocyte terminal differentiation and ultimately cross-links to form part of the cornefied envelope (Watt 1983, Eckert, Yaffe et al. 1993, Steinert and Marekov 1997). Thus it represents a gene present in stem cells, is functional in early developmental stages and holds structural properties, and subsequently could be involved in LR organisation.

The leucine-rich repeat serine/threonine-protein kinase 1-like gene relates to a number of functions activated by GTP binding (Korr, Toschi et al. 2006). The diaphanous formins are known to act as effectors of Rho GTPase binding, which may play a role in cell polarisation (Nakano, Imai et al. 2002, Wallar and Alberts 2003); therefore, it is possible that the GTPase activated kinases may exhibit downstream consequences from the loss of *Ldia2*. Additionally, a gene 'RPK1' in Arabidopsis, listed in Uniprot (http://www.uniprot.org/uniprot/Q9ZRF9) as a 'probable LRR receptor-like serine/threonine-protein kinase RPK1' has been implicated in embryonic pattern formation (Nodine, Yadegari et al. 2007).

Elongation factors are required for protein synthesis, and therefore unlikely to truly be absent in the *DD* individuals, however elongation factor 2 has been observed to co-localise with actin filaments, which could relate potential differential expression to chirality (Shestakova, Motuz et al. 1993). The identification of an elongation factor in the DE tags is of additional interest due to their frequent use as normalising control genes within relative quantification experiments such as qPCR.

The only tag with a description in the SUPER7 analysis was a keratin associated gene. In addition to the involucrin-like gene, the presence of two DE tags associated with keratinisation would provide greater support for a biological cause of the observed DE if each tag hadn't demonstrated the expression differentiation in opposite directions (Table 56).

Each of these tags holds the potential to be functionally associated with chiral dimorphism and as such should be tested further to explore differential expression relationships. As the majority of tags contain a pe contig, *L. stagnalis* qPCR assays could readily be designed to test the DE relationships with a greater number of replicates and also in the better controlled single-cell embryo tissue.

Overall the pattern of gene expression was very similar in both the FULLFAT and SUPER datasets (Figure 49, Figure 51). Due to the expected prevalence of PCR duplicates present in the FULLFAT datasets, only the SUPER datasets were presented here. However, it should be noted that generally

the SUPER datasets contained less tags and identified more DE tags than the FULLFAT datasets (Table 55). It is possible that the increase in significantly DE tags may result from the relaxing of the FDR correction due the reduced number of multiple comparisons performed in the SUPER datasets. Another possibility is that the sequence counts without flagged PCR duplicates fall below the detectability threshold in the SUPER datasets resulting in more presence/absence relationships.

Although the clone-filtered and non-clone-filtered datasets showed a general similar pattern of gene expression, the SUPER6 and SUPER7 datasets showed notable differences in the overall distribution of gene expression. As seen in the *dd-Dd* fold-change expression clusters of presence/absence relationships between *dd* and *Dd* groups, which are in opposite directions (Figure 52.3). This variation coupled with the potential library segregation shown in the MDS plots (Figure 49) highlights potential bias in the library sequencing and supports the analysis of library specific datasets.

The overall lack of DE likely indicates a failing of the eRAD sequence data to generate adequate quantitative gene expression data to identify statistically significant variation. However, it has additionally shown that the sinistral individuals have not shown a large-scale reduction of gene expression. Therefore, the pleiotropic effects known to occur in the sinistral developing *L. stagnalis* (Davison, Barton et al. 2009) do not appear to of stemmed from loss of function across a large number of genes.

Further analyses

As signposted throughout this chapter, there are apparently endless ways to analyse the wealth of bioinformatics data received from the eRAD libraries. The eRAD data was initially analysed as a genetic cross as this is how the snail samples were generated. In doing so, the stacks program only generates catalogue loci that are present within the parent. Consequently, an artificial superparent was used to generate the library catalogue without loss of information. However, the use of a superparent results in very high sequence counts and the inevitable inclusion of high counts of erroneous reads pooled from all individuals. Alternatively, the data could have been analysed as a population. Stacks population analysis generates the catalogue from all individuals and therefore would not have required the superparent and maintained restrictions of sequencing depth per individual. Although as the populations program has been designed specifically for population genetics applications it did not represent the relevant option for analysis.

Although not presented in the results it has been possible to identify candidate genes within the eRAD count data by performing local Blasts. This has revealed that the *Ldia2* candidate was captured

by the eRAD method yet failed to identify DE, further highlighting the inability of the current eRAD data analysis to identify DE. However, any candidate gene identified presently or in the future can be retrospectively checked against the eRAD sequence data to infer sequencing depth and possibly a transcriptomic sequence contig. It is also possible to export the normalised count data from edgeR and subsequently view relative expression of the specified tag.

In light of the potential differences revealed in L006 and L007 expression data a general linear model (GLM) approach could have been employed to assess the combined datasets including both genotype and sequencing library as factors (this is an option available in edgeR). However, the data would be unlikely to fit the model parameters due to the high and varied dispersion (Figure 49). Furthermore, as the libraries were analysed for DE separately, L006 and L007 raw data could have been prepared using different parameters in process_radtags. Due to the higher sequence quality of L007, sequence reads could have been trimmed to only 90 bases as opposed to 70 (Figure 46).

The RNA Seq data has recently been received (October 2015) and consequently the comparative analysis of sequence data can now be performed. This will provide essential information as to whether the limited capabilities of the eRAD sequence data here was due to the eRAD method itself or inherent limitations of the *L. stagnalis* ovotestis samples. It is possible that the eRAD method in general may not be able to provide reliable count data. A recent publication described the use of a novel sequencing method restriction-site associated RNA sequencing (RARseq), which is in principle the same as eRAD (Alabady, Rogers et al. 2015). RARseq was employed with the aim to facilitate genotype by sequencing methods in the pitcher plant with a large genome and do not appear to have utilised the quantitative capabilities of the RNA sequence data. The omission of quantitative analyses may be due to lack of interest or preparing for a second publication, although the lack of publications on eRAD from the Cresko group may indicate that reduced representation RNA sequencing data may not provide the quantitative capabilities anticipated.

The genotype analysis was not a priority for the eRAD data due the ongoing genomic analyses within the Davison research group. However, to perform a comprehensive assessment of the capabilities of the eRAD method this will have to be included, particularly in light of the apparent failings in DE analysis.

Conclusion

The eRAD trial was intended to be a comparative analysis with traditional RNA Seq. In the absence of the RNA Seq data the extent of the evaluation has been limited. The eRAD trial alone has provided some insight into the appropriate use of the technique, although it is difficult to identify where

failings are due to the samples or the method. Following preparation of the raw data and the removal of PCR clones, over 50% of reads were discarded per individual and some individuals over 80%. Coupled with the sample sequencing bias of unknown cause known to occur in RADSeq, the benefits of increased sampling depth per individual were variable. The high variability in sample sequencing representation may be one of the principle sources of the high variation observed in the count data.

The DE analysis failed to identify with confidence DE tags between chiral genotypes. This may reflect a true lack of DE between genotypes, however due to the high levels of variability within the count data this remains inconclusive. The lack of DE does reveal that sinistrals are not exhibiting a large-scale loss of function as may have been predicted from the adverse consequences of the sinistrality observed in developing *L. stagnalis*. Although the DE analysis was not highly informative, the eRAD sequencing data has enabled successful *de novo* assembly of transcriptomic contigs >700 bases in length. The contigs have been verified through further analyses, including the primer design in Chapter 3.

In the years that have passed since this project was initiated, the capabilities of *de novo* RNA Seq have improved as has the output of sequencing platforms. Consequently, the proposed benefits of eRAD to transcriptomic analyses may have diminished. However, no matter how much data can be gained from a sequencing lane, there will always be a benefit of reduced representation approaches enabling more samples to be analysed in less space. Furthermore, as demonstrated by 'RARseq' (Alabady, Rogers et al. 2015), reduced representation transcriptomic approaches have proven effective for genotyping complex genomes.

Chapter 5:
General discussion and conclusions

Biological inferences from observed patterns of gene expression

As introduced previously, the Davison research group finely mapped the chirality locus and subsequently identified seven candidate genes within the region (Liu, Davey et al. 2013, Davison et al. *awaiting publication*). One of the candidate genes, diaphanous related formin, *Ldia2*, was found to contain a frameshift mutation in the sinistral copy of the gene. The quantitative differences of *Ldia2* mRNA observed from the qPCR experiments described here indicate that the sinistral missense *Ldia2* transcript is rendered non-functional and is almost absent in the single-cell embryo.

Only a specific gene knockout experiment will prove that the loss of function of *Ldia2* results in sinistrality in *L. stagnalis*. Because such methods remain out of reach for this system, multiple experiments have been undertaken to prove that *Ldia2* is tightly associated with LR asymmetry determination and indicate beyond reasonable doubt that this association contributes to the observed phenotype.

Loss of function of *Ldia2* in developing *L. stagnalis* embryos has been demonstrated through drug-treatment experiments performed within the Davison group (Davison et al. *awaiting publication*). Application of SMIFH2, a drug that acts on formin FH2 domains, to dextral embryos prior to the third cell cleavage produces a phenotype that mimics that in sinistral development (Figure 3). The drug treatments strongly indicate that formin is directly involved in dominant spiral cleavage patterning and ultimately LR asymmetry determination. SMIFH2 however, will affect any gene bearing an FH2 domain. The lack of quantitative mRNA differences observed in the alternative formin candidate, *Ldia1*, tested in the qPCR experiments described here, coupled with the fine linkage-mapping of the chirality locus have provided valuable support that the loss of *Ldia2* alone results in the sinistral phenotype.

Of the seven candidate genes identified within the chirality locus only four were assessed in the qPCR experiments here (*Ldia1*, *Ldia2*, *Lfat1* and *Lfry*). This was due to the late identification of three additional candidate genes at the locus (June, 2015). All potentially associated genes within the chirality locus must be ruled out as candidates in order to support the phenotypic association of *Ldia2*. The lack of DE between chiral genotypes of any of the alternative proximal candidates in the

qPCR experiments here, has gone part of the way to achieving this. Further quantitative experiments recently performed in the Davison research group have not found evidence of DE in the remaining alternative candidates.

In addition to the experimental evidence supporting the association of *Ldia2* in LR asymmetry determination in *L. stagnalis*, the functional properties of diaphanous formin fit into a biological mechanism of symmetry-breaking. As previously described, diaphanous formin mediates the self-assembly of actin filaments, whilst the FH2 domain remains continuously associated with the elongating barbed-end of growing actin filaments (Kovar 2006 and references therein). It has recently been demonstrated that the actin cytoskeleton can self-organise into chiral patterning at the cellular level, a process which was additionally found to be dependent on formin function (Tee, Shemesh et al. 2015). This observation has provided a proven mechanism for intracellular symmetry-breaking via the formin-mediated actin-cytoskeleton.

The two actin-related proteins included in the qPCR experiments, were the only targets other than those of *Ldia2* to be identified as DE between chiral genotypes. Although the observed pattern of quantitative differences will require further experiments to define, the actin-related proteins were the only targets included in the experiment with a direct functional association with the formin. Therefore, their differential expression may provide insight into the predicted consequences of the non-functional formin in sinistral *L. stagnalis* and support for formin mediated symmetry-breaking.

Further to the suggested biological mechanism to support the association of *Ldia2* in chirality determination in *L. stagnalis*, an evolutionary mechanism has also been proposed. *L. stagnalis* belongs to the monophyletic group Lymnaeoidea, which contains four families, two dextral and two sinistral (Dayrat, Conrad et al. 2011). A recently constructed phylogeny placed *L. stagnalis* as a sister species to *Physa* with *Biomphalaria* as an outgroup, both of which are sinistral coiling snails (Davison et al. *awaiting publication*). Due to the conserved sequence similarity of *Ldia1* with the diaphanous formin sequence present in *Physa acuta* and *Biomphalaria glabrata*, *Ldia2* was predicted to be the result of a gene duplication having occurred in an ancestral *Lymnaea*. For these reasons it has been proposed that expression of *Ldia2* results in the dextral morph seen in *L. stagnalis*, and the subsequent loss results in reversion to the predicted ancestral sinistral morph (Davison et al, *awaiting publication*).

Aside from the specific findings contributing to the growing evidence of the association of diaphanous formin with LR asymmetry determination in *L. stagnalis*, for the most part the gene expression analyses presented have not identified significant differences between chiral genotypes.

Although the quantitative capabilities of the eRAD analyses are uncertain, they have demonstrated that the sinistral *L. stagnalis* are not exhibiting a wide-spread loss of gene function, which may have been predicted through the apparent pleiotropic effects of sinistrality in development, such as low hatch rate (Davison, Barton et al. 2009). Therefore, it may be that the negative consequences of sinistrality observed in development result from the single chirality locus. It is proposed that the expression of *Ldia1* in the developing sinistral embryo is providing a compensatory mechanism in the absence of the diaphanous-related formin, *Ldia2*. This provides an expanlation for why the knock-out of *Ldia2* is not fatal in all of the developing sinistral *L. stagnalis*, whilst some form of lag or inefficiency in this compensatory process, results in the negative consequences exhibited in some.

Further analyses

Expression patterns in later developmental stages

Possibly the most informative extension of this study would test for DE in later stage embryos. The lack of DE seen in the majority of GOIs in the single-cell embryo qPCR experiments may reflect insufficient developmental time to reveal consequences of the chiral genotype. The single-cell embryo was the most informative for identifying causal relationships and to support the role of *Ldia2* in LR asymmetry determination. However, many expression patterns resulting from the lack of *Ldia2*, or another mechanism, may only be apparent in later embryonic stages, especially around the definitive third-cell cleavage.

For example, the expression patterns observed in the actin-related proteins likely result from the varying levels of *Ldia2* present, due to the role of diaphanous formin in the self-regulation of actin assembly (Kovar 2006). Quantification experiments at later developmental stages would indicate whether this effect is amplified during the predicted extended time in the absence of diaphanous formin.

The assumption that the missense transcript of *Ldia2* is reduced due to NMD would predict a greater reduction in the transcript with increased time since egg-laying. This could be supported by quantitative assays of *Ldia2* present in sinistral embryos at successive cell cleavage steps and following the onset of zygotic transcription when it is predicted that new copies of *Ldia2* will be transcribed.

Localisation of gene expression

Due to the fundamental spatial component within the problem of LR asymmetry determination, localisation experiments will be essential to ascribe functional consequences of differential gene

expression. Although the *in situ* experiments I performed whilst at the Jackson laboratory proved inconclusive (a summary report is presented in the SI: S1), more recent experiments performed by Dr Daniel Jackson have shown asymmetric distribution of *Ldia2* and other early developmental transcripts from the two-cell stage (*data in preparation*). These findings are compelling indicators that the establishment of LR asymmetry occurs very early in development. Furthermore, the accessibility of whole-mount *in situ* hybridisation in early stage *L. stagnalis* embryos is improving thanks to ongoing optimisation within the Jackson group. Accordingly, the DE transcripts identified here, such as the actin-related proteins may be later supported with localisation analysis.

Validation of high-throughput differential expression analysis

Although the few DE expressed tags identified in the eRAD analysis appear unlikely to have resulted from true biological patterns of gene expression, the associated protein functions do hold potential to be involved in LR patterning in development. Therefore, the eRAD transcripts will be investigated further. Most of the DE tags generated a pe contig and subsequently tag-specific qPCR assays can be developed with relative ease.

The RNA Seq data has now been received (October 2015) facilitating the completion of the comparative analysis with eRAD. In addition to performing DE analyses, which may confirm or negate the conclusions regarding overall lack of DE between genotypes, the transcriptome assembly will be of interest. Once assembled, comparisons can be made regarding the quality of the contigs generated from each technique, such as average contig length and extent of annotation. These comparisons will hopefully indicate whether the increased depth of sequencing proposed by eRAD resulted in a higher quality *de novo* assembly.

Possibilities for gene-knockout

As previously stated, to prove *Ldia2* is the causal chirality gene would require gene-knockout. Due to the maternal effect of chirality determining factor in *L. stagnalis*, RNA interference mechanisms such as morpholinos will not be effective as the transcripts, and likely the gene products, are already present in the single cell embryo. The emerging capabilities of CRISPR-cas9 techniques have offered genome editing as a possibility in non-model organisms such as *L. stagnalis*. As the modified transcripts resulting from the CRISPR-cas method are irreversible, the ideal delivery would be to apply the CRISPR-cas modified transcript (cas-crRNA) to the 'mother' snail and subsequently observe the resulting offspring. However, delivery of the cas-crRNA to the reproductive organs remains a challenge due to their internal organisation within the external shell. Consequently, these methods will require a substantial investment to perform, beyond the current scope of the Davison research group, however the results would be highly informative.

Protein analyses

Throughout this thesis, gene expression has been inferred through the relative quantities of specific mRNA transcripts. Subsequently, differences in the relative quantity of mRNA between conditions have been used to indicate patterns of transcriptional regulation and link these to the observed phenotype. Whilst expression level analyses are highly informative, it is important to acknowledge that the quantity and structure of the resulting protein will hold greater functional consequences in relation to the phenotype. As previously described, it is possible that a missense mRNA transcript does not result in a reduction in the quantity of the transcript but produces a non-functional protein, detectable only via protein analysis. Alternatively, it possible that when the level of transcript available is reduced, translation will simply be increased in order to compensate. Consequently, it could be that a significant difference in the quantity of mRNA transcript does not result in a difference in the level of protein. Additional protein analyses would be ideal to support the findings here, especially with regards to the missense sinistral *Ldia2* transcript.

Candidate gene approaches vs whole transcriptome sequencing

The DE analyses within this project were performed over two different scales. The qPCR study assayed specific candidate genes under more controlled settings, whereas the eRAD study performed an explorative sweep of the (reduced) transcriptome. With the exception of the proximal candidates identified through previous linkage-mapping, the GOIs assessed through qPCR could be considered to represent a candidate gene approach to identifying associated genes. Each of the targets was chosen for their predicted function in LR asymmetry determination. Inevitably however, this can lead to 'closed-minded' experimental designs and ultimately wasted time on a 'wild-goose chase' to identify functional gene associations.

Alternatively, with the increasing availability of NGS technologies, the whole transcriptome can be scanned to identify DE patterns in thousands of targets simultaneously. However, such methods, as exemplified in the present study, can be riddled with uncertainty due to inherent errors such as unknown sequencing error, biased sequencing depth between transcripts and conditions, and perhaps the most challenging: false discovery rate due to multiple comparisons. Owing to these factors, the majority of DE patterns identified through high-throughput sequencing experiments require validification via alternative methods such as qPCR or microarrays, perhaps diminishing the high-throughput capabilities of the technique.

Of course candidate gene approaches can only be performed with prior knowledge of both the candidate gene's sequence and function. Accordingly, NGS methods can provide a valuable initial

investment to gain exploratory sequence information and an overall representation of expression patterns within the study system, simply not possible through previously available methods. Moreover, the necessity for validification of findings through multiple experimental methods does not negate either method, but increases support of both. In this study, the candidate gene approach was successful in highlighting DE in the actin-related proteins, which are directly related to the primary candidate gene, and a lack of DE within the other genes. However, to infer that the sinistral genotype is not demonstrating a wide-spread loss of function from a lack of DE in ten target genes would demand further evidence. Support for this interpretation was indicated through the lack of DE identified across >30,000 sequence tags in the eRAD data.

Applications of findings

This project has described five endogenous control genes confidently verified as stable in the single-cell embryo, ovotestis and foot tissue of chirally dimorphic *L. stagnalis*. The specified targets can provide other researchers of *L. stagnalis* rapid access to endogenous control genes suitable for relative qPCR in keeping with the MIQE guidelines (Bustin, Benes et al. 2009) which were previously lacking.

The trial of eRAD although awaiting comparison with RNA Seq, will provide an informative overview of the capabilities of this technique. The recently published paper on 'RARseq' (Alabady, Rogers et al. 2015) as previously stated, has not employed the quantitative capabilities of the technique, and as there are currently no publications of eRAD from the Cresko group, DE analysis via reduced representation RNA sequencing remains undescribed in the scientific literature. Finally, since the publication of the 'RARseq' paper it is arguable that the presently used term 'eRAD' should be changed to adhere to the term ascribed in the first publication of data analysis. However, eRAD has been used here in keeping with the term used by the Cresko group following the conference in which it was described in 2010 (National Center for Research Resources 2010).

References

Abu Shah, E. and K. Keren (2014). "Symmetry breaking in reconstituted actin cortices." Elife **3**: e01433

Adams, D. S., K. R. Robinson, T. Fukumoto, S. P. Yuan, R. C. Albertson, P. Yelick, L. Kuo, M. McSweeney and M. Levin (2006). "Early, H+-V-ATPase-dependent proton flux is necessary for consistent left-right patterning of non-mammalian vertebrates." Development **133**(9): 1657-1671.

Afzelius, B. A. (1976). "HUMAN SYNDROME CAUSED BY IMMOTILE CILIA." Science **193**(4250): 317-319.

Alabady, M. S., W. L. Rogers and R. L. Malmberg (2015). "Development of Transcriptomic Markers for Population Analysis Using Restriction Site Associated RNA Sequencing (RARseq)." Plos One **10**(8).

Anders, S. and W. Huber (2010). "Differential expression analysis for sequence count data." Genome Biology **11**(10).

Andersen, C. L., J. L. Jensen and T. F. Orntoft (2004). "Normalization of real-time quantitative reverse transcription-PCR data: A model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets." Cancer Research **64**(15): 5245-5250.

Andolfatto, P., D. Davison, D. Erezyilmaz, T. T. Hu, J. Mast, T. Sunayama-Morita and D. L. Stern (2011). "Multiplexed shotgun genotyping for rapid and efficient genetic mapping." Genome Research **21**(4): 610-617.

Andrews, S. (2010). "FastQC A Quality Control tool for High Throughput Sequence Data." http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

AppliedBiosystems (2008). "Guide to Performing Relative Quantitation of Gene Expression Using Real-Time Quantitative PCR."

AppliedBiosystems (2008). "Poly(A) Purist Kit Manual, Part Number AM.1916." P/N1916 Revision B.

AppliedBiosystems (2010). "TRI Reagent Solution Protocol."

AppliedBiosystems (2011). "Real time PCR: Understanding Ct." Application Note.

AppliedBiosystems (2014). Real-time PCR handbook.

Asami, T. (1993). "Genetic variation and evolution of coiling chirality in snails." Forma **8**: 263-276.

Asami, T., E. Gittenberger and G. Falkner (2008). "Whole-body enantiomorphy and maternal inheritance of chiral reversal in the pond snail Lymnaea stagnalis." Journal of Heredity **99**(5): 552-557.

Ashton, P. M., S. Nair, T. Dallman, S. Rubino, W. Rabsch, S. Mwaigwisya, J. Wain and J. O'Grady (2015). "MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island." Nature Biotechnology **33**(3): 296-+.

Aviv, H. and P. Leder (1972). "Purification of Biologically-Active Globin Messenger-Rna by Chromatography on Oligothymidylic-Acid-Cellulose." Proceedings of the National Academy of Sciences of the United States of America **69**(6): 1408-&.

Aw, S. and M. Levin (2008). "What's Left in Asymmetry?" Developmental Dynamics **237**(12): 3453-3463.

Ayscough, K. (1998). “Use of latrunculin-A, an actin monomer-binding drug”. Methods in Enzymology. **298**: 18-25.

Azami, S., A. Wagatsuma, H. Sadamoto, D. Hatakeyama, T. Usami, M. Fujie, R. Koyanagi, K. Azumi, Y. Fujito, K. Lukowiak and E. Ito (2006). "Altered gene activity correlated with long-term memory formation of conditioned taste aversion in Lymnaea." Journal of Neuroscience Research **84**(7): 1610-1620.

Azimzadeh, J., M. L. Wong, D. M. Downhour, A. S. Alvarado and W. F. Marshall (2012). "Centrosome Loss in the Evolution of Planarians." Science **335**(6067): 461-463.

Babcock, L. E. (2005). "Asymmetry in the fossil record." European Review **13**(2): 135-143.

Badano, J. L., N. Mitsuma, P. L. Beales and N. Katsanis (2006). The ciliopathies: An emerging class of human genetic disorders. Annual Review of Genomics and Human Genetics. **7:** 125-148.

Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, Z. A. Lewis, E. U. Selker, W. A. Cresko and E. A. Johnson (2008). "Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers." Plos One **3**(10).

Baker, M. (2011). "qPCR: quicker and easier but don't be sloppy." Nature Methods **8**(3): 207-212.

Baroux, C., D. Autran, C. S. Gillmor, D. Grimanelli and U. Grossniklaus (2008). The Maternal to Zygotic Transition in Animals and Plants. Control and Regulation of Stem Cells. B. Stillman, S. Stewart and T. Grodzicker. **73:** 89-100.

Barranco, P., J. Cabrero, J. P. M. Camacho and F. Pascual (1995). "CHROMOSOMAL BASIS FOR A BILATERAL GYNANDROMORPH IN PYCNOGASTER-INERMIS (RAMBUR, 1838) (ORTHOPTERA, TETTIGONIIDAE)." Contributions to Zoology **65**(2): 123-127.

Barski, A., S. Cuddapah, K. Cui, T.-Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev and K. Zhao (2007). "High-resolution profiling of histone methylations in the human genome." Cell **129**(4): 823-837.

Baum, B. (2006). "Left-right asymmetry: Actin-myosin through the looking glass." Current Biology **16**(13): R502-R504.

Bavan, S., V. A. Straub, T. E. Webb and S. J. Ennion (2012). "Cloning and Characterization of a P2X Receptor Expressed in the Central Nervous System of Lymnaea stagnalis." Plos One **7**(11).

Benjamini, Y., D. Drai, G. Elmer, N. Kafkafi and I. Golani (2001). "Controlling the false discovery rate in behavior genetics research." Behavioural Brain Research **125**(1-2): 279-284.

Benjamini, Y. and Y. Hochberg (1995). "CONTROLLING THE FALSE DISCOVERY RATE - A PRACTICAL AND POWERFUL APPROACH TO MULTIPLE TESTING." Journal of the Royal Statistical Society Series B-Methodological **57**(1): 289-300.

Bentley, D. R., S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. K. Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. H. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. J. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. L. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. D. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. A. Baybayan, V. A. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J.

A. Bridgham, R. C. Brown, A. A. Brown, D. H. Buermann, A. A. Bundu, J. C. Burrows, N. P. Carter, N. Castillo, M. C. E. Catenazzi, S. Chang, R. N. Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V. F. Fajardo, W. S. Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. A. Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish, C. D. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, K. Hoschler, S. Hurwitz, D. V. Ivanov, M. Q. Johnson, T. James, T. A. H. Jones, G. D. Kang, T. H. Kerelska, A. D. Kersey, I. Khrebtukova, A. P. Kindwall, Z. Kingsbury, P. I. Kokko-Gonzales, A. Kumar, M. A. Laurent, C. T. Lawley, S. E. Lee, X. Lee, A. K. Liao, J. A. Loch, M. Lok, S. J. Luo, R. M. Mammen, J. W. Martin, P. G. McCauley, P. McNitt, P. Mehta, K. W. Moon, J. W. Mullens, T. Newington, Z. M. Ning, B. L. Ng, S. M. Novo, M. J. O'Neill, M. A. Osborne, A. Osnowski, O. Ostadan, L. L. Paraschos, L. Pickering, A. C. Pike, A. C. Pike, D. C. Pinkard, D. P. Pliskin, J. Podhasky, V. J. Quijano, C. Raczy, V. H. Rae, S. R. Rawlings, A. C. Rodriguez, P. M. Roe, J. Rogers, M. C. R. Bacigalupo, N. Romanov, A. Romieu, R. K. Roth, N. J. Rourke, S. T. Ruediger, E. Rusman, R. M. Sanches-Kuiper, M. R. Schenker, J. M. Seoane, R. J. Shaw, M. K. Shiver, S. W. Short, N. L. Sizto, J. P. Sluis, M. A. Smith, J. E. S. Sohna, E. J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C. L. Tregidgo, G. Turcatti, S. vandeVondele, Y. Verhovsky, S. M. Virk, S. Wakelin, G. C. Walcott, J. W. Wang, G. J. Worsley, J. Y. Yan, L. Yau, M. Zuerlein, J. Rogers, J. C. Mullikin, M. E. Hurles, N. J. McCooke, J. S. West, F. L. Oaks, P. L. Lundberg, D. Klenerman, R. Durbin and A. J. Smith (2008). "Accurate whole human genome sequencing using reversible terminator chemistry." Nature **456**(7218): 53-59.

Bernard, M., H. Yoshioka, E. Rodriguez, M. Vanderrest, T. Kimura, Y. Ninomiya, B. R. Olsen and F. Ramirez (1988). "CLONING AND SEQUENCING OF PRO-ALPHA-1 (XI) COLLAGEN CDNA DEMONSTRATES THAT TYPE-XI BELONGS TO THE FIBRILLAR CLASS OF COLLAGENS AND REVEALS THAT THE EXPRESSION OF THE GENE IS NOT RESTRICTED TO CARTILAGENOUS TISSUE." Journal of Biological Chemistry **263**(32): 17159-17166.

Beyer, T., M. Danilchik, T. Thumberger, P. Vick, M. Tisler, I. Schneider, S. Bogusch, P. Andre, B. Ulmer, P. Walentek, B. Niesler, M. Blum and A. Schweickert (2012). "Serotonin Signaling Is Required for Wnt-Dependent GRP Specification and Leftward Flow in Xenopus." Current Biology **22**(1): 33-39.

Blencowe, B. J., S. Ahmad and L. J. Lee (2009). "Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes." Genes & Development **23**(12): 1379-1386.

Bornens, M. (2012). "The Centrosome in Cells and Organisms." Science **335**(6067): 422-426.

Bottke, W. (1986). "IMMUNO-LOCALIZATION OF FERRITIN POLYPEPTIDES IN OOCYTES AND SOMATIC TISSUE OF THE FRESH-WATER SNAILS LYMNAEA-STAGNALIS L AND PLANORBARIUS-CORNEUS L." Cell and Tissue Research **243**(2): 397-404.

Bottke, W. and R. R. Crichton (1984). "VITELLOGENIC FERRITIN OF LYMNAEA-STAGNALIS L (MOLLUSCA, GASTROPODA) DIFFERS IN STRUCTURE FROM SOMA CELL TYPE FERRITIN." Comparative Biochemistry and Physiology B-Biochemistry & Molecular Biology **77**(1): 57-61.

Bouck, A. and T. Vision (2007). "The molecular ecologist's guide to expressed sequence tags." Molecular Ecology **16**(5): 907-924.

Bouetard, A., A. L. Besnard, D. Vassaux, L. Lagadic and M. A. Coutellec (2013). "Impact of the redox-cycling herbicide diquat on transcript expression and antioxidant enzymatic activities of the freshwater snail Lymnaea stagnalis." Aquatic Toxicology **126**: 256-265.

Bouhaddioui, W., P. R. Provost and Y. Tremblay (2014). "Identification of Most Stable Endogenous Control Genes for MicroRNA Quantification in the Developing Mouse Lung." Plos One **9**(11).

Boycott, A. E. and C. Diver (1923). "On the inheritance of sinistrality in Limnaea peregra." Proceedings of the Royal Society of London Series B-Containing Papers of a Biological Character **95**(666): 207-213.

Boycott, A. E., C. Diver, S. L. Garstang, A. C. Hardy and F. M. Turner (1930). "The inheritance of sinistrality in Limnaea peregra (Mollusca, Pulmonata)." Philosophical Transactions of the Royal Society of London Series B-Containing Papers of a Biological Character **219**: 51-131.

Brown, N. A. and L. Wolpert (1990). "THE DEVELOPMENT OF HANDEDNESS IN LEFT RIGHT ASYMMETRY." Development **109**(1): 1-9.

Bullard, J. H., E. Purdom, K. D. Hansen and S. Dudoit (2010). "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments." Bmc Bioinformatics **11**.

Bulloch, A. G. M., C. Q. Diep, C. C. Logan, E. S. Bulloch, S. M. Robbins, J. Hislop and W. S. Sossin (2005). "Ltrk is differentially expressed in developing and adult neurons of the Lymnaea central nervous system." Journal of Comparative Neurology **487**(3): 240-254.

Burn, J. (1991). “Disturbance of Morphological Laterality in Humans”. Biological asymmetry and handedness. Wiley, Chichester (Ciba Foundation symposium 162). p282-299.

Bustin, S. A., V. Benes, J. Garson, J. Hellemans, J. Huggett, M. Kubista, R. Mueller, T. Nolan, M. W. Pfaffl, G. Shipley, C. T. Wittwer, P. Schjerling, P. J. Day, M. Abreu, B. Aguado, J. F. Beaulieu, A. Beckers, S. Bogaert, J. A. Browne, F. Carrasco-Ramiro, L. Ceelen, K. Ciborowski, P. Cornillie, S. Coulon, A. Cuypers, S. De Brouwer, L. De Ceuninck, J. De Craene, H. De Naeyer, W. De Spiegelaere, K. Deckers, A. Dheedene, K. Durinck, M. Ferreira-Teixeira, A. Fieuw, J. M. Gallup, S. Gonzalo-Flores, K. Goossens, F. Heindryckx, E. Herring, H. Hoenicka, L. Icardi, R. Jaggi, F. Javad, M. Karampelias, F. Kibenge, M. Kibenge, C. Kumps, I. Lambertz, T. Lammens, A. Markey, P. Messiaen, E. Mets, S. Morais, A. Mudarra-Rubio, J. Nakiwala, H. Nelis, P. A. Olsvik, C. Perez-Novo, M. Plusquin, T. Remans, A. Rihani, P. Rodrigues-Santos, P. Rondou, R. Sanders, K. Schmidt-Bleek, K. Skovgaard, K. Smeets, L. Tabera, S. Toegel, T. Van Acker, W. Van den Broeck, J. Van der Meulen, M. Van Gele, G. Van Peer, M. Van Poucke, N. Van Roy, S. Vergult, J. Wauman, M. Tshuikina-Wiklander, E. Willems, S. Zaccara, F. Zeka and J. Vandesompele (2013). "The need for transparency and good practices in the qPCR literature." Nature Methods **10**(11): 1063-1067.

Bustin, S. A., V. Benes, J. A. Garson, J. Hellemans, J. Huggett, M. Kubista, R. Mueller, T. Nolan, M. W. Pfaffl, G. L. Shipley, J. Vandesompele and C. T. Wittwer (2009). "The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments." Clinical Chemistry **55**(4): 611-622.

Calarcogillam, P. D., M. C. Siebert, R. Hubble, T. Mitchison and M. Kirschner (1983). "CENTROSOME DEVELOPMENT IN EARLY MOUSE EMBRYOS AS DEFINED BY AN AUTOANTIBODY AGAINST PERICENTRIOLAR MATERIAL." Cell **35**(3): 621-629.

Carneiro, K., C. Donnet, T. Rejtar, B. L. Karger, G. A. Barisone, E. Diaz, S. Kortagere, J. M. Lemire and M. Levin (2011). "Histone deacetylase activity is necessary for left-right patterning during vertebrate development." Bmc Developmental Biology **11**.

Carninci, P., T. Kasukawa, S. Katayama, J. Gough, M. C. Frith, N. Maeda, R. Oyama, T. Ravasi, B. Lenhard, C. Wells, R. Kodzius, K. Shimokawa, V. B. Bajic, S. E. Brenner, S. Batalov, A. R. R. Forrest, M. Zavolan, M. J. Davis, L. G. Wilming, V. Aidinis, J. E. Allen, X. Ambesi-Impiombato, R. Apweiler, R. N. Aturaliya, T. L. Bailey, M. Bansal, L. Baxter, K. W. Beisel, T. Bersano, H. Bono, A. M. Chalk, K. P. Chiu, V. Choudhary, A. Christoffels, D. R. Clutterbuck, M. L. Crowe, E. Dalla, B. P. Dalrymple, B. de Bono, G. Della Gatta, D. di Bernardo, T. Down, P. Engstrom, M. Fagiolini, G. Faulkner, C. F. Fletcher, T.

Fukushima, M. Furuno, S. Futaki, M. Gariboldi, P. Georgii-Hemming, T. R. Gingeras, T. Gojobori, R. E. Green, S. Gustincich, M. Harbers, Y. Hayashi, T. K. Hensch, N. Hirokawa, D. Hill, L. Huminiecki, M. Iacono, K. Ikeo, A. Iwama, T. Ishikawa, M. Jakt, A. Kanapin, M. Katoh, Y. Kawasawa, J. Kelso, H. Kitamura, H. Kitano, G. Kollias, S. P. T. Krishnan, A. Kruger, S. K. Kummerfeld, I. V. Kurochkin, L. F. Lareau, D. Lazarevic, L. Lipovich, J. Liu, S. Liuni, S. McWilliam, M. M. Babu, M. Madera, L. Marchionni, H. Matsuda, S. Matsuzawa, H. Miki, F. Mignone, S. Miyake, K. Morris, S. Mottagui-Tabar, N. Mulder, N. Nakano, H. Nakauchi, P. Ng, R. Nilsson, S. Nishiguchi, S. Nishikawa, F. Nori, O. Ohara, Y. Okazaki, V. Orlando, K. C. Pang, W. J. Pavan, G. Pavesi, G. Pesole, N. Petrovsky, S. Piazza, J. Reed, J. F. Reid, B. Z. Ring, M. Ringwald, B. Rost, Y. Ruan, S. L. Salzberg, A. Sandelin, C. Schneider, C. Schonbach, K. Sekiguchi, C. A. M. Semple, S. Seno, L. Sessa, Y. Sheng, Y. Shibata, H. Shimada, K. Shimada, D. Silva, B. Sinclair, S. Sperling, E. Stupka, K. Sugiura, R. Sultana, Y. Takenaka, K. Taki, K. Tammoja, S. L. Tan, S. Tang, M. S. Taylor, J. Tegner, S. A. Teichmann, H. R. Ueda, E. van Nimwegen, R. Verardo, C. L. Wei, K. Yagi, H. Yamanishi, E. Zabarovsky, S. Zhu, A. Zimmer, W. Hide, C. Bult, S. M. Grimmond, R. D. Teasdale, E. T. Liu, V. Brusic, J. Quackenbush, C. Wahlestedt, J. S. Mattick, D. A. Hume, C. Kai, D. Sasaki, Y. Tomaru, S. Fukuda, M. Kanamori-Katayama, M. Suzuki, J. Aoki, T. Arakawa, J. Iida, K. Imamura, M. Itoh, T. Kato, H. Kawaji, N. Kawagashira, T. Kawashima, M. Kojima, S. Kondo, H. Konno, K. Nakano, N. Ninomiya, T. Nishio, M. Okada, C. Plessy, K. Shibata, T. Shiraki, S. Suzuki, M. Tagami, K. Waki, A. Watahiki, Y. Okamura-Oho, H. Suzuki, J. Kawai, Y. Hayashizaki, F. Consortium and R. G. E. R. G. S (2005). "The transcriptional landscape of the mammalian genome." Science **309**(5740): 1559-1563.

Carter, C. J., C. Rand, I. Mohammad, A. Lepp, N. Vesprini, O. Wiebe, R. Carlone and G. E. Spencer (2015). "Expression of a Retinoic Acid Receptor (RAR)-Like Protein in the Embryonic and Adult Nervous System of a Protostome Species." Journal of Experimental Zoology Part B-Molecular and Developmental Evolution **324**(1): 51-67.

Casey, B. and B. P. Hackett (2000). "Left-right axis malformations in man and mouse." Current Opinion in Genetics & Development **10**(3): 257-261.

Catchen, J., P. A. Hohenlohe, S. Bassham, A. Amores and W. A. Cresko (2013). "Stacks: an analysis tool set for population genomics." Molecular Ecology **22**(11): 3124-3140.

Catchen, J. M., A. Amores, P. Hohenlohe, W. Cresko and J. H. Postlethwait (2011). "Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences." G3-Genes Genomes Genetics **1**(3): 171-182.

Chang, W. (2013). R Graphics Cookbook. California, US, O'Reilly Media, Inc.

Chen, T. H., J. J. Hsu, X. Zhao, C. Y. Guo, M. N. Wong, Y. Huang, Z. W. Li, A. Garfinkel, C. M. Ho, Y. Tintut and L. L. Demer (2012). "Left-Right Symmetry Breaking in Tissue Morphogenesis via Cytoskeletal Mechanics." Circulation Research **110**(4): 551-U117.

Cheney, R. E., M. K. Oshea, J. E. Heuser, M. V. Coelho, J. S. Wolenski, E. M. Espreafico, P. Forscher, R. E. Larson and M. S. Mooseker (1993). "BRAIN MYOSIN-V IS A 2-HEADED UNCONVENTIONAL MYOSIN WITH MOTOR-ACTIVITY." Cell **75**(1): 13-23.

Choi, M., U. I. Scholl, W. Ji, T. Liu, I. R. Tikhonova, P. Zumbo, A. Nayir, A. Bakkaloglu, S. Ozen, S. Sanjad, C. Nelson-Williams, A. Farhi, S. Mane and R. P. Lifton (2009). "Genetic diagnosis by whole exome capture and massively parallel DNA sequencing." Proceedings of the National Academy of Sciences of the United States of America **106**(45): 19096-19101.

Chua, G., M. D. Robinson, Q. Morris and T. R. Hughes (2004). "Transcriptional networks: reverse-engineering gene regulation on a global scale." Current Opinion in Microbiology **7**(6): 638-646.

Cokus, S. J., S. Feng, X. Zhang, Z. Chen, B. Merriman, C. D. Haudenschild, S. Pradhan, S. F. Nelson, M. Pellegrini and S. E. Jacobsen (2008). "Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning." Nature **452**(7184): 215-219.

Conesa, A., S. Gotz, J. M. Garcia-Gomez, J. Terol, M. Talon and M. Robles (2005). "Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research." Bioinformatics **21**(18): 3674-3676.

Cong, J. L., W. Geng, B. He, J. C. Liu, J. Charlton and P. N. Adler (2001). "The furry gene of Drosophila is important for maintaining the integrity of cellular extensions during morphogenesis." Development **128**(14): 2793-2802.

Cong, L., F. A. Ran, D. Cox, S. L. Lin, R. Barretto, N. Habib, P. D. Hsu, X. B. Wu, W. Y. Jiang, L. A. Marraffini and F. Zhang (2013). "Multiplex Genome Engineering Using CRISPR/Cas Systems." Science **339**(6121): 819-823.

Conti, E. and E. Izaurralde (2005). "Nonsense-mediated mRNA decay: molecular insights and mechanistic variations across species." Current Opinion in Cell Biology **17**(3): 316-325.

Corballis, M. C. and Morgan, M. J. (1978). On the biological basis of human laterality: I. Evidence for a maturational left–right gradient. Behavioral and Brain Sciences **1**: 261-269.

Crampton, H. E. (1894). "Reversal of cleavage in a sinistral gastropod." Annals of the New York Academy of Sciences **8**: 167-170.

De Brabander, M., Geuens, G., Nuydens, R., Willebrords, R., Aerts, F. and De Mey, J. (1986). “Microtubule dynamics during the cell cycle: the effects of taxol and nocodazole on the microtubule system of Pt K2 cells at different stages of the mitotic cycle”. International Review of Cytology. **101**: 215-274.

Davey, J. W., T. Cezard, P. Fuentes-Utrilla, C. Eland, K. Gharbi and M. L. Blaxter (2013). "Special features of RAD Sequencing data: implications for genotyping." Molecular Ecology **22**(11): 3151-3164.

Davison, A., N. H. Barton and B. Clarke (2009). "The effect of coil phenotypes and genotypes on the fecundity and viability of Partula suturalis and Lymnaea stagnalis: implications for the evolution of sinistral snails." Journal of Evolutionary Biology **22**(8): 1624-1635.

Davison, A., S. Chiba, N. H. Barton and B. Clarke (2005). "Speciation and gene flow between snails of opposite chirality." Plos Biology **3**(9): 1559-1571.

Davison, A., H. T. Frend, C. Moray, H. Wheatley, L. J. Searle and M. P. Eichhorn (2009). "Mating behaviour in Lymnaea stagnalis pond snails is a maternally inherited, lateralized trait." Biology Letters **5**(1): 20-22.

Dayrat, B., M. Conrad, S. Balayan, T. R. White, C. Albrecht, R. Golding, S. R. Gomes, M. G. Harasewych and A. M. de Frias Martins (2011). "Phylogenetic relationships and evolution of pulmonate gastropods (Mollusca): New insights from increased taxon sampling." Molecular Phylogenetics and Evolution **59**(2): 425-437.

de la Cruz, I. P., J. Z. Levin, C. Cummins, P. Anderson and H. R. Horvitz (2003). "sup-9, sup-10, and unc-93 may encode components of a two-pore K+ channel that coordinates muscle contraction in Caenorhabditis elegans." Journal of Neuroscience **23**(27): 9133-9145.

DeFranscesco, L. (1998). "Ode to Oligo(dT): Oligo(dT) Takes on a Variety of Faces in Kits for the Purification of mRNA." www.the-scientist.com Article No. 18946 1998.

Demuth, J. P. and M. J. Wade (2006). Experimental methods for measuring gene interactions. Annual Review of Ecology Evolution and Systematics. **37:** 289-316.

Dippold, H. C., M. M. Ng, S. E. Farber-Katz, S. K. Lee, M. L. Kerr, M. C. Peterman, R. Sim, P. A. Wiharto, K. A. Galbraith, S. Madhavarapu, G. J. Fuchs, T. Meerloo, M. G. Farquhar, H. L. Zhou and S. J. Field (2009). "GOLPH3 Bridges Phosphatidylinositol-4-Phosphate and Actomyosin to Stretch and Shape the Golgi to Promote Budding." Cell **139**(2): 337-351.

Downey, N. (2014). "Interpreting Melt Curves: An Indicator, Not a Diagnosis." Decoded articles: Core Concepts, Scientific Fundamentals Explained.

Duboc, V., E. Rottinger, F. Lapraz, L. Besnardeau and T. Lepage (2005). "Left-right asymmetry in the sea urchin embryo is regulated by nodal signaling on the right side." Developmental Cell **9**(1): 147-158.

Eckert, R. L., M. B. Yaffe, J. F. Crish, S. Murthy, E. A. Rorke and J. F. Welter (1993). "INVOLUCRIN - STRUCTURE AND ROLE IN ENVELOPE ASSEMBLY." Journal of Investigative Dermatology **100**(5): 613-617.

Eid, J., A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. Dewinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach and S. Turner (2009). "Real-Time DNA Sequencing from Single Polymerase Molecules." Science **323**(5910): 133-138.

Eisenberg, E. and E. Y. Levanon (2013). "Human housekeeping genes, revisited." Trends in Genetics **29**(10): 569-574.

Eisenstein, M. (2012). "Oxford Nanopore announcement sets sequencing sector abuzz." Nature Biotechnology **30**(4): 295-296.

Ekblom, R. and J. Galindo (2011). "Applications of next generation sequencing in molecular ecology of non-model organisms." Heredity **107**(1): 1-15.

Elliot, D. and M. Ladomery (2011). Molecular Biology of RNA., Oxford University Press Inc. New York.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler and S. E. Mitchell (2011). "A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species." Plos One **6**(5).

Emerson, K. J., C. R. Merz, J. M. Catchen, P. A. Hohenlohe, W. A. Cresko, W. E. Bradshaw and C. M. Holzapfel (2010). "Resolving postglacial phylogeography using high-throughput sequencing." Proceedings of the National Academy of Sciences of the United States of America **107**(37): 16196-16200.

Engel, M. H. and S. A. Macko (1997). "Isotopic evidence for extraterrestrial non-racemic amino acids in the Murchison meteorite." Nature **389**(6648): 265-268.

Essner, J. J., K. J. Vogan, M. K. Wagner, C. J. Tabin, H. J. Yost and M. Brueckner (2002). "Conserved function for embryonic nodal cilia." Nature **418**(6893): 37-38.

Etter, P. D., S. Bassham, P. A. Hohenlohe, E. A. Johnson and W. A. Cresko (2011). SNP Discovery and Genotyping for Evolutionary Genetics Using RAD Sequencing. Molecular Methods for Evolutionary Genetics. V. Orgogozo and M. V. Rockman. **772:** 157-178.

Etter, P. D., J. L. Preston, S. Bassham, W. A. Cresko and E. A. Johnson (2011). "Local De Novo Assembly of RAD Paired-End Contigs Using Short Sequencing Reads." Plos One **6**(4).

Evans, T., C. M. Wade, F. A. Chapman, A. D. Johnson and M. Loose (2014). "Acquisition of Germ Plasm Accelerates Vertebrate Evolution." Science **344**(6180): 200-203.

Fischer, E. (1894). "Sur la symétrie dans les phénomènes physique." J. Physique **3**(111): 393-416.

Fonseca, V. G., G. R. Carvalho, W. Sung, H. F. Johnson, D. M. Power, S. P. Neill, M. Packer, M. L. Blaxter, P. J. D. Lambshead, W. K. Thomas and S. Creer (2010). "Second-generation environmental sequencing unmasks marine metazoan biodiversity." Nature Communications **1**.

Freeman, G. and Lundelius, J. W. (1982). The developmental genetics of dextrality and sinistrality in the gastropod *Lymnaea peregra*. Wilhelm Roux's archives of developmental biology, **191**: 69-83.

Friedland, A. E., Y. B. Tzur, K. M. Esvelt, M. P. Colaiacovo, G. M. Church and J. A. Calarco (2013). "Heritable genome editing in C. elegans via a CRISPR-Cas9 system." Nature Methods **10**(8): 741-+.

Furusawa, T., S. Ikawa, N. Yanai and M. Obinata (2000). "Isolation of a novel PDZ-containing myosin from hematopoietic supportive bone marrow stromal cell lines." Biochemical and Biophysical Research Communications **270**(1): 67-75.

Gardner, M. (1990). The new ambidextrous universe: revised edition. New York, W H Freeman.

Gardner, R. L. (2010). "Normal Bias in the Direction of Fetal Rotation Depends on Blastomere Composition during Early Cleavage in the Mouse." Plos One **5**(3).

Garimberti, E. and S. Tosi (2010). Fluorescence in situ Hybridization (FISH), Basic Principles and Methodology. Fluorescence in Situ Hybridization. J. M. Bridger and E. V. Volpi. **659:** 3-20.

Gierlinski, M., C. Cole, P. Schofield, N. J. Schurch, A. Sherstnev, V. Singh, N. Wrobel, K. Gharbi, G. Simpson, T. Owen-Hughes, M. Blaxter and G. J. Barton (2015). "Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment." Bioinformatics (Oxford, England) **31**(22): 3625-3630.

Gittenberger, E. (1988). "Sympatric speciation in snails - a largely neglected model." Evolution **42**(4): 826-828.

Glazier, A. M., J. H. Nadeau and T. J. Aitman (2002). "Finding genes that underlie complex traits." Science **298**(5602): 2345-2349.

Goley, E. D. and M. D. Welch (2006). "The ARP2/3 complex: an actin nucleator comes of age." Nature Reviews Molecular Cell Biology **7**(10): 713-726.

Gournier, H., E. D. Goley, H. Niederstrasser, T. Trinh and M. D. Welch (2001). "Reconstitution of human Arp2/3 complex reveals critical roles of individual subunits in complex structure and activity." Molecular Cell **8**(5): 1041-1052.

Govind, C. K. (1989). "ASYMMETRY IN LOBSTER CLAWS." American Scientist **77**(5): 468-474.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W.

Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman and A. Regev (2011). "Full-length transcriptome assembly from RNA-Seq data without a reference genome." Nature Biotechnology **29**(7): 644-U130.

Graham, C. F., T. C. Glenn, A. G. McArthur, D. R. Boreham, T. Kieran, S. Lance, R. G. Manzon, J. A. Martino, T. Pierson, S. M. Rogers, J. Y. Wilson and C. M. Somers (2015). "Impacts of degraded DNA on restriction enzyme associated DNA sequencing (RADSeq)." Molecular Ecology Resources **15**(6): 1304-1315.

Grande, C., J. M. Martin-Duran, N. J. Kenny, M. Truchado-Garcia and A. Hejnol (2014). "Evolution, divergence and loss of the Nodal signalling pathway: new data and a synthesis across the Bilateria." International Journal of Developmental Biology **58**(6-8): 521-532.

Grande, C. and N. H. Patel (2009). "Nodal signalling is involved in left-right asymmetry in snails." Nature **457**(7232): 1007-1011.

Gray, H. (1918). Anatomy of the Human Body, 20th Edition. Philadelphia. , Lea & Febiger. Image downloaded from Bartleby.com, 2000. www.bartleby.com/107/. 13.11.2015..

Gros, J., K. Feistel, C. Viebahn, M. Blum and C. J. Tabin (2009). "Cell Movements at Hensen's Node Establish Left/Right Asymmetric Gene Expression in the Chick." Science **324**(5929): 941-944.

Grzechnik, P. and J. Kufel (2008). "Polyadenylation Linked to Transcription Termination Directs the Processing of snoRNA Precursors in Yeast." Molecular Cell **32**(2): 247-258.

Guo, S. and K. J. Kemphues (1996). "A non-muscle myosin required for embryonic polarity in Caenorhabditis elegans." Nature **382**(6590): 455-458.

Haas, B. J. and M. C. Zody (2010). "Advancing RNA-Seq analysis." Nature Biotechnology **28**(5): 421-423.

Hagmann, J. (1993). "PATTERN-FORMATION AND HANDEDNESS IN THE CYTOSKELETON OF HUMAN PLATELETS." Proceedings of the National Academy of Sciences of the United States of America **90**(8): 3280-3283.

Halbleib, J. M. and W. J. Nelson (2006). "Cadherins in development: cell adhesion, sorting, and tissue morphogenesis." Genes & Development **20**(23): 3199-3214.

Happle, R., H. Mittag and W. Kuster (1995). "THE CHILD NEVUS - A DISTINCT SKIN DISORDER." Dermatology **191**(3): 210-216.

Harbers, M. and P. Carninci (2005). "Tag-based approaches for transcriptome research and genome annotation." Nature Methods **2**(7): 495-502.

Harper, W. (1991). Kant on Incongruent Counterparts. The Philosophy of Right and Left. J. Van Cleve and R. Frederick, Springer Netherlands. **46:** 263-313.

Hatakeyama, D., H. Sadamoto and E. Ito (2004). "Real-time quantitative RT-PCR method for estimation of mRNA level of CCAAT/enhancer binding protein in the central nervous system of Lymnaea stagnalis." Acta Biologica Hungarica **55**(1-4): 157-161.

Heacock, A. M. and B. W. Agranoff (1977). "CLOCKWISE GROWTH OF NEURITES FROM RETINAL EXPLANTS." Science **198**(4312): 64-66.

Hellemans, J., G. Mortier, A. De Paepe, F. Speleman and J. Vandesompele (2007). "qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data." Genome Biology **8**(2).

Hemmati-brivanlou, A., D. Frank, M. E. Bolce, B. D. Brown, H. L. Sive and R. M. Harland (1990). "Localization of specific mRNAs in *Xenopus* embryos by whole-mount *in situ* hybridization." Development **110**(2): 325-330.

Hendricks, J. R. (2009). "Sinistral snail shells in the sea: developmental causes and consequences." Lethaia **42**(1): 55-66.

Henri, H., M. Cariou, G. Terraz, S. Martinez, A. El Filali, M. Veyssiere, L. Duret and S. Charlat (2015). "Optimization of multiplexed RADseq libraries using low-cost adaptors." Genetica **143**(2): 139-143.

Henry, J. Q. (2014). "Spiralian model systems." International Journal of Developmental Biology **58**: 389-401.

Hibbeler, S., J. P. Scharsack and S. Becker (2008). "Housekeeping genes for quantitative expression studies in the three-spined stickleback Gasterosteus aculeatus." BMC Molecular Biology **9:**1.

Hill, A. B. (1965). "Environment and disease - association or causation." Proceedings of the Royal Society of Medicine-London **58**(5): 295-300.

Hoebe, K. and B. Beutler (2008). "Forward genetic analysis of TLR-signaling pathways: An evaluation." Advanced Drug Delivery Reviews **60**(7): 824-829.

Hohenlohe, P. A., S. Bassham, P. D. Etter, N. Stiffler, E. A. Johnson and W. A. Cresko (2010). "Population Genomics of Parallel Adaptation in Threespine Stickleback using Sequenced RAD Tags." Plos Genetics **6**(2).

Hori, M. (1993). "FREQUENCY-DEPENDENT NATURAL-SELECTION IN THE HANDEDNESS OF SCALE-EATING CICHLID FISH." Science **260**(5105): 216-219.

Hoso, M., Y. Kameda, S. P. Wu, T. Asami, M. Kato and M. Hori (2010). "A speciation gene for left-right reversal in snails results in anti-predator adaptation." Nature Communications **1**.

Hosoiri, Y., Y. Harada and R. Kuroda (2003). "Construction of a backcross progeny collection of dextral and sinistral individuals of a freshwater gastropod, Lymnaea stagnalis." Development Genes and Evolution **213**(4): 193-198.

Houchmandzadeh, B., E. Wieschaus and S. Leibler (2002). "Establishment of developmental precision and proportions in the early Drosophila embryo." Nature **415**(6873): 798-802.

Hozumi, S., R. Maeda, K. Taniguchi, M. Kanai, S. Shirakabe, T. Sasamura, P. Speder, S. Noselli, T. Aigaki, R. Murakami and K. Matsuno (2006). "An unconventional myosin in Drosophila reverses the default handedness in visceral organs." Nature **440**(7085): 798-802.

Huang, X. H., Q. Feng, Q. Qian, Q. Zhao, L. Wang, A. H. Wang, J. P. Guan, D. L. Fan, Q. J. Weng, T. Huang, G. J. Dong, T. Sang and B. Han (2009). "High-throughput genotyping by whole-genome resequencing." Genome Research **19**(6): 1068-1076.

Illumina (2011). "HiSeq Sequencing Systems." Specification sheet: Illumina sequencing Pub. No. 770-2010-014. Current as of 02 May 2011.

Illumina (2015). "HiSeq 3000/HiSeq 4000 Sequencing Systems." Specification sheet: Illumina sequencing Pub. No. 770-2014-057. Current as of Aug 2015.

Jacob, F., R. Guertler, S. Naim, S. Nixdorf, A. Fedier, N. F. Hacker and V. Heinzelmann-Schwarz (2013). "Careful Selection of Reference Genes Is Required for Reliable Performance of RT-qPCR in Human Normal and Cancer Cell Lines." Plos One **8**(3).

Janky, R. s., J. van Helden and M. M. Babu (2009). "Investigating transcriptional regulation: From analysis of complex networks to discovery of cis-regulatory elements." Methods **48**(3): 277-286.

Jeong, Y. J., H. W. Choi, H. S. Shin, X. S. Cui, N. H. Kim, G. L. Gerton and J. H. Jun (2005). "Optimization of real time RT-PCR methods for the analysis of gene expression in mouse eggs and preimplantation embryos." Molecular Reproduction and Development **71**(3): 284-289.

Johnson, M. S. (1982). "Polymorphism For Direction Of Coil In Partula-Suturalis - Behavioral Isolation And Positive Frequency-Dependent Selection." Heredity **49**(OCT): 145-151.

Johnson, S. M., F. J. Tan, H. L. McCullough, D. P. Riordan and A. Z. Fire (2006). "Flexibility and constraint in the nucleosome core landscape of Caenorhabditis elegans chromatin." Genome Research **16**(12): 1505-1516.

Kadler, K. E., D. F. Holmes, J. A. Trotter and J. A. Chapman (1996). "Collagen fibril formation." Biochemical Journal **316**: 1-11.

Karsenti, E. and I. Vernos (2001). "Cell cycle - The mitotic spindle: A self-made machine." Science **294**(5542): 543-547.

Kartagener, M. and P. Stucki (1962). "Bronchiectasis With Situs Inversus." Archives of Pediatrics **79**(6): 193-&.

Keene, D. R., J. T. Oxford and N. P. Morris (1995). "Ultrastructural-Localization Of Collagen Type-Ii, Type-Ix, And Type-Xi In The Growth-Plate Of Human Rib And Fetal Bovine Epiphyseal Cartilage - Type-Xi Collagen Is Restricted To Thin Fibrils." Journal of Histochemistry & Cytochemistry **43**(10): 967-979.

Kelvin, L. W. T. (1904). Baltimore lectures on molecular dynamics and the wave theory of light. London, Cambridge University Press.

Kilner, P. J., G. Z. Yang, A. J. Wilkes, R. H. Mohiaddin, D. N. Firmin and M. H. Yacoub (2000). "Asymmetric redirection of flow through the heart." Nature **404**(6779): 759-761.

Klar, A. J. S. (1994). "A model for specification of the left-right axis in vertebrates." Trends in Genetics **10**(11): 392-396.

Klar, A. J. S. (2008). "Support for the selective chromatid segregation hypothesis advanced for the mechanism of left-right body axis development in mice." Breast disease **29**.

Koch, R. (1882). Die Aetiologie der Tuberculose. Reprinted in Clark D (ed) (1942). Source book of medical history. New York. , Dover Publications.

Koene, J. M. and J. Cosijn (2012). "Twisted sex in an hermaphrodite: mirror-image mating behaviour is not learned." Journal of Molluscan Studies **78**: 308-311.

Konig, A., R. Happle, D. Bornholdt, H. Engel and K. H. Grzeschik (2000). "Mutations in the NSDHL gene, encoding a 3 beta-hydroxysteroid dehydrogenase, cause CHILD syndrome." American Journal of Medical Genetics **90**(4): 339-346.

Korr, D., L. Toschi, P. Donner, H. D. Pohlenz, B. Kreft and B. Weiss (2006). "LRRK1 protein kinase activity is stimulated upon binding of GTP to its Roc domain." Cellular Signalling **18**(6): 910-920.

Kovar, D. R. (2006). "Molecular details of formin-mediated actin assembly." Current Opinion in Cell Biology **18**(1): 11-17.

Kubat, Z. (2007). "Chromosome walking with BAC clones as a method of genome mapping." Plant Soil and Environment **53**(10): 447-450.

Kuijk, E. W., L. du Puy, H. T. A. van Tol, H. P. Haagsman, B. Colenbrander and B. A. J. Roelen (2007). "Validation of reference genes for quantitative RT-PCR studies in porcine oocytes and preimplantation embryos." Bmc Developmental Biology **7**.

Kuroda, R. (2014). "How a Single Gene Twists a Snail." Integrative and Comparative Biology **54**(4): 677-687.

Kuroda, R., B. Endo, M. Abe and M. Shimizu (2009). "Chiral blastomere arrangement dictates zygotic left-right asymmetry pathway in snails." Nature **462**(7274): 790-U112.

Kvam, V. M., P. Lu and Y. Si (2012). "A COMPARISON OF STATISTICAL METHODS FOR DETECTING DIFFERENTIALLY EXPRESSED GENES FROM RNA-SEQ DATA." American Journal of Botany **99**(2): 248-256.

Lee, J. S., W. S. Kwon, M. S. Rahman, S. J. Yoon, Y. J. Park and M. G. Pang (2015). "Actin-related protein 2/3 complex-based actin polymerization is critical for male fertility." Andrology **3**(5): 937-946.

Levin, J. Z., M. Yassour, X. Adiconis, C. Nusbaum, D. A. Thompson, N. Friedman, A. Gnirke and A. Regev (2010). "Comprehensive comparative analysis of strand-specific RNA sequencing methods." Nature Methods **7**(9): 709-U767.

Levin, M. (1998). "Left-right asymmetry and the chick embryo." Seminars in Cell & Developmental Biology **9**(1): 67-76.

Levin, M. (2003). “Motor protein control of ion flux is an early step in embryonic left–right asymmetry”. BioEssays **25**:1002–1010.

Levin, M. (2005). "Left-right asymmetry in embryonic development: a comprehensive review." Mechanisms of Development **122**(1): 3-25.

Levin, M., R. L. Johnson, C. D. Stern, M. Kuehn and C. Tabin (1995). "A MOLECULAR PATHWAY DETERMINING LEFT-RIGHT ASYMMETRY IN CHICK EMBRYOGENESIS." Cell **82**(5): 803-814.

Levin, M. and A. R. Palmer (2007). "Left-right patterning from the inside out: widespread evidence for intracellular control." Bioessays **29**(3): 271-287.

Levin, M., T. Thorlin, K. R. Robinson, T. Nogi and M. Mercola (2002). "Asymmetries in H+/K+-ATPase and cell membrane potentials comprise a very early step in left-right patterning." Cell **111**(1): 77-89.

Li, F. and H. N. Higgs (2003). "The mouse formin mDia1 is a potent actin nucleation factor regulated by autoinhibition." Current Biology **13**(15): 1335-1340.

Li, P., Y. Piao, H. S. Shon and K. H. Ryu (2015). "Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data." Bmc Bioinformatics **16**.

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

Liu, M. M., J. W. Davey, R. Banerjee, J. Han, F. Yang, A. Aboobaker, M. L. Blaxter and A. Davison (2013). "Fine Mapping of the Pond Snail Left-Right Asymmetry (Chirality) Locus Using RAD-Seq and Fibre-FISH." Plos One **8**(8).

Liu, M. M., J. W. Davey, D. J. Jackson, M. L. Blaxter and A. Davison (2014). "A conserved set of maternal genes? Insights from a molluscan transcriptome." International Journal of Developmental Biology **58**(6-8): 501-511.

Liu, Y., P. Dodds, G. Emilion, A. J. Mungall, I. Dunham, S. Beck, R. S. Wells, F. M. L. Charnock and T. S. Ganesan (2002). "The human homologue of unc-93 maps to chromosome 6q27 - characterisation and analysis in sporadic epithelial ovarian cancer." Bmc Genetics **3**.

Livak, K. J. and T. D. Schmittgen (2001). "Analysis of relative gene expression data using real-time quantitative PCR and the 2(T)(-Delta Delta C) method." Methods **25**(4): 402-408.

Lo, H. S., Z. N. Wang, Y. Hu, H. H. Yang, S. Gere, K. H. Buetow and M. P. Lee (2003). "Allelic variation in gene expression is common in the human genome." Genome Research **13**(8): 1855-1862.

Lobikin, M., G. Wang, J. S. Xu, Y. W. Hsieh, C. F. Chuang, J. M. Lemire and M. Levin (2012). "Early, nonciliary role for microtubule proteins in left-right patterning is conserved across kingdoms." Proceedings of the National Academy of Sciences of the United States of America **109**(31): 12586-12591.

Loman, N. J., R. V. Misra, T. J. Dallman, C. Constantinidou, S. E. Gharbia, J. Wain and M. J. Pallen (2012). "Performance comparison of benchtop high-throughput sequencing platforms." Nature Biotechnology **30**(5): 434-+.

Lykke-Anderson, S. and T. H. Jensen (2015). "Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes." Nature Reviews Molecular Cell Biology **16**(11): 665-677.

Lykken, D. T. (1968). "STATISTICAL SIGNIFICANCE IN PSYCHOLOGICAL RESEARCH." Psychological Bulletin **70**(3P1): 151-&.

MacArthur, D. G., T. A. Manolio, D. P. Dimmock, H. L. Rehm, J. Shendure, G. R. Abecasis, D. R. Adams, R. B. Altman, S. E. Antonarakis, E. A. Ashley, J. C. Barrett, L. G. Biesecker, D. F. Conrad, G. M. Cooper, N. J. Cox, M. J. Daly, M. B. Gerstein, D. B. Goldstein, J. N. Hirschhorn, S. M. Leal, L. A. Pennacchio, J. A. Stamatoyannopoulos, S. R. Sunyaev, D. Valle, B. F. Voight, W. Winckler and C. Gunter (2014). "Guidelines for investigating causality of sequence variants in human disease." Nature **508**(7497): 469-476.

Mahoney, N. M., G. Goshima, A. D. Douglass and R. D. Vale (2006). "Making microtubules and mitotic spindles in cells without functional centrosomes." Current Biology **16**(6): 564-569.

Manner, J. (2001). "Does an equivalent of the "ventral node" exist in chick embryos? A scanning electron microscopic study." Anatomy and Embryology **203**(6): 481-490.

Maravillas-Montero, J. L. and L. Santos-Argumedo (2012). "The myosin family: unconventional roles of actin-dependent molecular motors in immune cells." Journal of Leukocyte Biology **91**(1): 35-46.

Mardis, E. R. (2011). "A decade's perspective on DNA sequencing technology." Nature **470**(7333): 198-203.

Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. T. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt,

G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. G. Yu, R. F. Begley and J. M. Rothberg (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature **437**(7057): 376-380.

Marioni, J. C., C. E. Mason, S. M. Mane, M. Stephens and Y. Gilad (2008). "RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays." Genome Research **18**(9): 1509-1517.

Martin, K. C. and A. Ephrussi (2009). "mRNA Localization: Gene Expression in the Spatial Dimension." Cell **136**(4): 719-730.

Martindale, M. Q. (2005). “The evolution of metazoan axial properties”. Nature Reviews Genetics **6**: 917-927.

Mason, S. F. (1991). Origins of the handedness of biological molecules. Biological Asymmetry and Handedness. Wiley, Chichester (Ciba Foundation symposium 162) . p3-15.

Matsuzaki, F., T. Ohshiro, H. Ikeshima-Kataoka and H. Izumi (1998). "miranda localizes staufen and prospero asymmetrically in mitotic neuroblasts and epithelial cells in early Drosophila embryogenesis." Development **125**(20): 4089-4098.

Mattick, J. S. and I. V. Makunin (2006). "Non-coding RNA." Human Molecular Genetics **15**: R17-R29.

McGettigan, P. A. (2013). "Transcriptomics in the RNA-seq era." Current Opinion in Chemical Biology **17**(1): 4-11.

McGrath, J., S. Somlo, S. Makova, X. Tian and M. Brueckner (2003). "Two populations of node monocilia initiate left-right asymmetry in the mouse." Cell **114**(1): 61-73.

McIntyre, L. M., K. K. Lopiano, A. M. Morse, V. Amin, A. L. Oberg, L. J. Young and S. V. Nuzhdin (2011). "RNA-seq: technical variability and sampling." Bmc Genomics **12**.

McKernan, K. J., H. E. Peckham, G. L. Costa, S. F. McLaughlin, Y. T. Fu, E. F. Tsung, C. R. Clouser, C. Duncan, J. K. Ichikawa, C. C. Lee, Z. Zhang, S. S. Ranade, E. T. Dimalanta, F. C. Hyland, T. D. Sokolsky, L. Zhang, A. Sheridan, H. N. Fu, C. L. Hendrickson, B. Li, L. Kotler, J. R. Stuart, J. A. Malek, J. M. Manning, A. A. Antipova, D. S. Perez, M. P. Moore, K. C. Hayashibara, M. R. Lyons, R. E. Beaudoin, B. E. Coleman, M. W. Laptewicz, A. E. Sannicandro, M. D. Rhodes, R. K. Gottimukkala, S. Yang, V. Bafna, A. Bashir, A. MacBride, C. Alkan, J. M. Kidd, E. E. Eichler, M. G. Reese, F. M. De la Vega and A. P. Blanchard (2009). "Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding." Genome Research **19**(9): 1527-1541.

McManus, C. (2002). Right hand, left hand: The origins of asymmetry in brains, bodies, atoms and cultures. London, UK, Weidenfeld & Nicolson**:** pp 121-145.

McManus, C. (2002). Right hand, left hand: The origins of asymmetry in brains, bodies, atoms and cultures. London, UK, Weidenfeld & Nicolson pp. 1-15.

McManus, C. (2002). Right hand, left hand: The origins of asymmetry in brains, bodies, atoms and cultures. London, UK, Weidenfeld & Nicolson.

McMillen, T. and A. Goriely (2002). "Tendril perversion in intrinsically curved rods." Journal of Nonlinear Science **12**(3): 241-281.

Mehta, A. D., R. S. Rock, M. Rief, J. A. Spudich, M. S. Mooseker and R. E. Cheney (1999). "Myosin-V is a processive actin-based motor." Nature **400**(6744): 590-593.

Meneely, P. (2009). Advanced genetic analysis: Genes, genomes and networks in eukaryotes. pp 8-27. New York, Oxford University Press Inc.

Mercer, T. R., M. E. Dinger and J. S. Mattick (2009). "Long non-coding RNAs: insights into functions." Nature Reviews Genetics **10**(3): 155-159.

Meshcheryakov, V. N. (1990). The Common Pond Snail *Lymnaea stagnalis*. Animal Species for Developmental Studies: Volume 1 Invertebrates. T. A. Dettlaff and S. G. Vassetzky. US, Springer**:** pp 69-132.

Mikheyev, A. S. and M. M. Y. Tin (2014). "A first look at the Oxford Nanopore MinION sequencer." Molecular Ecology Resources **14**(6): 1097-1102.

Miller, M. R., J. P. Dunham, A. Amores, W. A. Cresko and E. A. Johnson (2007). "Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers." Genome Research **17**(2): 240-248.

Minoche, A. E., J. C. Dohm and H. Himmelbauer (2011). "Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems." Genome Biology **12**(11).

Miwa, H., A. Miyake, Y. Kouta, A. Shimada, Y. Yamashita, Y. Nakayama, H. Yamauchi, M. Konishi and N. Itoh (2009). "A novel neural-specific BMP antagonist, Brorin-like, of the Chordin family." Febs Letters **583**(22): 3643-3648.

Mogilner, A. and B. Fogelson (2015). "Cytoskeletal Chirality: Swirling Cells Tell Left from Right." Current Biology **25**(12): R501-R503.

Morokuma, J., M. Ueno, H. Kawanishi, H. Saiga and H. Nishida (2002). "HrNodal, the ascidian nodal-related gene, is expressed in the left side of the epidermis, and lies upstream of HrPitx." Development Genes and Evolution **212**(9): 439-446.

Morrill, J. B. (1982). Developmental Biology of the Pulmonate Gastropod, Lymnaea. In Developmental Biology of Freshwater Invertebrates (Ed. Harrisson, F. W.). Alan R. Liss, Inc. New York. pp 399-483

Murray, J. and B. Clarke (1966). "The inheritance of polymorphic shell characters in partula (gastropoda)." Genetics **54**(5).

Nagashima, K., S. Torii, Z. H. Yi, M. Igarashi, K. Okamoto, T. Takeuchi and T. Izumi (2002). "Melanophilin directly links Rab27a and myosin Va through its distinct coiled-coil regions." Febs Letters **517**(1-3): 233-238.

Nakano, K., J. Imai, R. Arai, A. Toh-e, Y. Matsui and I. Mabuchi (2002). "The small GTPase Rho3 and the diaphanous/formin For3 function in polarized cell growth in fission yeast." Journal of Cell Science **115**(23): 4629-4639.

National Center for Research Resources, N. I. o. H. (2010). Realizing the Scientific Potential of Transcriptomics in Aquatic Models, Portland, Oregon, US.

Neu-Yilik, G., N. H. Gehring, M. W. Hentze and A. E. Kulozik (2004). "Nonsense-mediated mRNA decay: from vacuum cleaner to Swiss army knife." Genome Biology **5**(4).

Neville, A. C. (1976). "THE INSTITUTE OF BIOLOGYS LONDON ENGLAND STUDIES IN BIOLOGY NO 67 ANIMAL ASYMMETRY." Institute of Biology's Studies in Biology: 60.

Nielsen, K. L., A. L. Hogh and J. Emmersen (2006). "DeepSAGE - digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples." Nucleic Acids Research **34**(19).

Nobelprize.org (Nobel Media AB 2014.). "The Nobel Prize in Chemistry 2001 - Popular Information". <http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2001/popular.html>Web. 13 Nov 2015. ."

Nodine, M. D., R. Yadegari and F. E. Tax (2007). "RPK1 and TOAD2 are two receptor-like kinases redundantly required for Arabidopsis embryonic pattern formation." Developmental Cell **12**(6): 943-956.

Nonaka, S., Y. Tanaka, Y. Okada, S. Takeda, A. Harada, Y. Kanai, M. Kido and N. Hirokawa (1998). "Randomization of left-right asymmetry due to loss of nodal cilia generating leftward flow of extraembryonic fluid in mice lacking KIF3B motor protein." Cell **95**(6): 829-837.

Okada, Y., S. Nonaka, Y. Tanaka, Y. Saijoh, H. Hamada and N. Hirokawa (1999). "Abnormal nodal flow precedes situs inversus in iv and inv mice." Molecular Cell **4**(4): 459-468.

Okada, Y., S. Takeda, Y. Tanaka, J. C. I. Belmonte and N. Hirokawa (2005). "Mechanism of nodal flow: A conserved symmetry breaking event in left-right axis determination." Cell **121**(4): 633-644.

Okumura, T., H. Utsuno, J. Kuroda, E. Gittenberger, T. Asami and K. Matsuno (2008). "The Development and Evolution of Left-Right Asymmetry in Invertebrates: Lessons From Drosophila and Snails." Developmental Dynamics **237**(12): 3497-3515.

Oshlack, A. and M. J. Wakefield (2009). "Transcript length bias in RNA-seq data confounds systems biology." Biology Direct **4**.

Oviedo, N. J. and Levin, M. (2007). “Gap Junctions Provide New Links in Left-Right Patterning”. Cell. **129**: 645-647.

Page, G. P., V. George, R. C. Go, P. Z. Page and D. B. Allison (2003). ""Are we there yet?": Deciding when one has demonstrated specific genetic causation in complex diseases and quantitative traits." American Journal of Human Genetics **73**(4): 711-719.

Palmer, M. and E. Prediger. (2015). "Assessing RNA quality." Ambion, TechNotes, 2015.

Pearl, J. (2000). Causality: models, reasoning, and inference. New York, NY, USA, Cambridge University Press

Peeters, H. and K. Devriendt (2006). "Human laterality disorders." European Journal of Medical Genetics **49**(5): 349-362.

Pellino, M., T. F. Sharbel, M. Mau, S. Amiteye and J. M. Corral (2011). "Selection of reference genes for quantitative real-time PCR expression studies of microdissected reproductive tissues in apomictic and sexual Boechera." BMC Research Notes **4**(303).

Peterson, B. K., J. N. Weber, E. H. Kay, H. S. Fisher and H. E. Hoekstra (2012). "Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species." Plos One **7**(5).

Pfaffl, M. W. (2001). "A new mathematical model for relative quantification in real-time RT-PCR." Nucleic Acids Research **29**(9).

Pfaffl, M. W., G. W. Horgan and L. Dempfle (2002). "Relative expression software tool (REST (c)) for group-wise comparison and statistical analysis of relative expression results in real-time PCR." Nucleic Acids Research **30**(9).

Pfaffl, M. W., A. Tichopad, C. Prgomet and T. P. Neuvians (2004). "Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper - Excel-based tool using pair-wise correlations." Biotechnology Letters **26**(6): 509-515.

Pfaffl, M. W., J. Vandesompele and M. Kubista (2009). Data Analysis Software.

Pollard, T. D. (2007). Regulation of actin filament assembly by Arp2/3 complex and formins. Annual Review of Biophysics and Biomolecular Structure. **36:** 451-477.

Ponder, W. F. and D. R. Lindberg (1997). "Towards a phylogeny of gastropod molluscs: Analysis using morphological characters." Zoological Journal of the Linnean Society **119**(2): 83-265.

PrimerDesign (2014). "geNorm primer only kit handbook. HB01.02.05."

PrimerDesign (2014). "geNorm(TM) kit with Perfect Probe(TM) handbook. HB01.01.02."

Pushkarev, D., N. F. Neff and S. R. Quake (2009). "Single-molecule sequencing of an individual human genome." Nature Biotechnology **27**(9): 847-U101.

Qiagen (2007). "RNeasy micro handbook, 2nd Edition." **Dec. 2007**.

Quail, M. A., M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow and Y. Gu (2012). "A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers." Bmc Genomics **13**.

Quick, J., A. R. Quinlan and N. J. Loman (2014). "A reference bacterial genome dataset generated on the MinION portable single-molecule nanopore sequencer." GigaScience **3**.

Radonic, A., S. Thulke, I. M. Mackay, O. Landt, W. Siegert and A. Nitsche (2004). "Guideline to reference gene selection for quantitative real-time PCR." Biochemical and Biophysical Research Communications **313**(4): 856-862.

Redowicz, M. J. (2007). "Unconventional myosins in muscle." European Journal of Cell Biology **86**(9): 549-558.

Reise, H., M. Benke and J. M. C. Hutchinson (2002). "A sinistral specimen of the terrestrial slug Arion lusitanicus (Gastropoda: Pulmonata: Arionidae)." Malakologische Abhandlungen (Dresden) **20**(2): 247-252.

Reuter, J. A., D. V. Spacek and M. P. Snyder (2015). "High-Throughput Sequencing Technologies." Molecular Cell **58**(4): 586-597.

Ribeiro, M., M. Schofield, I. Kemenes, P. R. Benjamin, M. O'Shea and S. A. Korneev (2010). "Atypical Guanylyl Cyclase from the Pond Snail Lymnaea Stagnalis: Cloning, Sequence Analysis and Characterization of Expression." Neuroscience **165**(3): 794-800.

Rieu, I. and S. J. Powers (2009). "Real-Time Quantitative RT-PCR: Design, Calculations, and Statistics." Plant Cell **21**(4): 1031-1033.

Rigon, F., G. Manica, F. Guma, M. Achaval and M. C. Faccioni-Heuser (2010). "Ultrastructural features of the columellar muscle and contractile protein analyses in different muscle groups of Megalobulimus abbreviatus (Gastropoda, Pulmonata)." Tissue & Cell **42**(1): 53-60.

Robertson, R. (1993). "SNAIL HANDEDNESS - THE COILING DIRECTIONS OF GASTROPODS." Research & Exploration **9**(1): 104-119.

Robinson, M., D. McCarthy, Y. Chen and G. K. Smyth (2012). "edgeR: differential expression analysis of digital gene expression data. User's guide." First edition, 17th September 2008, Revised version 17th December 2012.

Robinson, M. D., D. J. McCarthy and G. K. Smyth (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." Bioinformatics **26**(1): 139-140.

Robinson, M. D. and A. Oshlack (2010). "A scaling normalization method for differential expression analysis of RNA-seq data." Genome Biology **11**(3).

Robinson, M. D. and G. K. Smyth (2007). "Moderated statistical tests for assessing differences in tag abundance." Bioinformatics **23**(21): 2881-2887.

Robinson, M. D. and G. K. Smyth (2008). "Small-sample estimation of negative binomial dispersion, with applications to SAGE data." Biostatistics **9**(2): 321-332.

Rosenbaum, P. R. (1995). "Quantiles in nonrandom samples and observational studies." Journal of the American Statistical Association **90**(432): 1424-1431.

Rothberg, J. M., W. Hinz, T. M. Rearick, J. Schultz, W. Mileski, M. Davey, J. H. Leamon, K. Johnson, M. J. Milgrew, M. Edwards, J. Hoon, J. F. Simons, D. Marran, J. W. Myers, J. F. Davidson, A. Branting, J. R. Nobile, B. P. Puc, D. Light, T. A. Clark, M. Huber, J. T. Branciforte, I. B. Stoner, S. E. Cawley, M. Lyons, Y. T. Fu, N. Homer, M. Sedova, X. Miao, B. Reed, J. Sabina, E. Feierstein, M. Schorn, M. Alanjary, E. Dimalanta, D. Dressman, R. Kasinskas, T. Sokolsky, J. A. Fidanza, E. Namsaraev, K. J. McKernan, A. Williams, G. T. Roth and J. Bustillo (2011). "An integrated semiconductor device enabling non-optical genome sequencing." Nature **475**(7356): 348-352.

Rowe, H. C., S. Renaut and A. Guggisberg (2011). "RAD in the realm of next-generation sequencing technologies." Molecular Ecology **20**(17): 3499-3502.

Rubin, D. B. (1991). "PRACTICAL IMPLICATIONS OF MODES OF STATISTICAL-INFERENCE FOR CAUSAL EFFECTS AND THE CRITICAL ROLE OF THE ASSIGNMENT MECHANISM." Biometrics **47**(4): 1213-1234.

Ruppert, E. E., R. S. Fox and R. Barnes, . D. (2004). Invertebrate zoology: A functional evolutionary approach. Seventh Edition. Belmont, California, Brooks/Cole Thompson Learning**:** pp. 197.

Sander, J. D. and J. K. Joung (2014). "CRIRISPR-Cas systems for editing, regulating and targeting genomes." Nature Biotechnology **32**(4): 347-355.

Sanger, F., G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, J. C. Fiddes, C. A. Hutchison, P. M. Slocombe and M. Smith (1977). "NUCLEOTIDE-SEQUENCE OF BACTERIOPHAGE PHICHI174 DNA." Nature **265**(5596): 687-695.

Sanger, F., S. Nicklen and A. R. Coulson (1977). "DNA SEQUENCING WITH CHAIN-TERMINATING INHIBITORS." Proceedings of the National Academy of Sciences of the United States of America **74**(12): 5463-5467.

Schefe, J. H., K. E. Lehmann, I. R. Buschmann, T. Unger and H. Funke-Kaiser (2006). "Quantitative real-time RT-PCR data analysis: current concepts and the novel "gene expression's C-T difference" formula." Journal of Molecular Medicine-Jmm **84**(11): 901-910.

Schier, A. F. (2009). "Nodal Morphogens." Cold Spring Harbor Perspectives in Biology **1**(5).

Schilthuizen, M., P. G. Craze, A. S. Cabanban, A. Davison, J. Stone, E. Gittenberger and B. J. Scott (2007). "Sexual selection maintains whole-body chiral dimorphism in snails." Journal of Evolutionary Biology **20**(5): 1941-1949.

Schilthuizen, M. and A. Davison (2005). "The convoluted evolution of snail chirality." Naturwissenschaften **92**(11): 504-515.

Schilthuizen, M., B. J. Scott, A. S. Cabanban and P. G. Craze (2005). "Population structure and coil dimorphism in a tropical land snail." Heredity **95**(3): 216-220.

Schurch, N. J., P. Schofield, M. Gierliński, C. Cole, A. Sherstnev, V. Singh, N. Wrobel, K. Gharbi, G. G. Simpson, T. Owen-Hughes, M. Blaxter and G. J. Barton (2015). "Evaluation of tools for differential gene expression analysis by RNA-seq on a 48 biological replicate experiment." eprint arXiv:1505.02017. Bibliographic code: 2015arXiv150502017S.

Sellers, J. R. (2000). "Myosins: a diverse superfamily." Biochimica Et Biophysica Acta-Molecular Cell Research **1496**(1): 3-22.

Sellers, J. R. and C. Veigel (2006). "Walking with myosin V." Current Opinion in Cell Biology **18**(1): 68-73.

Serre, D., S. Gurd, B. Ge, R. Sladek, D. Sinnett, E. Harmsen, M. Bibikova, E. Chudin, D. L. Barker, T. Dickinson, J. B. Fan and T. J. Hudson (2008). "Differential allelic expression in the human genome: A robust approach to identify genetic and epigenetic Cis-acting mechanisms regulating gene expression." Plos Genetics **4**(2).

Shen, M. M. (2007). "Nodal signaling: developmental roles and regulation." Development **134**(6): 1023-1034.

Shestakova, E. A., L. P. Motuz, A. A. Minin and L. P. Gavrilova (1993). "STUDY OF LOCALIZATION OF THE PROTEIN-SYNTHESIZING MACHINERY ALONG ACTIN FILAMENT BUNDLES." Cell Biology International **17**(4): 409-416.

Shibazaki, Y., M. Shimizu and R. Kuroda (2004). "Body handedness is directed by genetically determined cytoskeletal dynamics in the early embryo." Current Biology **14**(16): 1462-1467.

Silver, N., S. Best, J. Jiang and S. L. Thein (2006). "Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time PCR." Bmc Molecular Biology **7**.

Sindelka, R., Z. Ferjentsik and J. Jonak (2006). "Developmental expression profiles of Xenopus laevis reference genes." Developmental Dynamics **235**(3): 754-758.

Slomovic, S., D. Laufer, D. Geiger and G. Schuster (2006). "Polyadenylation of ribosomal RNA in human cells." Nucleic Acids Research **34**(10): 2966-2975.

Speder, P., G. Adam and S. Noselli (2006). "Type ID unconventional myosin controls left-right asymmetry in Drosophila." Nature **440**(7085): 803-807.

Speder, P., A. Petzoldt, M. Suzanne and S. Noselli (2007). "Strategies to establish left/right asymmetry in vertebrates and invertebrates." Current Opinion in Genetics & Development **17**(4): 351-358.

St Johnston, D., D. Beuchle and C. Nussleinvolhard (1991). "Staufen, a gene required to localize maternal RNAs in the Drosophila egg." Cell **66**(1): 51-63.

Stefani, G. and F. J. Slack (2008). "Small non-coding RNAs in animal development." Nature Reviews Molecular Cell Biology **9**(3): 219-230.

Steinert, P. M. and L. N. Marekov (1997). "Direct evidence that involucrin is a major early isopeptide crosslinked component of the keratinocyte cornified cell envelope." Journal of Biological Chemistry **272**(3): 2021-2030.

Streisin.G, Y. Okada, J. Emrich, J. Newton, A. Tsugita, E. Terzaghi and M. Inouye (1966). "FRAMESHIFT MUTATIONS AND GENETIC CODE." Cold Spring Harbor Symposia on Quantitative Biology **31**: 77-&.

Sturtevant, A. H. (1923). "Inheritance of direction of coiling in Limnaea." Science **58**(1501): 269-270.

Sutcharit, C., T. Asami and S. Panha (2007). "Evolution of whole-body enantiomorphy in the tree snail genus Amphidromus." Journal of Evolutionary Biology **20**(2): 661-672.

Suzuki, S. T. (2000). "Recent progress in protocadherin research." Experimental Cell Research **261**(1): 13-18.

Tabin, C. J. and K. J. Vogan (2003). "A two-cilia model for vertebrate left-right axis specification." Genes & Development **17**(1): 1-6.

Taki, F. A., A. A. Abdel-Rahman and B. H. Zhang (2014). "A Comprehensive Approach to Identify Reliable Reference Gene Candidates to Investigate the Link between Alcoholism and Endocrinology in Sprague-Dawley Rats." Plos One **9**(5).

Tanoue, T. and M. Takeichi (2004). "Mammalian Fat1 cadherin regulates actin dynamics and cell-cell contact." Journal of Cell Biology **165**(4): 517-528.

Tanoue, T. and M. Takeichi (2005). "New insights into Fat cadherins." Journal of Cell Science **118**(11): 2347-2353.

Tarazona, S., F. Garcia-Alcalde, J. Dopazo, A. Ferrer and A. Conesa (2011). "Differential expression in RNA-seq: A matter of depth." Genome Research **21**(12): 2213-2223.

Taylor, K. H., R. S. Kramer, J. W. Davis, J. Guo, D. J. Duff, D. Xu, C. W. Caldwell and H. Shi (2007). "Ultradeep bisulfite sequencing analysis of DNA methylation patterns in multiple gene promoters by 454 sequencing." Cancer Research **67**(18): 8511-8518.

Taylor, S., M. Wakem, G. Dijkman, M. Alsarraj and M. Nguyen (2010). "A practical approach to RT-qPCR-Publishing data that conform to the MIQE guidelines." Methods **50**(4): S1-S5.

Tee, Y. H., T. Shemesh, V. Thiagarajan, R. F. Hariadi, K. L. Anderson, C. Page, N. Volkmann, D. Hanein, S. Sivaramakrishnan, M. M. Kozlov and A. D. Bershadsky (2015). "Cellular chirality arising from the self-organization of the actin cytoskeleton." Nature Cell Biology **17**(4): 445-+.

Teer, J. K. and J. C. Mullikin (2010). "Exome sequencing: the sweet spot before whole genomes." Human Molecular Genetics **19**: R145-R151.

ThermoScientific (2010). "Thermo Scientific NanoDrop Spectrophotometers, Nucleic Acid.": 14.

Thompson, S. P. (1910). The life of William Thompson, Baron Kelvin of Largs. London: , Macmillan.

Tian, T. and A. M. Meng (2006). "Nodal signals pattern vertebrate embryos." Cellular and Molecular Life Sciences **63**(6): 672-685.

Torgersen, J. (1950). "SITUS INVERSUS, ASYMMETRY, AND TWINNING." American Journal of Human Genetics **2**(4): 361-370.

Ueshima, R. and T. Asami (2003). "Single-gene speciation by left-right reversal - A land-snail species of polyphyletic origin results from chirality constraints on mating." Nature **425**(6959): 679-679.

Untergasser, A., I. Cutcutache, T. Koressaar, J. Ye, B. C. Faircloth, M. Remm and S. G. Rozen (2012). "Primer3-new capabilities and interfaces." Nucleic Acids Research **40**(15).

Utsuno, H., T. Asami, T. J. M. Van Dooren and E. Gittenberger (2011). "INTERNAL SELECTION AGAINST THE EVOLUTION OF LEFT-RIGHT REVERSAL." Evolution **65**(8): 2399-2411.

van den Biggelaar, J. A. M. (1991). "Asymmetries during molluscan embryogenesis." Biological asymmetry and handedness. Wiley, Chichester (Ciba Foundation symposium 162). p128-142.

van Kesteren, R. E., C. Carter, H. M. G. Dissel, J. van Minnen, Y. Gouwenberg, N. I. Syed, G. E. Spencer and A. B. Smit (2006). "Local synthesis of actin-binding protein beta-thymosin regulates neurite outgrowth." Journal of Neuroscience **26**(1): 152-157.

van Nierop, P., S. Bertrand, D. W. Munno, Y. Gouwenberg, J. van Minnen, J. D. Spafford, N. I. Syed, D. Bertrand and A. B. Smit (2006). "Identification and functional expression of a family of nicotinic acetylcholine receptor subunits in the central nervous system of the mollusc Lymnaea stagnalis." Journal of Biological Chemistry **281**(3): 1680-1691.

VanBatenburg, F. H. D. and E. Gittenberger (1996). "Ease of fixation of a change in coiling: Computer experiments on chirality in snails." Heredity **76**: 278-286.

Vandenberg, L. N. (2012). "Laterality defects are influenced by timing of treatments and animal model." Differentiation **83**(1): 26-37.

Vandenberg, L. N., J. M. Lemire and M. Levin (2013). "It's never too early to get it Right: A conserved role for the cytoskeleton in left-right asymmetry." Communicative & integrative biology **6**(6).

Vandenberg, L. N. and M. Levin (2010). "Consistent left-right asymmetry cannot be established by late organizers in Xenopus unless the late organizer is a conjoined twin." Development **137**(7): 1095-1105.

Vandenberg, L. N. and M. Levin (2010). "Far From Solved: A Perspective on What We Know About Early Mechanisms of Left-Right Asymmetry." Developmental Dynamics **239**(12): 3131-3146.

Vandenberg, L. N. and M. Levin (2013). "A unified model for left-right asymmetry? Comparison and synthesis of molecular models of embryonic laterality." Developmental Biology **379**(1): 1-15.

Vandesompele, J., K. De Preter, F. Pattyn, B. Poppe, N. Van Roy, A. De Paepe and F. Speleman (2002). "Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes." Genome Biology **3**(7).

VanGuilder, H. D., K. E. Vrana and W. M. Freeman (2008). "Twenty-five years of quantitative PCR for gene expression analysis." Biotechniques **44**(5): 619-626.

Velculescu, V. E., L. Zhang, B. Vogelstein and K. W. Kinzler (1995). "SERIAL ANALYSIS OF GENE-EXPRESSION." Science **270**(5235): 484-487.

Vera, J. C., C. W. Wheat, H. W. Fescemyer, M. J. Frilander, D. L. Crawford, I. Hanski and J. H. Marden (2008). "Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing." Molecular Ecology **17**(7): 1636-1647.

Vermeij, G. J. (1975). "EVOLUTION AND DISTRIBUTION OF LEFT-HANDED AND PLANISPIRAL COILING IN SNAILS." Nature **254**(5499): 419-420.

Vicente-Manzanares, M., X. F. Ma, R. S. Adelstein and A. R. Horwitz (2009). "Non-muscle myosin II takes centre stage in cell adhesion and migration." Nature Reviews Molecular Cell Biology **10**(11): 778-790.

Vogan, K. J. and C. J. Tabin (1999). "Developmental biology - A new spin on handed asymmetry." Nature **397**(6717): 295-+.

Vondarl, M., P. M. Harrison and W. Bottke (1994). "CDNA CLONING AND DEDUCED AMINO-ACID-SEQUENCE OF 2 FERRITINS - SOMA FERRITIN AND YOLK FERRITIN, FROM THE SNAIL LYMNAEA-STAGNALIS L." European Journal of Biochemistry **222**(2): 353-366.

Wade, C. M., P. B. Mordan and F. Naggs (2006). "Evolutionary relationships among the Pulmonate land snails and slugs (Pulmonata, Stylommatophora)." Biological Journal of the Linnean Society **87**(4): 593-610.

Wagatsuma, A., H. Sadamoto, T. Kitahashi, K. Lukowiak, A. Urano and E. Ito (2005). "Determination of the exact copy numbers of particular mRNAs in a single cell by quantitative real-time RT-PCR." Journal of Experimental Biology **208**(12): 2389-2398.

Walentek, P., T. Beyer, T. Thumberger, A. Schweickert and M. Blum (2012). "ATP4a Is Required for Wnt-Dependent Foxj1 Expression and Leftward Flow in Xenopus Left-Right Development." Cell Reports **1**(5): 516-527.

Wallar, B. J. and A. S. Alberts (2003). "The formins: active scaffolds that remodel the cytoskeleton." Trends in Cell Biology **13**(8): 435-446.

Wan, L. Q., K. Ronaldson, M. Park, G. Taylor, Y. Zhang, J. M. Gimble and G. Vunjak-Novakovic (2011). "Micropatterned mammalian cells exhibit phenotype-specific left-right asymmetry." Proceedings of the National Academy of Sciences of the United States of America **108**(30): 12295-12300.

Wang, Z., M. Gerstein and M. Snyder (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nature Reviews Genetics **10**(1): 57-63.

Watanabe, H., H. A. Schmidt, A. Kuhn, S. K. Hoger, Y. Kocagoz, N. Laumann-Lipp, S. Ozbek and T. W. Holstein (2014). "Nodal signalling determines biradial asymmetry in Hydra." Nature **515**: 112-115.

Watt, F. M. (1983). "INVOLUCRIN AND OTHER MARKERS OF KERATINOCYTE TERMINAL DIFFERENTIATION." Journal of Investigative Dermatology **81**(1): S100-S103.

Weier, H. U. G., M. Wang, K. M. Greulich, Y. Zhu, J. F. Cheng and J. W. Gray (1995). "Quantitative DNA fiber mapping: A rapid technique for high resolution physical map assembly." American Journal of Human Genetics **57**(4 SUPPL.): A16.

Weigel, D. and M. Nordborg (2005). "Natural variation in arabidopsis. How do we find the causal genes?" Plant Physiology **138**(2): 567-568.

Welch, M. D., A. H. DePace, S. Verma, A. Iwamatsu and T. J. Mitchison (1997). "The human Arp2/3 complex is composed of evolutionarily conserved subunits and is localized to cellular regions of dynamic actin filament assembly." Journal of Cell Biology **138**(2): 375-384.

Whittaker, C. A. and R. O. Hynes (2002). "Distribution and evolution of von Willebrand/integrin a domains: Widely dispersed adhesion and elsewhere." Molecular Biology of the Cell **13**(10): 3369-3387.

Wickham, H. (2009). ggplot2: elegant graphics for data analysis. . New York, Springer

Wilusz, J. E., H. Sunwoo and D. L. Spector (2009). "Long noncoding RNAs: functional surprises from the RNA world." Genes & Development **23**(13): 1494-1504.

Wolf, J. B. W., T. Bayer, B. Haubold, M. Schilhabel, P. Rosenstiel and D. Tautz (2010). "Nucleotide divergence vs. gene expression differentiation: comparative transcriptome sequencing in natural isolates from the carrion crow and its hybrid zone with the hooded crow." Molecular Ecology **19**: 162-175.

Wolpert, L. (1991). "Final general discussion". Biological asymmetry and handedness. Wiley, Chichester (Ciba Foundation symposium 162). p312-315.

Wood, D. L. A., K. Nones, A. Steptoe, A. Christ, I. Harliwong, F. Newell, T. J. C. Bruxner, D. Miller, N. Cloonan and S. M. Grimmond (2015). "Recommendations for Accurate Resolution of Gene and Isoform Allele-Specific Expression in RNA-Seq Data." Plos One **10**(5).

Wood, W. B. (1997). "Left-right asymmetry in animal development." Annual Review of Cell and Developmental Biology **13**: 53-82.

Wu, X. F., G. Jung and J. A. Hammer (2000). "Functions of unconventional myosins." Current Opinion in Cell Biology **12**(1): 42-51.

www.rdml.org. "Real-time PCR Data Markup Language website."

Xie, F. L., P. Xiao, D. L. Chen, L. Xu and B. H. Zhang (2012). "miRDeepFinder: a miRNA analysis tool for deep sequencing of plant small RNAs." Plant Molecular Biology **80**(1): 75-84.

Xu, J. S., A. Van Keymeulen, N. M. Wakida, P. Carlton, M. W. Berns and H. R. Bourne (2007). "Polarity reveals intrinsic cell chirality." Proceedings of the National Academy of Sciences of the United States of America **104**(22): 9296-9300.

Yamada, A., M. Yoshio, K. Oiwa and L. Nyitray (2000). "Catchin, a novel protein in molluscan catch muscles, is produced by alternative splicing from the myosin heavy chain gene." Journal of Molecular Biology **295**(2): 169-178.

Yamashita, R. A., J. R. Sellers and J. B. Anderson (2000). "Identification and analysis of the myosin superfamily in Drosophila: a database approach." Journal of Muscle Research and Cell Motility **21**(6): 491-505.

Yang, Y. J., R. M. Graze, B. M. Walts, C. M. Lopez, H. V. Baker, M. L. Wayne, S. V. Nuzhdin and L. M. McIntyre (2011). "Partitioning Transcript Variation in Drosophila: Abundance, Isoforms, and Alleles." G3-Genes Genomes Genetics **1**(6): 427-436.

Yu, J. K., L. Z. Holland and N. D. Holland (2002). "An amphioxus nodal gene (AmphiNodal) with early symmetrical expression in the organizer and mesoderm and later asymmetrical expression associated with left-right axis formation." Evolution & Development **4**(6): 418-425.

Zerbino, D. R. and E. Birney (2008). "Velvet: Algorithms for de novo short read assembly using de Bruijn graphs." Genome Research **18**(5): 821-829.

Zhou, X. L., H. Sasaki, L. Lowe, B. L. M. Hogan and M. R. Kuehn (1993). "NODAL IS A NOVEL TGF-BETA-LIKE GENE EXPRESSED IN THE MOUSE NODE DURING GASTRULATION." Nature **361**(6412): 543-547.

List of Tables

List of Figures

List of Boxes

List of Equations

Supplementary Information

Supplementary Table of Contents

S1. Whole-mount in situ hybridisation experiments

Whole mount in situ hybridisation techniques were developed for a number of the GOIs assessed in Chapter 3 during a research trip to the Jackson laboratory at the Georg-August-Universität Göttingen. The technique however, required a significant amount of optimisation and consistent results were not obtained during the time available. Included here is the final report submitted to the funding body, summarising the findings. The primary candidate genes have not yet been published and so the genes of interest (GOIs) were anonymously referred to as GOI A – GOI D. The corresponding gene descriptions are presented in Table S1.

Table S1 Gene name and associated protein description corresponding to the anonymous GOI IDs throughout the report.

Anonymous ID	Candidate Gene and description
GOI A	*Lmhc*, heavy chain myosin
GOI B	*Larp 2/3 1a*, actin-related protein 2/3 complex
GOI C	*Ldia2*, diaphanous related formin (includes frameshift mutation in the sinistral copy)
GOI D	*Ldia1*, diaphanous related formin

BOEHRINGER INGELHEIM FONDS TRAVEL GRANT: FINAL REPORT

Awardee: Harriet F Johnson, 2014

SUMMARY

This research trip was intended to measure the localised expression of a set of candidate genes associated with left-right asymmetry during development of the pond snail *Lymnaea stagnalis*. The results of the whole mount *in situ* experiments were too variable to assume any significance and it is apparent that this technique necessitates rigorous optimising, which was beyond the scope of this research stay. However, I have gained a wealth of experience in the manipulation of developmental stages of *L.stagnalis,* which has provided a substantial contribution to further work since returning to the University of Nottingham. In addition to gaining laboratory skills, this research exchange has enhanced collaborative relationships with the Jackson lab, whilst also forming new networks with those at the Georg-August-Universität Göttingen.

KEY SKILLS GAINED

- Cloned gene of interest
- Staging of *L.stagnalis* embryos
- Decapsulation and fixing of *L.stagnalis* embryos
- Whole-mount *in situ* hybridisation technique including riboprobe synthesis
- Use of automated robot - InsituPro VSi
- Northern blot analysis
- RNA extraction from *L.stagnalis* embryos

DESCRIPTION OF SCIENTIFIC ACTIVITIES

Aims

Consistent left-right (LR) asymmetry of the visceral organs is a highly conserved feature of animal development. Deviations from normal LR patterning can result in serious clinical consequences and may affect 1 in 5000 live births (Casey and Hackett 2000), yet much uncertainty remains regarding the mechanisms of LR axis specification during development. To gain a deep evolutionary understanding of development, a wide variety of model organisms are required. We are using snails to understand LR asymmetry, because their "chirality" is variable and determined at a very early stage in development.

Through collaboration with Daniel Jackson within the Courant Research Centre at the Georg-August-Universität Göttingen, I intended to use *in situ* expression techniques to reveal localisation and potential changes in gene expression that take place during the early development of LR variable, or "mirror image" snails.

In Situ Experiments

Riboprobe Synthesis

I arrived with several candidate gene sequences identified from previous analyses of *Lymnaea stagnalis* ovotestis tissue, to provide targets for whole-mount *in situ* hybridisation (WMISH) experiments in *L.stagnalis* embryos.

I was able to successfully clone specific cDNA fragments from four genes of interest (hereafter referred to as GOI A-D), from which complementary RNA binding probes (riboprobes) for use in WMISH were then synthesised. The riboprobe is a specific sequence of single-stranded RNA, labelled with digoxigenin-UTP (DIG), generated from a directional polymerase. The direction/orientation of the riboprobe is essential for it to complement correctly to the transcript in the tissue.

Sample preparation

WMISH allows for whole embryos to be analysed in the experiment. To prepare the tissues for analysis, eggs must first be decapsulated to isolate the embryos, which are then preserved/ fixed in paraformaldehyde.

In order for the gene-specific riboprobes to bind to their complimentary RNA transcripts within the embryo, the tissue must be permeable. Depending on the size of the embryo, further steps in the fixation protocol are required to increase permeability of the tissue. Generally early cleavage stage embryos do not require permeabilisation, however the fixation protocol is still being optimised for *L.stagnalis*.

Whole-mount in situ hybrisidation

The lengthy hybridisation steps were carried out via an automated robot: the *Intavis* 'InsituPro VSi'. To briefly summarise the protocol here, the prepared samples were first incubated with a hybridisation buffer at 50°C to which the gene-specific riboprobes were then added. Following a minimum incubation of ten hours allowing the riboprobes to hybridise to their complimentary targets, the excess riboprobe was washed away. The sample was then incubated with an antibody (anti-DIG), which was incorporated into the riboprobe bound to the specific transcript within the tissue.

Outside of the automated robot the sample was treated with a stain mix of NBT (nitro-blue tetrazolium chloride) and BCIP (5-bromo-4-chloro-3'-indolyphosphate). The stain produces a purple-blue colour in the presence of the antibody and consequently provides a signal of localised expression of the specific transcript in the tissue. A darker stain generally infers a higher level of transcription. This step is highly variable and time sensitive and so is performed manually, allowing for continuous observation.

Results

The results from the WMISH experiments were highly variable and it became apparent that this technique requires optimising specific for each riboprobe and developmental stage.

In developmental stages more than 24 hours post-cleavage, positive control genes performed as expected and showed consistent staining patterns, whereas the early cleavage stages, although exhibiting well preserved morphology, predominantly failed to stain or demonstrated unreliable staining patterns (Figure S1).

Analyses of the genes of interest in later developmental stage embryos were also more successful than those of early cleavage stages. GOI A in late stages displayed a consistent staining pattern

focused around the defined foot and mantle tissue. Embryos of 1-3 days post-cleavage, show a concentrated stain, however it is difficult to identify any key features or consistent staining pattern (Figure S2).

GOI B in late stages showed a weaker signal but a staining pattern similar to that of GOI A (Figure S3). It was in the early cleavage stages that GOI B gave an interesting signal. It appeared that GOI B was expressed ubiquitously in the early two-cell stage, yet this was reduced to one-sided expression in the late two-cell stage (Figure S4, c & d). However, it was difficult to reproduce this result, or in fact any staining in the early cleavage stages (Figure S4, e & f).

GOI C exhibited a potential sided difference in signal intensity in the four-cell stage embryo (Figure S5) and also demonstrated the same one-sided staining pattern as GOI B at the two-cell stage. However there were also difficulties in successful staining of this GOI from two alternative riboprobes. GOI D failed to produce a signal in any of WMISH experiments.

The one-sided staining pattern was also observed at the two-cell stage for the positive controls genes, which were expected to be ubiquitously expressed (Figure S1, d & e). It is unknown the reason for this effect, however due to the low intensity of the signal, coupled with the fact that approximately 50% of embryos failed to stain, it is unlikely to be a true signal. As such the staining patterns of all the early cleavage stages are considered inconclusive.

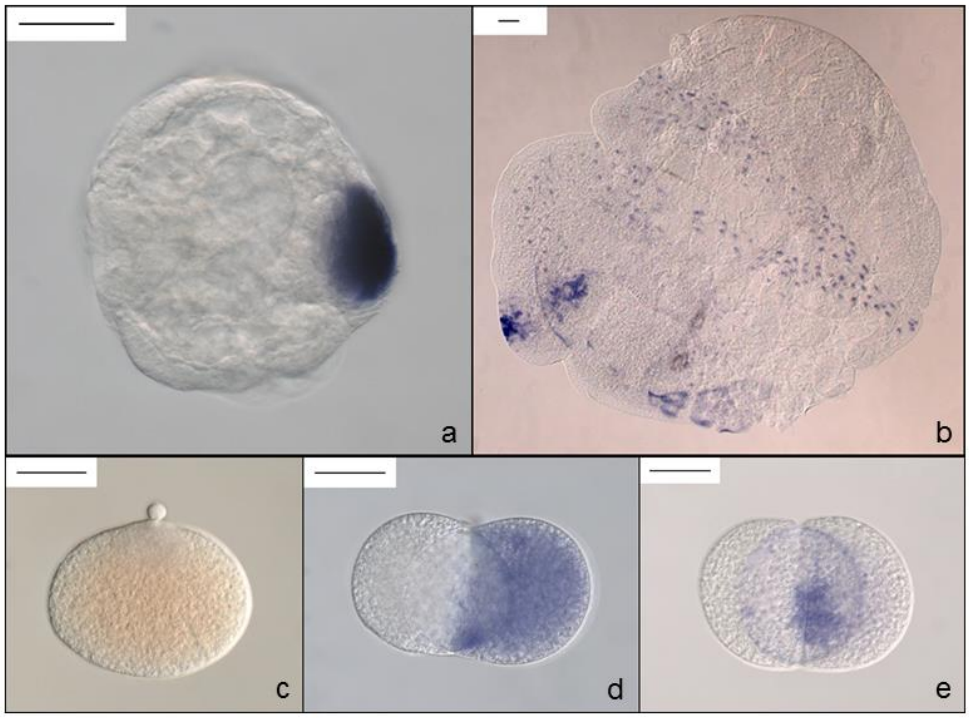


Figure S1. *In situ* expression staining of various positive control genes for appropriate developmental stages of *L.stagnalis*. Anonymous control gene in 1-3 day old embryo (a). Beta tubulin expression in 5-6 day old embryo (b). No signal from anonymous maternal transcript in 1 cell embryo (c). One-sided signal at 2-cell stage from anonymous maternal transcript (d, e). Scale bar represents 50µm in all images.

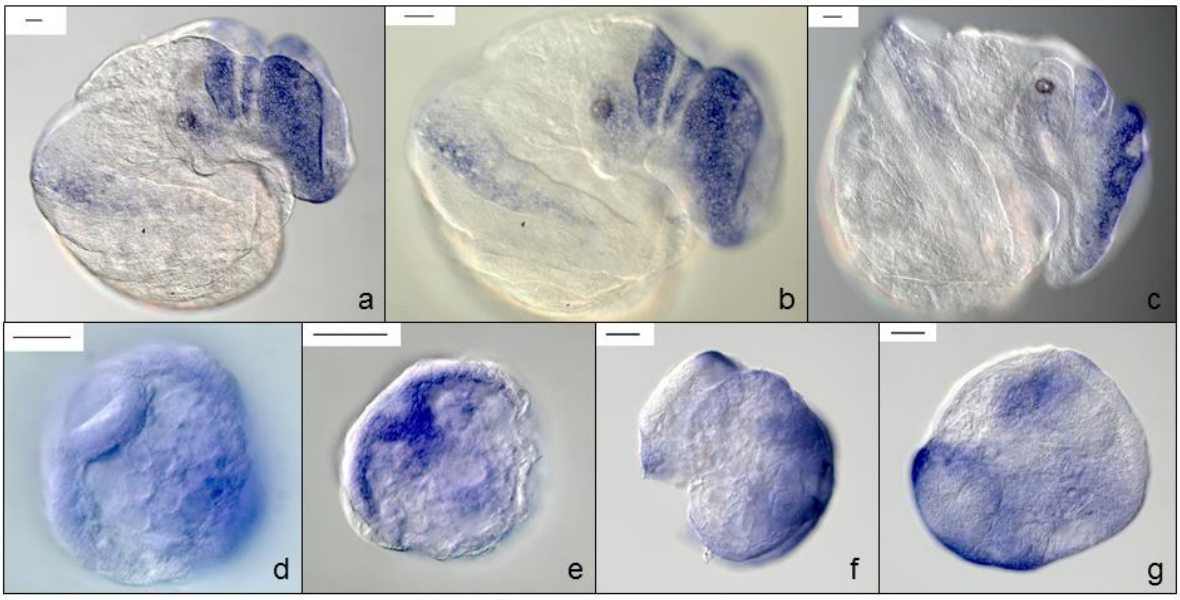


Figure S2. *In situ* expression staining of GOI A in *L.stagnalis* 5-6 day old embryos (a, b, c) and 1-3 day old embryos (d, e, f, g). Scale bar represents 50µm in all images.

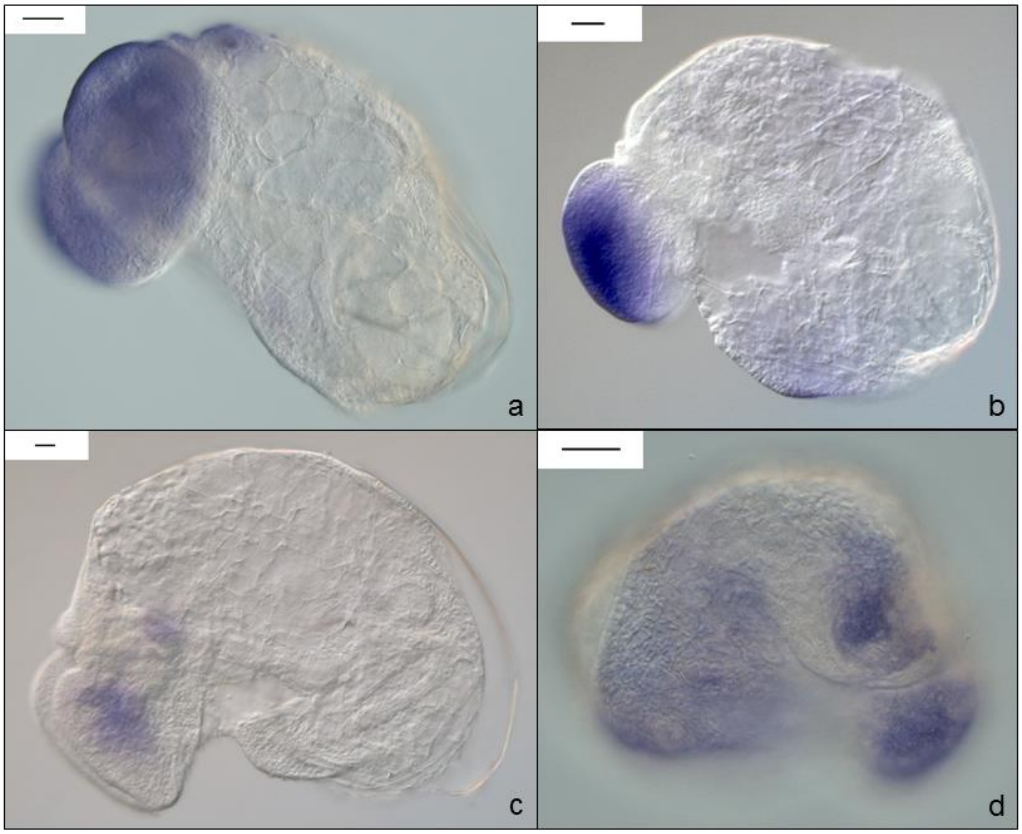


Figure S3. *In situ* expression staining of GOI B in *L.stagnalis* 4-6 day old embryos. Scale bar represents 50µm in all images.

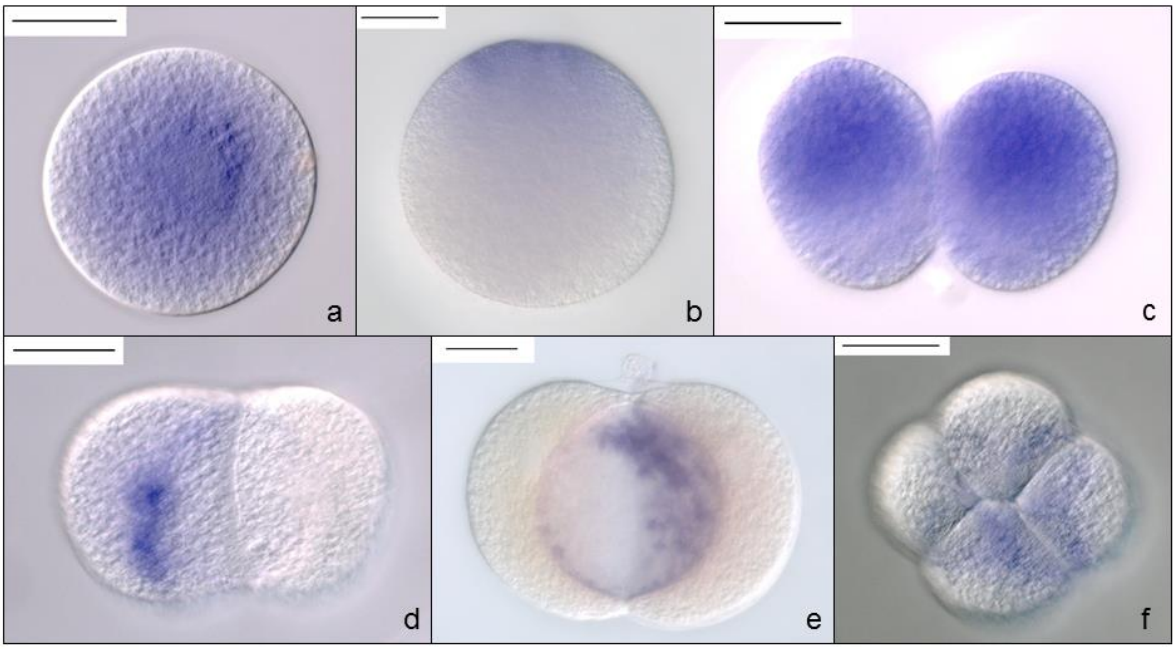


Figure S4. *In situ* expression staining of GOI B in *L.stagnalis* in early cleavage stages; one cell (a, b), early two cell (c), late two cell (d, e) and four cell (f) embryos. Scale bar represents 50µm in all images.

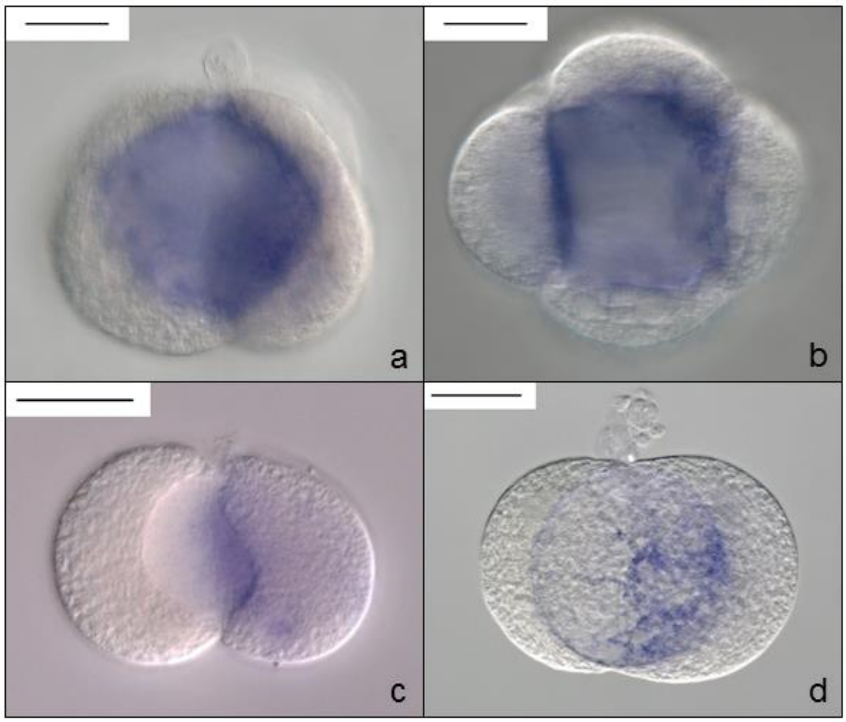


Figure S5. *In situ* expression staining of GOI C in *L.stagnalis* in early cleavage stages; four cell (a, b) & late two cell (c, d). Scale bar represents 50µm in all images.

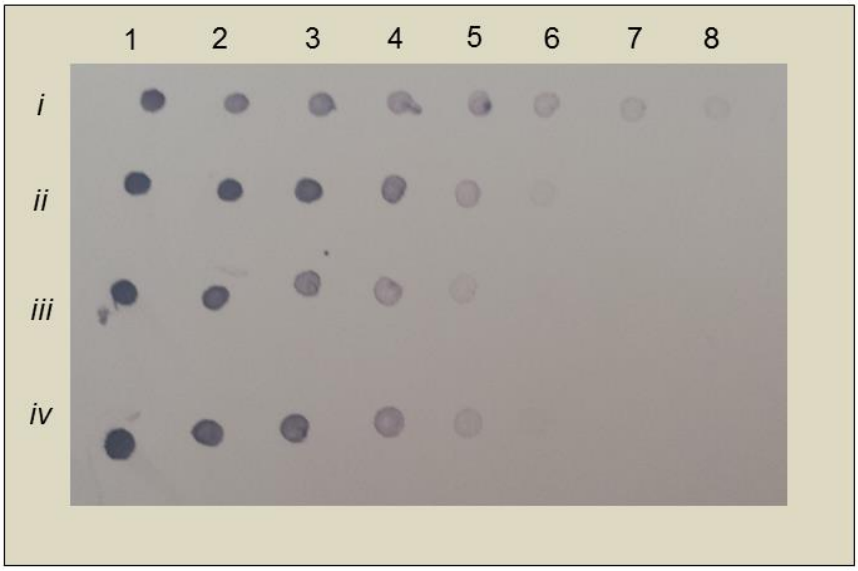


Figure S6. Seven step serial dilutions of riboprobes for GOI C (ii), GOI D (iii) & GOI B (iv) at dilution factor 1:5 indicated by numbers 1 (full concentration) to 8, compared to control DIG-labelled RNA (i) at dilution factor 1:2.

Experimental Controls

In addition the positive controls run in the WMISH experiments, further experimental controls were performed to verify the riboprobe function and attempt to ascertain the reason for the repeated failure in some of the WMISH experiments.

Dot-blot test

Firstly, if the riboprobe has not successfully incorporated the DIG-label required for successful antibody hybridisation and staining, there will be no signal from the WMISH experiment. Equally, failure can occur if the riboprobe is at a too low concentration to be detected.

A 'dot-blot' test was carried out to verify the concentration of the riboprobes (Figure S6). Seven-step serial dilutions at a dilution factor of 1:5 were prepared of each DIG-labelled riboprobe. These were applied to a nylon membrane, treated with the anti-DIG containing hybridisation buffer and finally stained with the NBT/BCIP stain solution. The darker stain indicates a higher riboprobe concentration. The concentration of the riboprobes for the GOIs could then be compared to that of an RNA control (Figure S6). All of the riboprobes tested showed a clear signal at dilutions comparable to the concentrations used in the WMISH experiments (quantified using a *Nanodrop* spectrophotometer) and consequently indicate a functional DIG-label concentration. It can therefore be assumed that this is not the reason for failure in the WMISH experiments.

Northern blot test

Having confirmed the riboprobes contain the label required for successful staining, Northern blot analyses were performed in order to test whether the riboprobes can successfully hybridise to their complimentary RNA transcripts.

Total RNA isolated from three different tissues (buccal mass, foot & mantle) of adult L.stagnalis were each heat denatured and size differentiated via electrophoresis through a formamide-agarose gel (Figure 7, a, b, c). The size-segregated bands of RNA were then transferred onto a nylon membrane to be treated with a specific riboprobe, antibody and finally NBT/BCIP stain, essentially following the WMISH protocol. The resulting stain firstly indicates that the riboprobe is functional and the transcript is present in the sample, but also allows the size of the transcript to be inferred by comparing the distance travelled to that of an RNA marker of known size (Table 1). The size of the transcript can be more accurately predicted by semi-log plotting of the distance travelled relative to a marker of known size (plots available on request).

A previously successful riboprobe was tested on the first Northern blot test to provide a positive control, which produced a strong, discrete banded stain (Figure S7, a*i* (right)). The experimental

riboprobe for GOI B however, produced a much fainter and less distinct stain (Figure S7, a*i* (left)). It should be noted that the stain is in a very common size region for transcripts, as can be seen by the bright smear on the fluorescent gel scan (Figure S7, a, b, c). The stain could be a result of non-specific binding due to the high density of transcripts in this region, or it could be less distinct because the transcript has not been able to segregate effectively due to the cluster of fragments.

No stain was produced for GOI C (Figure S7, b*i* (left)) indicating either that the riboprobe is not functional or that the transcript is not present in the sample. This riboprobe was removed from further experiments and provides a possible explanation for failure in the WMISH, whereas the riboprobe for GOI D (which failed all WMISH experiments) did produce a banded stain, which appears to be too discrete to be background noise, however it is very light (Figure S7, b*i* (right)). This could indicate the transcript is in low abundance and may require an increased concentration of the riboprobe.

Alternative riboprobes for GOI B and GOIC were also tested. The riboprobe for GOI B produced the same smeared stain and as such remains inconclusive (Figure S7, c*i* (left)). The riboprobe for GOI C did produce a discrete but faint stain (Figure S7, c*i* (right)), similar to that produced by GOI D. It is likely therefore that the transcript is present within the sample yet is in low abundance. It must be acknowledged that the final lane in the GOI C analysis contained approximately half the concentration of RNA as the adjacent lanes (Figure S7, c). This was due to having insufficient sample available.

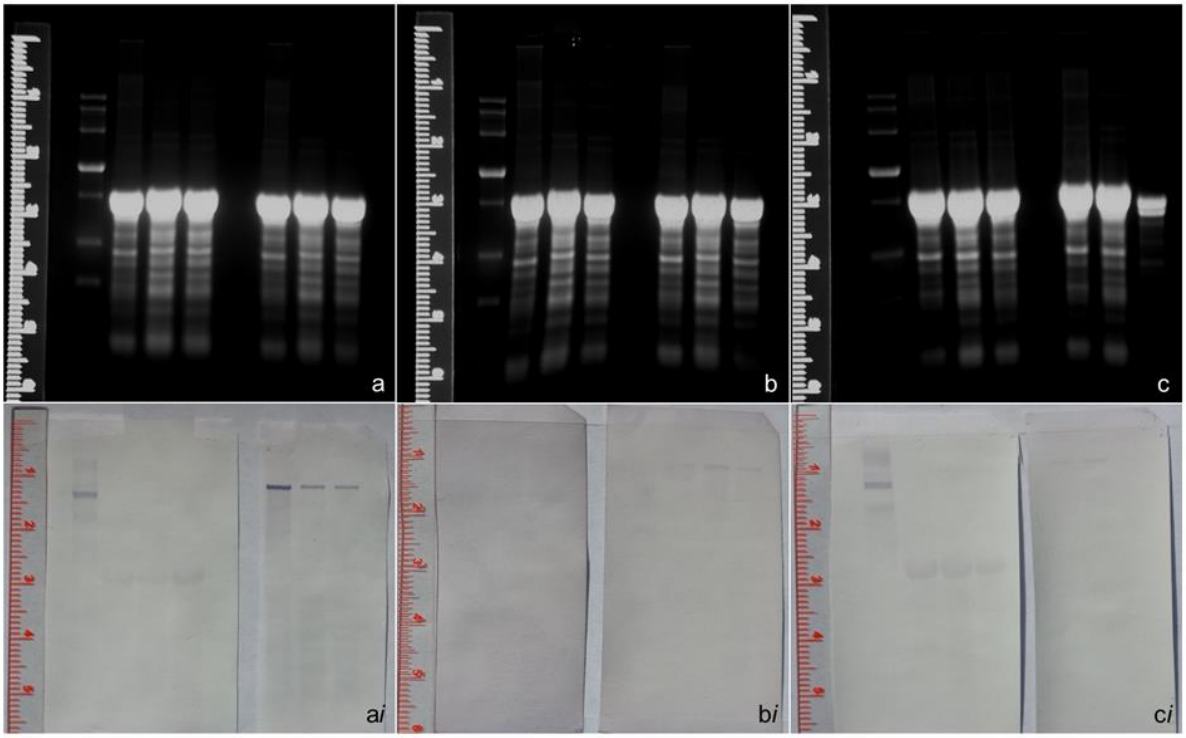


Figure S7. Northern blot analysis: **Top:** a, b & c show size differentiated total RNA from adult *L.stagnalis* tissues (left-right: SS RNA Ladder, buccal mass, foot, mantle, no sample, buccal mass, foot, mantle). **Bottom:** NBT/BCIP staining of; GOI B (left) and positive control (right) (a*i)*; GOI C (left) and GOI D (right) (*bi)*; GOI B (left) and GOI C (right) (c*i)*.

Table S2. Northern blot measures of distance migrated in millimetres (mm) via gel electrophoresis of single stranded RNA of known length (SS RNA Ladder) in kilobases (KB) and riboprobe-specified transcripts (GOI (left) GOI (right)) for each gel depicted in Figure S7 a, b & c.

Gel	a	b	c
SS RNA Ladder (KB)	Distance (mm)	Distance (mm)	Distance (mm)
9	10	9.5	10
7	12	11.5	12.5
5	16	15.5	15.5
3	22.5	22.5	22.8
2	27	27	27.5
1	35	36.5	35.5
0.5	41.8	44.5	43
GOI (left)	29.5	n/a	28.3
GOI (right)	12.5	11	9.2

RNA extraction from embryos

Another possible reason for failure may be that the transcript is not present in the tissue used in the WMISH experiment. The riboprobes were generated from sequences originally obtained from ovotestis tissue from adult *L.stagnalis*. The Northern blot tests were also performed on RNA from adult tissues. As such, it is possible that the specific transcript may not be present in the early stage embryos.

To verify the transcripts are in fact present in the early stage embryos, total RNA was extracted from a pool of one to four cell stage embryos. Complementary DNA (cDNA) was then synthesised from the resulting total RNA using random hexamers. Fragments specific to the riboprobes of GOI B, C & D were successfully amplified via a standard PCR reaction from embryonic and adult tissue (the same adult samples used in the Northern blot analysis, which functioned as a positive control) (Figure S8). It was therefore confirmed that all of the GOIs which failed WMISH experiments were present in the early stage embryos, and that was not the reason for failure.

It is interesting to note that the sequence specific to the riboprobe for GOI C, which failed the Northern blot, displays a very faint band in the embryonic sample, only clearly visible with increased UV exposure (Figure S8, 2). This could indicate that the transcript is in very low abundance in the embryo. It is also important to note that the PCR of the alternative riboprobe sequence GOI C, amplified a different sized fragment in the embryonic sample, compared to that of the adult tissue sample (Figure S8, 4 & 8). When the gel scan is over-exposed, it is apparent that the adult sample has amplified multiple fragments from the same primer pair. This may highlight a true difference in transcripts present in embryonic and adult tissues, although the PCR products must be sequenced to confirm specificity.

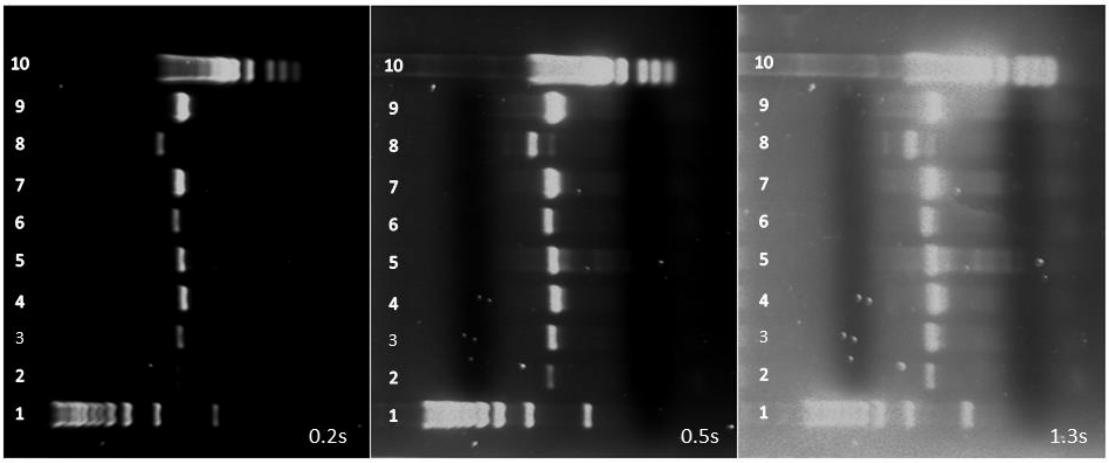


Figure S8. Riboprobe specific cDNA sequences amplified via PCR from embryonic (2-5) and adult (6-9) tissues of *L.stagnalis* at UV exposure 0.2 seconds, 0.5 seconds & 1.3 seconds. 1: 1 Kb ladder. 2 & 6: GOI C. 3 & 7: GOI D. 4 & 8: GOI C (alternative riboprobe). 5 & 9: GOI B. 10: 100bp ladder.

CONCLUSIONS

The results of WMISH can be highly variable and the method requires optimising, which is very time intensive and beyond the scale of this visit. In light of the successful experimental controls, it appears reasons for failure in the early cleavage stages may reside in the fixing technique. When fixing later stage embryos, which demonstrated the expected positive controls (Figure S1, a & b), the samples were gently shaken during incubations, whereas the more sensitive early cleavage stage embryos were not, in the hope to minimise potential damage. It is possible that the viscous ooplasm which surrounds the embryo within the egg may have not been completely removed, resulting in inefficient fixing treatment. Although this seems unlikely due to the extremely well preserved morphology throughout the WMISH experiments (for example see Figure S1, c).

ADDITIONAL BENEFITS

Overall from the research visit, I have gained not only new laboratory expertise but also experience working in a foreign research department, which has been both enlightening and rewarding.

During my time in Göttingen, I was welcomed by all in the department and learned a lot more about the higher education and research system in Germany, including attending the successful PhD defence of one of the students (now a postdoctoral fellow) in the Jackson group. However my time was not only spent working. I was able to enjoy many aspects of the city of Göttingen and take back with me a new love for spätzle and hefeweizen.

REFERENCES

Casey, B. and B. P. Hackett (2000). "Left–right axis malformations in man and mouse." Current opinion in genetics & development **10**(3): 257-261.

S2. Genotyping PCR primers

Chapter 2: PCR: 1315-507

F: 5'-GAGGAGAGGTTTGATTTCATTGAT-3'
R: 5'-CATTCCGCAAACTCTCCATT-3'

Marker RAD04, developed in Liu, Davey et al. (2013)

Chapter 3: PCR: b3g FP1, F8R8

F: 5'-YGGRCCAACATTTATTTYCGTTAC-3'
R: 5'-GTCATGGAMATGGTGCAGAG-3'

Developed in Davison et al, *awaiting publication.* Non-standard bases represent IUPAC ambiguity codes.

S3. Total RNA visual quality assessment

Following each RNA extraction method an aliquot (2 µl) of total RNA was visualised via agarose gel electrophoresis, using ethidium bromide as a fluorescent marker. It has been stated that molluscs (amongst other organisms) do not demonstrate the same sized 18S and 28S rRNA bands commonly found in total RNA of mammals (Barcia, Lopez-Garcia et al. 1997). However, the gels run here were not performed as RNA degradation gels, which are required to specifically size the fragments and therefore the actual size of the bands is unknown. The RNA was examined for quality via presence of distinct bands representing the abundant rRNA (Figure S9). If the sample was degraded these specific sizes would be variable and seen as a smear down the gel. The presence of distinct bands indicated that the samples were of good quality.

A selection of total RNA samples extracted from the ovotestis via TRI Reagent protocol is presented in Figure S9: 1. In order to test the stability of the RNA, an aliquot of each sample was incubated at 65°C for 2 hours, whilst the remaining sample was stored at -80°C (the method of storage employed for all RNA samples). This rather extreme test resulted in overall degradation of the RNA (Figure S9:1b).The RNA extracted from a pooled single-cell embryo sample via the RNeasy kit is shown in Figure S9: 2. Although present at a much lower concentration than those shown from ovotestis, two distinct bands can be seen.

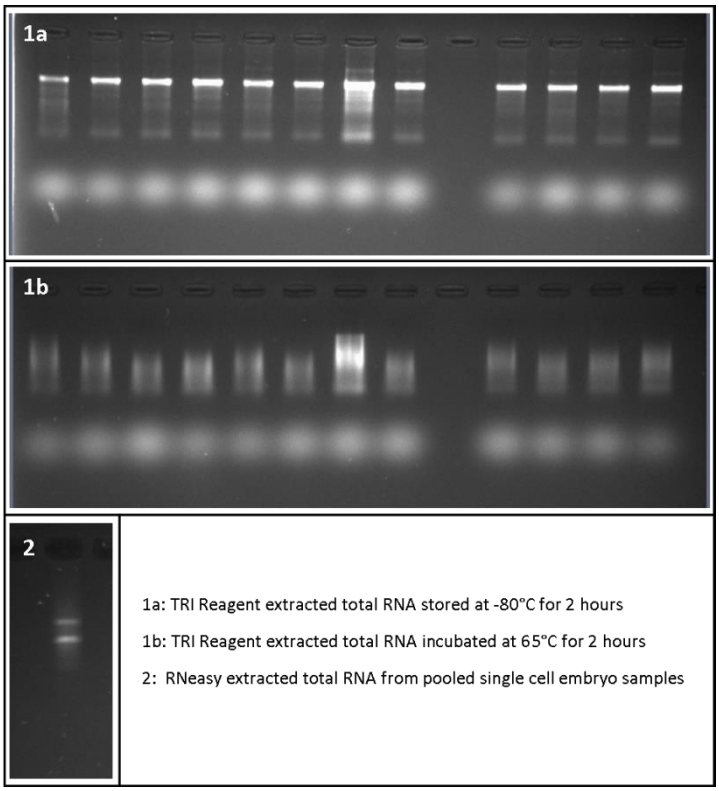


Figure S9 UV visualisation of total RNA via agarose gel electrophoresis. Representative samples are shown from two different methods of RNA extraction and storage conditions.

S4. Predicted expression level of endogenous control genes

Following research in the Davison lab of conserved maternal transcripts across species (Liu, Davey et al. 2014), the genes identified as present in the one cell *L. stagnalis* embryo had information available regarding their associated human housekeeping gene data and expression levels. The expression levels for the six genes selected as candidates for endogenous controls were found to show neither very high nor very low levels of gene expression.

Table S3 Expression level of the human housekeeping gene associated with the six candidate endogenous control genes in L. stagnalis. The sequence description and gene abbreviation (Abv.) are presented in addition to the accession number of each human housekeeping gene NCBI Blast hit and expression level calculated from data published in (Eisenberg and Levanon 2013).

Abv.	Sequence description	Human housekeeping blast hit	Expression level
Lube2	ubiquitin protein	NM_003336	0.467
Lhis2a	histone-like	NM_012412	0.490
Lef1a	elongation factor 1 alpha	NM_006620	0.649
Lywhaz	14-3-3 zeta	NM_006761	0.697
Lacads	acyl-coenzyme a c-2 to c-3 short chain	NM_014049	0.706
Lrpl14	ribosomal protein l14	NM_003973	0.938

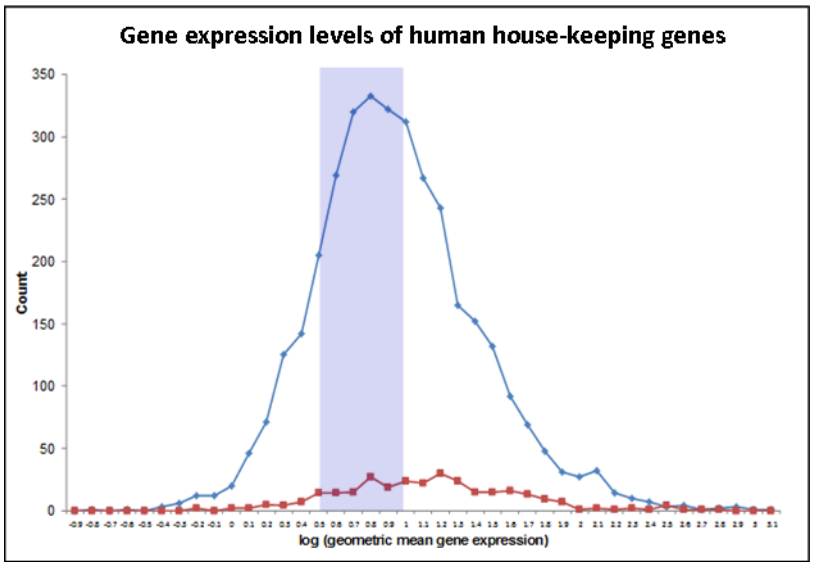


Figure S10 Histogram plot showing log geometric mean gene expression level for conserved (red) and non-conserved (blue) maternal transcripts. The range of expression level of the six candidate genes selected for endogenous controls in *L.stagnalis* is indicated by the shaded area. Human gene expression data originally from (Eisenberg and Levanon 2013).

S5. Sample assessment PCRs:

Two PCRs were used to test samples for genomic DNA carryover. The intron-specific primer sequences are as follows:

CAND F1: 5'- CAAAACCTGGCAATGCTACTG -3'

CAND R1: 5'- ACGTAGGGTTGAAAGTCATGC -3'

PARA F1: 5'- ACCTCTCAGCAACCTTAGGC -3'

PARA F1: 5'- TGAAAGTATCCCAGTCCATGC -3'

All cDNA samples were tested firstly for cDNA amplification to verify the functionality of the cDNA. Following the positive control for cDNA amplification, another PCR tested for the presence of genomic carryover by attempting to amplify a sequence specific to the intron.

S5.1 Ovotestis

Each ovotestis sample amplified a bright product in the intronic PCR experiment indicating the presence of genomic DNA carryover, however no sample amplified two products in the qPCR primers test and so any genomic contamination is believed to be outcompeted by the more abundant cDNA template. There was a notable exception in sample 10629, where no band was detectable in the intronic PCR, however this is assumed to be due to loading error. Amplification was observed in the negative control of the *Lmhc* reaction in Figure S11.2a, however there were no issues of cross-contamination apparent in the qPCR experiments. Therefore this is assumed to be a result of the specific primer dilution or water aliquot used in this individual PCR.

Sample 10630 and 10631 were not included in the PCR experiment shown in Figure S11. However positive amplification of the samples had already been observed, although 10631 did suffer from what is assumed to be low concentration, resulting in a failure to amplify in a number of experiments. Due to the unanimous presence of genomic carryover in the ovotestis samples it is assumed that these two samples also had a similar level of genomic carryover.

S5.2 Foot

Each foot sample amplified a product in the intronic PCR experiment indicating the presence of genomic DNA carryover, however the bands appeared less bright than that of the genomic control sample (Figure S12.b) and compared to those in the ovotestis (Figure S11.2b, although it should be noted that these are two separate gels and therefore have some level of variation in the UV exposure and amount of ethidium bromide included and as such are not a direct comparison).

Again no sample amplified two products in the qPCR primers test and so any genomic contamination is believed to be outcompeted by the more abundant cDNA template. There was a notable exception in sample 11352, where no band was detectable in the intronic PCR. Although this could represent a complete absence of genomic carryover in this sample, it is more likely a consequence of loading error or PCR failure.

S5.3 Embryo

Each embryo sample PCR is presented in Figure S13. No sample produced a PCR product detectable via UV gel electrophoresis (gels were examined for longer UV exposure time to ensure no product was present – images not presented). The sinistral samples (11282, 11284, 11287, 11283, 11301, and 11303) showed notably fainter bands for the amplification of *Ldia2*.

Amplification was again observed in the negative control in *Lmhc* experiment shown in Figure S13.3a, again due to the lack of issues of cross-contamination apparent in the qPCR experiments and the same problem observed in the ovotestis and foot *Lmhc* test PCR, this is assumed to be a result of the specific primer dilution or water aliquot used. A band was also seen in the negative control of Figure S13.2b, due to the lack of any bands in the other wells this is assumed to be due to well crossover within the gel loading as opposed to contamination within the reaction.

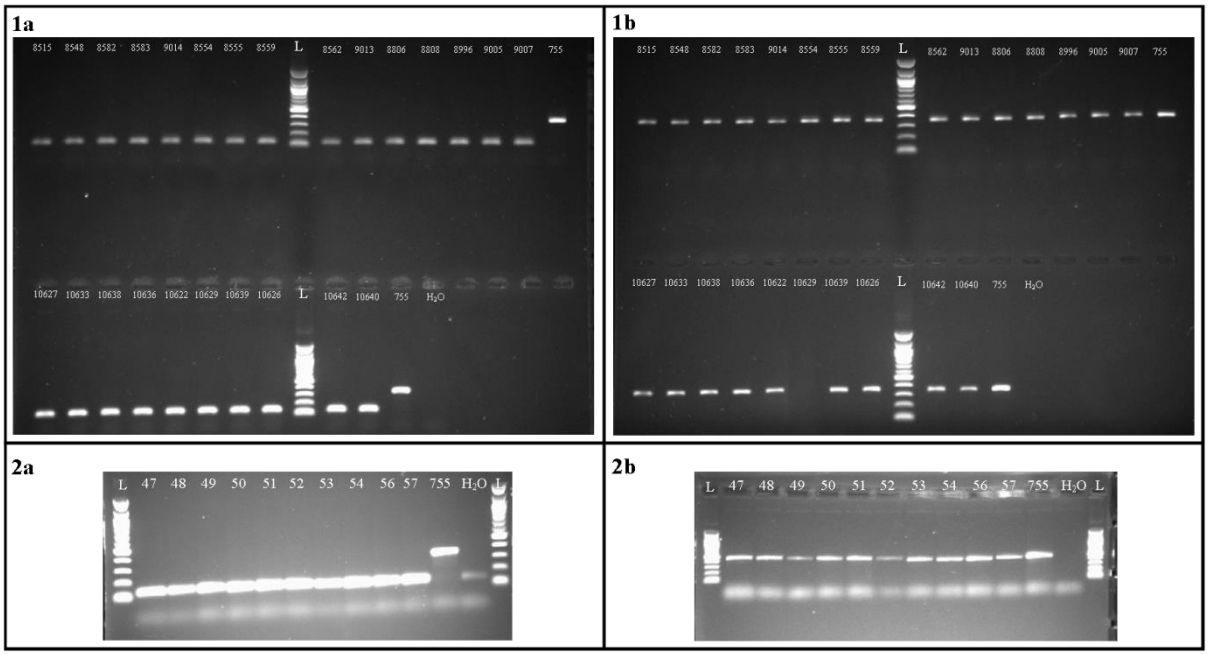


Figure S11 UV visualisation via agarose gel electrophoresis of PCR products amplified from each of the ovotestis cDNA samples with the exception of (10630, 10631) a genomic control sample (755) and a negative control (H_2O), specific to *Lmhc* qPCR target (1a, 2a) and intron specific target (1b, 2b) and 100 bp DNA size marker (L).

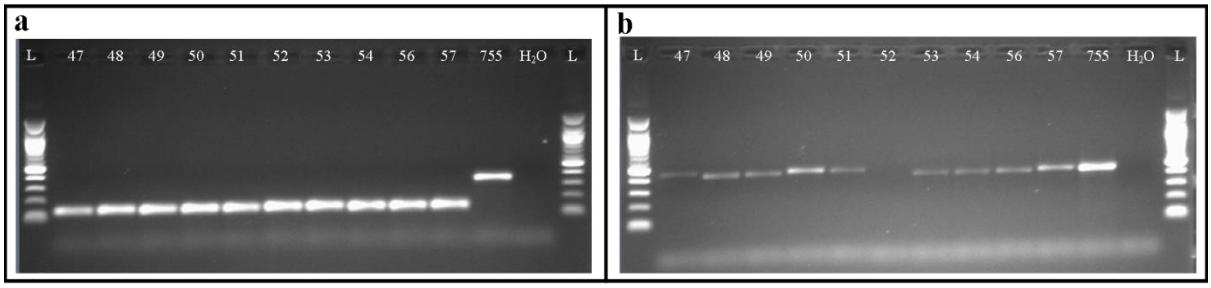


Figure S12 UV visualisation via agarose gel electrophoresis of PCR products amplified from each of the foot cDNA samples with an additional genomic control sample (755) and negative control (H2O), specific to *Lmhc* qPCR target (a,) and intron specific target (b) and 100 bp DNA size marker (L).

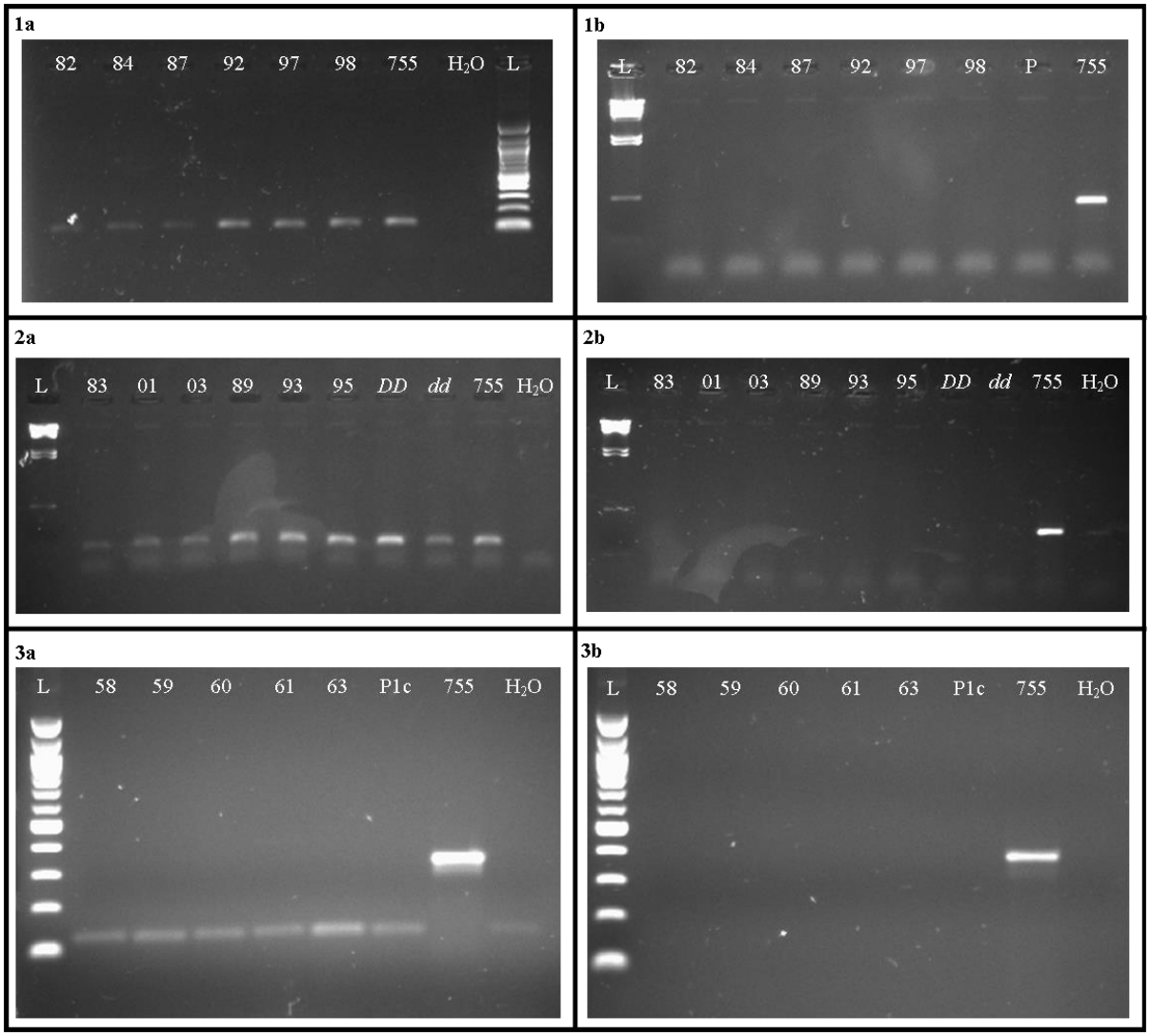


Figure S13 UV visualisation via agarose gel electrophoresis of PCR products amplified from each of the embryo cDNA samples (11282-11303; 11358-11363, including the one cell pooled sample P1c), a genomic control sample (755) and a negative control (H_2O), specific to *Ldia2* (1a, 2a) and *Lmhc* (3a) qPCR targets and intron specific targets (1b, 2b, 3b) and 100 bp DNA size marker (L).

S5.4 DNase treatments across tissues

Every embryo tissue sample used within these experiments failed to produce a PCR product from the intronic specific PCR whereas each of the ovotestis and foot samples exhibited genomic contamination via the amplification of intron-specific PCR products, with the exception of one foot sample 11352, and one ovotestis sample 10629. However due to the intensity of the bands in the other samples, it is more likely that the absence of amplification represents a PCR fail as opposed to a genuine lack of genomic DNA within the sample.

The failure of the DNaseI treatment to remove all genomic DNA from the foot tissue yet succeeded in the embryo tissue, is of note. One key difference between the samples is their extraction method. The embryos were extracted via the RNeasy micro kit, having been stored in RNA later, whereas the foot tissue samples were firstly extracted using TRI Reagent® and subsequently re-extracted using the RNeasy kit. The TRI Reagent® RNA extraction method used to isolate the foot tissue RNA samples will inevitably to result in some level of phenol carryover and potential ethanol carryover. The level of such impurities can be inferred from the nanodrop 260/230 values (Table 2 – main document).

Although some of the foot samples 260/230 values are below the recommended 1.8 value (ThermoScientific 2010) the embryo sample values are generally far lower. It may be more informative to look at the Nanodrop data of the foot samples prior to the re-extraction. These are presented in Table S4 and show similarly, the 260/230 values are often lower than the recommended values yet not as low as those seen in the embryo samples. It is important to note that the values for the embryo samples may be skewed by the generally much lower RNA concentration; due to the relative measure of the spectrophotometer peaks, the same amount of carry over impurities will impact a low RNA concentration sample more than a high RNA concentration sample.

The DNase treatment within the RNeasy extraction protocol has a limited loading capacity (although this was not exceeded according to the user manual, (Qiagen 2007)) which may have limited its effectiveness in the foot samples. It is probable, therefore that the absence of genomic carryover in the embryo samples may simply be due to the lower starting quantity.

It is acknowledged by the suppliers of DNase treatments that no DNase is capable of removing all traces of genomic DNA. However, the failure of two alternative DNase treatments to remove carryover genomic DNA from ten of the ovotestis samples is somewhat surprising (Figure S11). The failure may be due to residual phenol from the TRI Reagent protocol inhibiting the active enzyme.

The generally low 260/230 ratios seen in the ovotestis samples (Table 17 – main document) supports this (AppliedBiosystems 2010).

Finally, the low concentration of the embryo samples may be the sole reason for their apparent lack of genomic carryover. As opposed to the samples genuinely being purer than the foot or ovotestis samples, it may be that the starting material is reduced such that the level genomic carryover is insufficient to generate a PCR product detectable via gel electrophoresis.

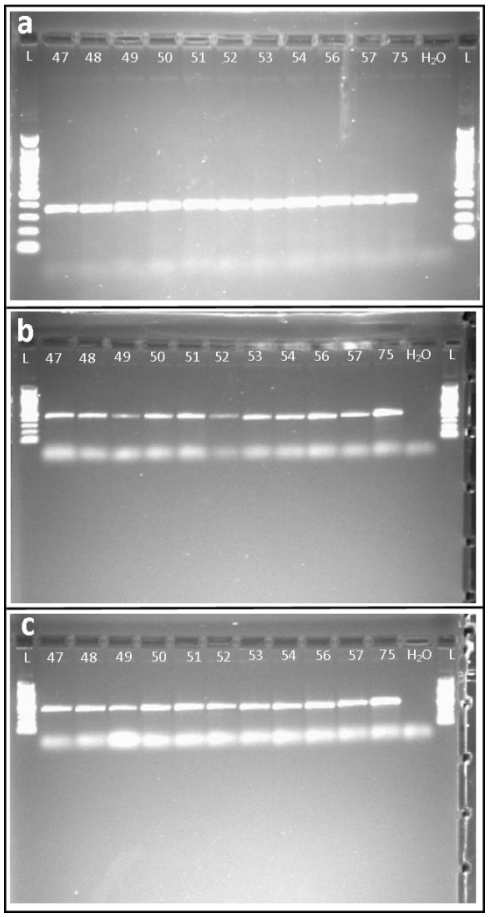


Figure S14 Intronic PCR test of DNase treated ovotestis samples 11347-11357, with genomic DNA positive control (755) and negative control (H_2O). PCR products amplified from cDNA generated fromTRI Reagent extracted total RNA (a), DNA-free™ (Ambion) treated TRI Reagent extracted total RNA (b) and Precision DNase (PrimerDesign) treated following DNA-free™ treated TRI Reagent extracted total RNA (c).

Table S4 Summary of RNeasy re-extraction of foot samples. Table shows: Sample descriptive information including sample ID (ID) and genotype (Geno). A summary of the Nanodrop quantification of the untreated TRI Reagent extracted total RNA. A summary of the volume (µl added) and mass (µg added) of total RNA added to the RNeasy extraction protocol. Nanodrop data is presented for the following RNeasy re-extracted total RNA and a final summary of the µg of total RNA eluted into 14 µl, and the percentage retained of the RNA input (% ret.). Finally the difference between the 260/280 and 260/230 ratios from the re-extracted sample compared to the original sample are presented (Δ260/280, Δ260/230). A positive value indicates the ratios are higher in the re-extracted samples.

Sample Description		Untreated total RNA			RNeasy extraction		RNeasy re-extracted			Efficiency			
ID	Geno	ng/µl	260/280	260/230	µl added	µg added	ng/µl	260/280	260/230	µg eluted (14µl)	% ret.	Δ260/280	Δ260/230
11355	*dd*	194.24	1.96	1.40	25.0	4.86	207.17	2.08	1.52	2.90	59.73	0.12	0.12
11355	*dd*	194.24	1.96	1.40	5.0	0.97	24.89	2.10	1.23	0.35	35.88	0.14	-0.17
11355	*dd*	194.24	1.96	1.40	10.0	1.94	56.77	2.31	1.16	0.79	40.92	0.35	-0.24
11347	*DD*	379.55	1.95	2.06	5.3	2.00	72.02	1.95	1.40	1.01	50.41	0.00	-0.66
11348	*dd*	335.73	1.92	2.22	6.0	2.00	74.69	2.03	2.12	1.05	52.28	0.11	-0.10
11349	*dd*	415.29	1.91	2.27	4.8	2.00	70.75	2.05	2.03	0.99	49.53	0.14	-0.24
11350	*DD*	296.31	1.98	1.12	6.7	2.00	49.54	1.96	1.51	0.69	34.68	-0.02	0.39
11351	*DD*	254.38	1.97	1.55	7.9	2.00	85.33	2.19	2.26	1.19	59.73	0.22	0.71
11352	*DD*	552.47	2.01	1.41	3.6	2.00	67.62	2.10	1.95	0.95	47.33	0.09	0.54
11353	*dd*	511.76	1.99	1.82	3.9	2.00	69.20	2.41	1.36	0.97	48.44	0.42	-0.46
11354	*dd*	646.13	1.99	2.15	3.1	2.00	78.20	2.07	1.72	1.09	54.74	0.08	-0.43
11356	*dd*	234.32	1.92	1.84	8.5	2.00	76.98	2.09	1.98	1.08	53.89	0.17	0.14
11357	*DD*	255.02	1.95	1.35	7.8	2.00	62.15	2.23	1.25	0.87	43.51	0.28	-0.10

S6. Description of omitted data points from qPCR raw data

Chapter 2: Endogenous control gene experiments

To reduce experimental noise in the final dataset, data points resulting in an average Cq value with high SD were omitted from the final analyses. All bar one of the omitted data points were from the Lef1a assessment. In the embryo analysis, samples 11295, 11287 and 11303 each exhibited high SD for the amplification of Lef1a. One perceived outlier of each sample was omitted from analysis, resulting in an average Cq value calculated from only two data points without high SD (Table 7). In the foot tissue analysis, perceived outliers were removed from five samples which exhibited high SD in Lef1a, namely 11348, 11350, 11351, 11352, 11357. In sample 11347, two replicates failed to amplify a product for Lef1a and therefore only one Cq value contributed to the average Cq value, and as such it does not strictly represent an average Cq (Table 8). In the ovotestis analysis, one of the technical repeats for sample 10642 in Lef1a was flagged as an outlier and subsequently removed from the final dataset (Table 9).

In the Lrpl14 analysis, one of the technical repeats for embryo sample 11292 was flagged as an outlier and subsequently removed from analysis (Table 7).

S7. Sanger sequencing protocol

No cloning was performed in order to sequence the qPCR amplification products. Sanger sequencing initiated from both the forward and reverse primers enabled sufficient capture of the full qPCR amplicon sequence. Sanger sequencing was performed on pooled single cell embryo samples. The six endogenous control gene amplicons were sequenced from a pool of Dd samples. *Ldia1* 3'UTR and *Ldia2* 3'UTR were sequenced from a pool of DD samples, whereas *Ldia2* ORF was sequenced from a pool of DD samples and another of dd samples to ensure that the same transcript was being amplified from both genotypes.

The qPCR amplicons were amplified via standard non-quantitative PCR as described in Box 2 (main document) using the specified primer pair (10 mM) and 3 µl cDNA (1:30 dilution). PCR products were then cleaned using exonuclease SAP (shrimp alkaline phosphatase) protocol to remove leftover primers and dNTPs. 1 µl SAP (NEB), 0.15 µl Exo I (NEB) and 1.35 µl 10x buffer (provided with Exo I) was added to each 20 µl PCR reaction and incubated at 37 °C for 40 minutes followed by heat inactivation at 80 °C for 15 minutes.

Cleaned PCR products were prepared for Sanger sequencing using the BigDye® Terminator kit v.3.1 (ThermoFisher Scientific). 3 µl of PCR product was added to 1 µl Big Dye, 2 µl 5x buffer, 0.5 µl specific primer, each reaction was made up to 10 µl with PCR grade water. 6 µl of the *Ldia2 ORF, dd* template PCR product was added to the sequencing reaction to accommodate the low quantity transcript. Samples were then incubated for 25 cycles of 95 °C for 30 seconds; 50 °C for 20 seconds and 60 °C for 3 minutes. Prepared templates were then set to The GenePool (University of Edinburgh) to be sequenced.

S8. qPCR analysis; summary statistics

S8.1 Genotype Analysis

S8.1.1 LOGNRQ, embryo, genotypic groups

```
> by(LOGNRQembryo, LOGNRQembryo$Geno, summary)
LOGNRQembryo$Geno: dd
      Ldia2_3'            Ldia1_3'            Ldia2_ORF              Lfry              Larp2/3 1a
 Min.   :-2.789      Min.   :-0.07757    Min.   :-1.748      Min.   :-0.078063    Min.   :-0.023905
 1st Qu.:-2.376      1st Qu.: 0.02273    1st Qu.:-1.693      1st Qu.:-0.040849    1st Qu.:-0.022077
 Median :-2.023      Median : 0.07571    Median :-1.519      Median :-0.018880    Median :-0.014248
 Mean   :-2.198      Mean   : 0.05379    Mean   :-1.539      Mean   :-0.028470    Mean   :-0.006666
 3rd Qu.:-1.999      3rd Qu.: 0.10550    3rd Qu.:-1.407      3rd Qu.:-0.009005    3rd Qu.: 0.008762
 Max.   :-1.869      Max.   : 0.12841    Max.   :-1.326      Max.   :-0.001114    Max.   : 0.020663
      Lfat1               Larp2/3 3
 Min.   :-0.0565351  Min.   :-0.119148
 1st Qu.:-0.0219427  1st Qu.:-0.109192
 Median :-0.0095018  Median :-0.096031
 Mean   :-0.0004645  Mean   :-0.074226
 3rd Qu.: 0.0175076  3rd Qu.:-0.035321
 Max.   : 0.0723309  Max.   :-0.005485
------------------------------------------------------------------------------
LOGembryo$Geno: Dd
 Geno
      Ldia2_3'            Ldia1_3'            Ldia2_ORF              Lfry              Larp2/3 1a
 Min.   :-0.3551     Min.   :-0.090559   Min.   :-0.3830     Min.   :-0.089078    Min.   :-0.00392
 1st Qu.:-0.2616     1st Qu.:-0.021767   1st Qu.:-0.3721     1st Qu.:-0.013292    1st Qu.: 0.01623
 Median :-0.2434     Median :-0.002141   Median :-0.3528     Median :-0.008009    Median : 0.03867
 Mean   :-0.2565     Mean   :-0.005565   Mean   :-0.3211     Mean   :-0.017515    Mean   : 0.05181
 3rd Qu.:-0.2153     3rd Qu.: 0.022330   3rd Qu.:-0.2995     3rd Qu.:-0.005204    3rd Qu.: 0.09862
 Max.   :-0.2070     Max.   : 0.064314   Max.   :-0.1979     Max.   : 0.028010    Max.   : 0.10946
      Lfat1               Larp2/3 3
 Min.   :-0.081742   Min.   :-0.2602
 1st Qu.:-0.035784   1st Qu.:-0.2414
 Median :-0.015828   Median :-0.2001
 Mean   :-0.020851   Mean   :-0.1901
 3rd Qu.:-0.008401   3rd Qu.:-0.1453
 Max.   : 0.037499   Max.   :-0.1036
------------------------------------------------------------------------------
LOGNRQembryo$Geno: DD
      Ldia2_3'            Ldia1_3'            Ldia2_ORF              Lfry
 Min.   :-0.091723   Min.   :-0.05367    Min.   :-0.0648373   Min.   :-0.07480
 1st Qu.:-0.015450   1st Qu.:-0.03565    1st Qu.:-0.0404973   1st Qu.:-0.02558
 Median :-0.003644   Median :-0.01585    Median :-0.0068293   Median :-0.00785
 Mean   :-0.006848   Mean   : 0.00806    Mean   : 0.0008603   Mean   :-0.01058
 3rd Qu.: 0.002166   3rd Qu.: 0.01701    3rd Qu.: 0.0494676   3rd Qu.: 0.01708
 Max.   : 0.073208   Max.   : 0.14803    Max.   : 0.0671442   Max.   : 0.03314
  Larp2/3 1a              Lfat1               Larp2/3 3
 Min.   :-0.004734   Min.   :-0.153447   Min.   :-0.11530
 1st Qu.: 0.001232   1st Qu.:-0.053768   1st Qu.:-0.08315
 Median : 0.018875   Median :-0.024267   Median :-0.05446
 Mean   : 0.026872   Mean   :-0.042658   Mean   :-0.05674
 3rd Qu.: 0.053201   3rd Qu.:-0.003911   3rd Qu.:-0.03122
 Max.   : 0.068219   Max.   : 0.006761   Max.   : 0.00000
```

S8.1.2 LOGNRQ, foot, genotypic groups

```
> by(LOGNRQfoot, LOGNRQfoot$Geno, summary)
LOGNRQfoot$Geno: dd
  Ldia2 ORF           Ldia2 3'            Ldia1 3'            Lfry              Larp2/3 1a
 Min.   :0.2114      Min.   :0.2160      Min.   :-0.047015   Min.   :0.07089   Min.   :0.1611
 1st Qu.:0.2478      1st Qu.:0.2228      1st Qu.:-0.039590   1st Qu.:0.15546   1st Qu.:0.2133
 Median :0.3807      Median :0.2625      Median :-0.034053   Median :0.16817   Median :0.3138
 Mean   :0.3649      Mean   :0.3064      Mean   :-0.007333   Mean   :0.18609   Mean   :0.2641
 3rd Qu.:0.4852      3rd Qu.:0.4141      3rd Qu.: 0.015022   3rd Qu.:0.23435   3rd Qu.:0.3153
 Max.   :0.4994      Max.   :0.4164      Max.   : 0.068972   Max.   :0.30159   Max.   :0.3170
     Lfat1               Larp2/3 3           Lmhc                Lcol11a 2/1         Lmhc nm
 Min.   :0.1245      Min.   :-0.018402   Min.   :0.02748     Min.   :0.4984      Min.   :-0.17144
 1st Qu.:0.2239      1st Qu.:-0.014050   1st Qu.:0.26669     1st Qu.:0.5224      1st Qu.: 0.04490
 Median :0.2777      Median :-0.013202   Median :0.33842     Median :0.7056      Median : 0.06835
 Mean   :0.2548      Mean   :-0.003229   Mean   :0.30067     Mean   :0.6539      Mean   : 0.05351
 3rd Qu.:0.3188      3rd Qu.: 0.013299   3rd Qu.:0.41177     3rd Qu.:0.7553      3rd Qu.: 0.13757
 Max.   :0.3293      Max.   : 0.016212   Max.   :0.45900     Max.   :0.7877      Max.   : 0.18817
      Lmyo5a              Lstau               Lmyo18a             Lunc93a
 Min.   :0.1315      Min.   :0.1075      Min.   :0.1858      Min.   :0.3639
 1st Qu.:0.3270      1st Qu.:0.2238      1st Qu.:0.2402      1st Qu.:0.6853
 Median :0.4161      Median :0.4010      Median :0.2926      Median :0.8084
 Mean   :0.3961      Mean   :0.3158      Mean   :0.3403      Mean   :0.7177
 3rd Qu.:0.5255      3rd Qu.:0.4195      3rd Qu.:0.4684      3rd Qu.:0.8597
 Max.   :0.5807      Max.   :0.4269      Max.   :0.5146      Max.   :0.8710
------------------------------------------------------------------------------
LOGNRQfoot$Geno: DD
  Ldia2 ORF           Ldia2 3'            Ldia1 3'            Lfry              Larp2/3 1a
 Min.   :0.00000     Min.   :0.00000     Min.   :-0.09353    Min.   :-0.007567 Min.   :0.0000
 1st Qu.:0.04776     1st Qu.:0.07302     1st Qu.:-0.04504    1st Qu.: 0.000000 1st Qu.:0.1126
 Median :0.09369     Median :0.20973     Median : 0.00000    Median : 0.091471 Median :0.1373
 Mean   :0.21100     Mean   :0.24431     Mean   : 0.00434    Mean   : 0.168440 Mean   :0.1785
 3rd Qu.:0.28648     3rd Qu.:0.30587     3rd Qu.: 0.06549    3rd Qu.: 0.309935 3rd Qu.:0.2797
 Max.   :0.62706     Max.   :0.63292     Max.   : 0.09479    Max.   : 0.448362 Max.   :0.3628
     Lfat1               Larp2/3 3           Lmhc                Lcol11a 2/1         Lmhc nm
 Min.   :-0.008095   Min.   :-0.071342   Min.   :-0.002715   Min.   :0.0000      Min.   :-0.168247
 1st Qu.: 0.000000   1st Qu.:-0.031524   1st Qu.: 0.000000   1st Qu.:0.1988      1st Qu.:-0.002348
 Median : 0.119185   Median :-0.021033   Median : 0.007698   Median :0.3342      Median : 0.000000
 Mean   : 0.169448   Mean   :-0.025690   Mean   : 0.149604   Mean   :0.3814      Mean   : 0.096722
 3rd Qu.: 0.308724   3rd Qu.:-0.004552   3rd Qu.: 0.353810   3rd Qu.:0.6800      3rd Qu.: 0.321287
 Max.   : 0.427424   Max.   : 0.000000   Max.   : 0.389229   Max.   :0.6942      Max.   : 0.332917
      Lmyo5a              Lstau               Lmyo18a             Lunc93a
 Min.   :0.00000     Min.   :0.00000     Min.   :-0.2125     Min.   :0.0000
 1st Qu.:0.08935     1st Qu.:0.06114     1st Qu.: 0.0000     1st Qu.:0.3515
 Median :0.13919     Median :0.09675     Median : 0.2454     Median :0.4041
 Mean   :0.30580     Mean   :0.24688     Mean   : 0.1756     Mean   :0.5460
 3rd Qu.:0.56861     3rd Qu.:0.45755     3rd Qu.: 0.3238     3rd Qu.:0.7402
 Max.   :0.73186     Max.   :0.61893     Max.   : 0.5211     Max.   :1.2340
```

S8.1.3 LOGNRQ, ovotestis, genotypic groups

```
> by(LOGNRQovo, LOGNRQovo$Geno, summary)
LOGovoNRQ$Geno: dd
      Larp2/3 1a          Lmhc              Lmyo5a            Lstau
 Min.   :-0.34974   Min.   :-0.55018   Min.   :-0.33487   Min.   :-0.24599
 1st Qu.:-0.12548   1st Qu.:-0.35939   1st Qu.:-0.20314   1st Qu.: 0.04934
 Median :-0.04604   Median :-0.13875   Median :-0.02504   Median : 0.11825
 Mean   :-0.01315   Mean   :-0.15845   Mean   :-0.05647   Mean   : 0.17579
 3rd Qu.: 0.11142   3rd Qu.:-0.01496   3rd Qu.: 0.06165   3rd Qu.: 0.29631
 Max.   : 0.31669   Max.   : 0.45416   Max.   : 0.19353   Max.   : 0.62779
     Lcol11a 2/1        Lmyo18a            Lunc93a           Ldia2 3'          Ldia1 3'
 Min.   :-0.27625   Min.   :-0.7414   Min.   :-0.9525   Min.   :-0.38807   Min.   :-0.19039
 1st Qu.: 0.07487   1st Qu.:-0.4134   1st Qu.:-0.3542   1st Qu.:-0.26730   1st Qu.:-0.05385
 Median : 0.21484   Median :-0.3571   Median :-0.2561   Median :-0.08594   Median : 0.06249
 Mean   : 0.23819   Mean   :-0.3496   Mean   :-0.3274   Mean   :-0.09879   Mean   : 0.08123
 3rd Qu.: 0.35203   3rd Qu.:-0.2103   3rd Qu.:-0.2368   3rd Qu.: 0.04130   3rd Qu.: 0.23552
 Max.   : 1.02542   Max.   :-0.1546   Max.   : 0.1275   Max.   : 0.26436   Max.   : 0.31009
      Lfat1            Ldia2 ORF            Lfry            Larp2/3 3           Lmhc nm
 Min.   :-0.49161   Min.   :-0.44863   Min.   :-0.471614   Min.   :-0.353702   Min.   :-0.22548
 1st Qu.:-0.33428   1st Qu.:-0.31353   1st Qu.:-0.147309   1st Qu.:-0.266463   1st Qu.:-0.17633
 Median :-0.25492   Median :-0.18641   Median :-0.030168   Median :-0.154449   Median :-0.08901
 Mean   :-0.24353   Mean   :-0.18286   Mean   :-0.090992   Mean   :-0.146146   Mean   :-0.08616
 3rd Qu.:-0.13020   3rd Qu.:-0.02540   3rd Qu.: 0.003432   3rd Qu.:-0.003215   3rd Qu.:-0.02188
 Max.   :-0.04637   Max.   : 0.03727   Max.   : 0.071142   Max.   : 0.078576   Max.   : 0.08226
------------------------------------------------------------------------------
LOGNRQovo$Geno: Dd
      Larp2/3 1a          Lmhc              Lmyo5a            Lstau            Lcol11a 2/1
 Min.   :-0.19441   Min.   :-0.56082   Min.   :-0.22185   Min.   :-0.03999   Min.   :-0.21783
 1st Qu.:-0.08710   1st Qu.:-0.37694   1st Qu.:-0.11635   1st Qu.: 0.13694   1st Qu.:-0.15386
 Median : 0.04981   Median :-0.10929   Median :-0.04375   Median : 0.21550   Median :-0.01230
 Mean   : 0.03833   Mean   :-0.15627   Mean   :-0.04850   Mean   : 0.30682   Mean   : 0.04433
 3rd Qu.: 0.12508   3rd Qu.: 0.09248   3rd Qu.: 0.02210   3rd Qu.: 0.39616   3rd Qu.: 0.18630
 Max.   : 0.27593   Max.   : 0.14750   Max.   : 0.14001   Max.   : 0.93284   Max.   : 0.41413
     Lmyo18a            Lunc93a           Ldia2 3'          Ldia1 3'            Lfat1
 Min.   :-0.46542   Min.   :-0.943373   Min.   :-0.25109   Min.   :-0.3829677   Min.   :-0.47937
 1st Qu.:-0.40222   1st Qu.:-0.343708   1st Qu.:-0.07225   1st Qu.:-0.0852932   1st Qu.:-0.39524
 Median :-0.23273   Median :-0.273670   Median : 0.05402   Median : 0.0749014   Median :-0.25533
 Mean   :-0.23387   Mean   :-0.312446   Mean   : 0.01320   Mean   : 0.0001706   Mean   :-0.25076
 3rd Qu.:-0.11718   3rd Qu.:-0.181672   3rd Qu.: 0.11179   3rd Qu.: 0.1272953   3rd Qu.:-0.13027
 Max.   : 0.09513   Max.   : 0.000555   Max.   : 0.25367   Max.   : 0.2272405   Max.   :-0.01427
    Ldia2 ORF            Lfry            Larp2/3 3           Lmhc nm
 Min.   :-0.44784   Min.   :-0.29791   Min.   :-0.681697   Min.   :-0.26647
 1st Qu.:-0.21791   1st Qu.:-0.16630   1st Qu.:-0.329444   1st Qu.:-0.23620
 Median :-0.09025   Median :-0.08976   Median :-0.099196   Median :-0.16315
 Mean   :-0.11169   Mean   :-0.07512   Mean   :-0.209761   Mean   :-0.15522
 3rd Qu.: 0.01169   3rd Qu.: 0.01273   3rd Qu.:-0.066938   3rd Qu.:-0.07375
 Max.   : 0.13728   Max.   : 0.14246   Max.   :-0.005092   Max.   :-0.03683
------------------------------------------------------------------------------
LOGNRQovo$Geno: DD
      Larp2/3 1a          Lmhc              Lmyo5a            Lstau
 Min.   :-0.206561   Min.   :-0.400025   Min.   :-0.256630   Min.   :-0.33172
 1st Qu.:-0.087084   1st Qu.:-0.114728   1st Qu.:-0.136653   1st Qu.:-0.17034
 Median : 0.010405   Median : 0.051047   Median :-0.088532   Median : 0.04775
 Mean   : 0.002281   Mean   : 0.005812   Mean   :-0.070966   Mean   : 0.05983
 3rd Qu.: 0.118965   3rd Qu.: 0.188980   3rd Qu.:-0.007127   3rd Qu.: 0.21141
 Max.   : 0.247851   Max.   : 0.246892   Max.   : 0.123435   Max.   : 0.63063
   Lcol11a 2/1        Lmyo18a            Lunc93a           Ldia2 3'          Ldia1 3'
 Min.   :-0.02996   Min.   :-0.666647   Min.   :-0.5906   Min.   :-0.30242   Min.   :-0.10354
 1st Qu.: 0.01195   1st Qu.:-0.503825   1st Qu.:-0.3682   1st Qu.: 0.01899   1st Qu.:-0.01910
 Median : 0.18167   Median :-0.386568   Median :-0.1969   Median : 0.11562   Median : 0.07498
 Mean   : 0.22960   Mean   :-0.345847   Mean   :-0.2001   Mean   : 0.10992   Mean   : 0.07400
 3rd Qu.: 0.45715   3rd Qu.:-0.174298   3rd Qu.:-0.1138   3rd Qu.: 0.16346   3rd Qu.: 0.17035
 Max.   : 0.62965   Max.   : 0.005941   Max.   : 0.2984   Max.   : 0.38321   Max.   : 0.24759
      Lfat1            Ldia2 ORF            Lfry            Larp2/3 3           Lmhc nm
 Min.   :-0.45524   Min.   :-0.35733   Min.   :-0.29261   Min.   :-0.30764   Min.   :-0.28073
 1st Qu.:-0.31218   1st Qu.:-0.05926   1st Qu.:-0.10569   1st Qu.:-0.24986   1st Qu.:-0.17335
 Median :-0.15514   Median :-0.01832   Median :-0.03847   Median :-0.02437   Median :-0.07940
 Mean   :-0.17107   Mean   :-0.01963   Mean   :-0.03417   Mean   :-0.08937   Mean   :-0.11644
 3rd Qu.:-0.02428   3rd Qu.: 0.06520   3rd Qu.: 0.03385   3rd Qu.: 0.01449   3rd Qu.:-0.04382
 Max.   : 0.12736   Max.   : 0.21574   Max.   : 0.16830   Max.   : 0.12145   Max.   : 0.01234
```

S8.2 Tissue Analysis

S8.2.1 Three tissues, DD

```
> by(DDLOGNRQ3, DDLOGNRQ3$Tissue, summary)
DDLOGNRQ3$Tissue: 1cell
 Geno         Tissue     Ldia2 3'          Ldia2 ORF         Ldia2 3'            Lfry
 DD:6   1cell     :6   Min.   :0.3593   Min.   :0.2211   Min.   :-1.0416   Min.   :-0.5460
        Foot      :0   1st Qu.:0.4392   1st Qu.:0.2488   1st Qu.:-1.0337   1st Qu.:-0.4929
        Ovotestis:0   Median :0.4479   Median :0.2791   Median :-1.0140   Median :-0.4748
                      Mean   :0.4525   Mean   :0.2952   Mean   :-0.9859   Mean   :-0.4734
                      3rd Qu.:0.4697   3rd Qu.:0.3528   3rd Qu.:-0.9840   3rd Qu.:-0.4415
                      Max.   :0.5469   Max.   :0.3758   Max.   :-0.8316   Max.   :-0.4154
      Lfat1            Larp2/3 1a       Larp2/3 3
 Min.   :-0.9597   Min.   :-1.211   Min.   :-0.7125
 1st Qu.:-0.8573   1st Qu.:-1.187   1st Qu.:-0.6725
 Median :-0.8444   Median :-1.176   Median :-0.6438
 Mean   :-0.8550   Mean   :-1.162   Mean   :-0.6505
 3rd Qu.:-0.8325   3rd Qu.:-1.127   3rd Qu.:-0.6207
 Max.   :-0.7912   Max.   :-1.106   Max.   :-0.6078
------------------------------------------------------------------------
DDLOGNRQ3$Tissue: Foot
 Geno         Tissue        Ldia2 3'          Ldia2 ORF          Ldia2 3'            Lfry
 DD:5   1cell     :0   Min.   :-0.256886   Min.   :-0.37865   Min.   :-0.14318   Min.   :-0.24169
        Foot      :5   1st Qu.:-0.170208   1st Qu.:-0.31724   1st Qu.:-0.12520   1st Qu.:-0.23560
        Ovotestis:0   Median :-0.064006   Median :-0.30181   Median :-0.06331   Median :-0.16707
                      Mean   : 0.005266   Mean   :-0.14981   Mean   :-0.04112   Mean   :-0.05541
                      3rd Qu.: 0.079013   3rd Qu.:-0.06214   3rd Qu.: 0.06151   3rd Qu.: 0.09828
                      Max.   : 0.438417   Max.   : 0.31079   Max.   : 0.06456   Max.   : 0.26905
      Lfat1            Larp2/3 1a        Larp2/3 3
 Min.   :-0.16105   Min.   :-0.38230   Min.   :-0.2634
 1st Qu.:-0.15549   1st Qu.:-0.26185   1st Qu.:-0.1767
 Median :-0.05871   Median :-0.25607   Median :-0.1752
 Mean   : 0.02624   Mean   :-0.18598   Mean   :-0.1831
 3rd Qu.: 0.17771   3rd Qu.:-0.07258   3rd Qu.:-0.1661
 Max.   : 0.32875   Max.   : 0.04293   Max.   :-0.1339
------------------------------------------------------------------------
DDLOGNRQ3$Tissue: Ovotestis
 Geno          Tissue          Ldia2 3'          Ldia2 ORF          Ldia2 3'            Lfry
 DD:14   1cell     : 0   Min.   :-0.12941   Min.   :-0.05493   Min.   :-0.082785   Min.   :-0.08779
         Foot      : 0   1st Qu.: 0.03671   1st Qu.: 0.05516   1st Qu.:-0.046868   1st Qu.: 0.03651
         Ovotestis:14   Median : 0.09176   Median : 0.22323   Median : 0.002798   Median : 0.09398
                        Mean   : 0.15167   Mean   : 0.19107   Mean   : 0.020820   Mean   : 0.09309
                        3rd Qu.: 0.27406   3rd Qu.: 0.31516   3rd Qu.: 0.065596   3rd Qu.: 0.11011
                        Max.   : 0.44953   Max.   : 0.38183   Max.   : 0.191467   Max.   : 0.29169
      Lfat1             Larp2/3 1a         Larp2/3 3
 Min.   :-0.118579   Min.   :-0.02433   Min.   :-0.12863
 1st Qu.: 0.009445   1st Qu.: 0.11747   1st Qu.:-0.05264
 Median : 0.120728   Median : 0.27707   Median : 0.06423
 Mean   : 0.126823   Mean   : 0.28661   Mean   : 0.02863
 3rd Qu.: 0.231337   3rd Qu.: 0.43263   3rd Qu.: 0.10102
 Max.   : 0.422312   Max.   : 0.66345   Max.   : 0.13062
```

S8.2.2 Three tissues, dd

```
> by(SSLOGNRQ3, SSLOGNRQ3$Tissue, summary)
SSLOGNRQ3$Tissue: 1cell
 Geno         Tissue     Ldia2 3'          Ldia2 ORF          Ldia1 3'            Lfry
 dd:6   1cell     :6   Min.   :-2.301   Min.   :-1.439   Min.   :-1.0648   Min.   :-0.5311
        Foot      :0   1st Qu.:-1.915   1st Qu.:-1.390   1st Qu.:-0.9458   1st Qu.:-0.4970
        Ovotestis:0   Median :-1.563   Median :-1.214   Median :-0.9092   Median :-0.4718
                      Mean   :-1.728   Mean   :-1.234   Mean   :-0.9299   Mean   :-0.4810
                      3rd Qu.:-1.527   3rd Qu.:-1.119   3rd Qu.:-0.8915   3rd Qu.:-0.4667
                      Max.   :-1.396   Max.   :-1.004   Max.   :-0.8525   Max.   :-0.4408
      Lfat1            Larp2/3 1a        Larp2/3 3
 Min.   :-0.8591   Min.   :-1.225   Min.   :-0.7159
 1st Qu.:-0.8250   1st Qu.:-1.194   1st Qu.:-0.6955
 Median :-0.8068   Median :-1.180   Median :-0.6841
 Mean   :-0.8024   Mean   :-1.185   Mean   :-0.6589
 3rd Qu.:-0.7703   3rd Qu.:-1.168   3rd Qu.:-0.6099
 Max.   :-0.7529   Max.   :-1.161   Max.   :-0.5847
------------------------------------------------------------------------
SSLOGNRQ3$Tissue: Foot
 Geno         Tissue        Ldia2 3'          Ldia2 ORF          Ldia1 3'            Lfry
 dd:5   1cell     :0   Min.   :-0.05519   Min.   :-0.18152   Min.   :-0.11720   Min.   :-0.18511
        Foot      :5   1st Qu.:-0.03113   1st Qu.:-0.12795   1st Qu.:-0.10740   1st Qu.:-0.07059
        Ovotestis:0   Median : 0.03549   Median : 0.03187   Median :-0.07175   Median :-0.05638
                      Mean   : 0.06193   Mean   :-0.00130   Mean   :-0.05818   Mean   :-0.04314
                      3rd Qu.: 0.17771   3rd Qu.: 0.13218   3rd Qu.:-0.03008   3rd Qu.: 0.01087
                      Max.   : 0.18278   Max.   : 0.13892   Max.   : 0.03552   Max.   : 0.08550
      Lfat1            Larp2/3 1a        Larp2/3 3
 Min.   :-0.05090   Min.   :-0.21832   Min.   :-0.1864
 1st Qu.: 0.06581   1st Qu.:-0.18334   1st Qu.:-0.1733
 Median : 0.14224   Median :-0.04708   Median :-0.1680
 Mean   : 0.10624   Mean   :-0.10576   Mean   :-0.1660
 3rd Qu.: 0.17599   3rd Qu.:-0.04291   3rd Qu.:-0.1586
 Max.   : 0.19808   Max.   :-0.03715   Max.   :-0.1437
------------------------------------------------------------------------
SSLOGNRQ3$Tissue: Ovotestis
 Geno          Tissue          Ldia2 3'          Ldia2 ORF          Ldia1 3'            Lfry
 dd:14   1cell     : 0   Min.   :-0.29794   Min.   :-0.190263   Min.   :-0.26370   Min.   :-0.29597
         Foot      : 0   1st Qu.:-0.20419   1st Qu.:-0.066498   1st Qu.:-0.05913   1st Qu.:-0.01411
         Ovotestis:14   Median :-0.04102   Median :-0.002847   Median : 0.04095   Median : 0.03111
                        Mean   :-0.06819   Mean   : 0.016705   Mean   : 0.01691   Mean   : 0.02512
                        3rd Qu.: 0.03545   3rd Qu.: 0.084821   3rd Qu.: 0.12528   3rd Qu.: 0.06440
                        Max.   : 0.16661   Max.   : 0.291789   Max.   : 0.19665   Max.   : 0.25410
      Lfat1             Larp2/3 1a         Larp2/3 3
 Min.   :-0.32064   Min.   :-0.20491   Min.   :-0.35075
 1st Qu.:-0.01700   1st Qu.: 0.06902   1st Qu.:-0.12333
 Median : 0.03099   Median : 0.20705   Median :-0.02411
 Mean   : 0.04322   Mean   : 0.26004   Mean   :-0.03928
 3rd Qu.: 0.13244   3rd Qu.: 0.46988   3rd Qu.: 0.02401
 Max.   : 0.33473   Max.   : 0.75647   Max.   : 0.17853
```

S8.2.3 Two tissues, DD

```
> by(DDLOGNRQ2, DDLOGNRQ2$Tissue, summary)
DDLOGNRQ2$Tissue: Foot
 Geno          Tissue       Lmhc             Lmhc nm           Lcol11a 2/1         Lmyo5a
 DD:5   Foot     :5   Min.   :0.5168   Min.   :-0.31930   Min.   :0.1667   Min.   :0.008179
        Ovotestis:0   1st Qu.:0.5364   1st Qu.:-0.13421   1st Qu.:0.3792   1st Qu.:0.111181
                      Median :0.5578   Median :-0.12290   Median :0.4841   Median :0.130526
                      Mean   :0.7038   Mean   :-0.01964   Mean   :0.5660   Mean   :0.331825
                      3rd Qu.:0.9202   3rd Qu.: 0.22874   3rd Qu.:0.8909   3rd Qu.:0.606823
                      Max.   :0.9880   Max.   : 0.24946   Max.   :0.9091   Max.   :0.802418
     Lmyo18a            Lstau             Lunc93a
 Min.   :-0.37328   Min.   :-0.29615   Min.   :-1.4512
 1st Qu.:-0.17443   1st Qu.:-0.22135   1st Qu.:-1.0861
 Median : 0.05412   Median :-0.21625   Median :-1.0639
 Mean   : 0.01897   Mean   :-0.03143   Mean   :-0.8874
 3rd Qu.: 0.17936   3rd Qu.: 0.19144   3rd Qu.:-0.6810
 Max.   : 0.40908   Max.   : 0.38516   Max.   :-0.1549
------------------------------------------------------------------------
DDLOGNRQ2$Tissue: Ovotestis
 Geno          Tissue       Lmhc             Lmhc nm           Lcol11a 2/1         Lmyo5a
 DD:14   Foot     : 0   Min.   :-0.517681   Min.   :-0.17128   Min.   :-0.310938   Min.   :0.03377
         Ovotestis:14   1st Qu.:-0.178545   1st Qu.:-0.06786   1st Qu.:-0.144665   1st Qu.:0.13448
                        Median :-0.003189   Median :-0.03740   Median :-0.052170   Median :0.18873
                        Mean   :-0.033270   Mean   :-0.03030   Mean   : 0.004545   Mean   :0.20724
                        3rd Qu.: 0.111410   3rd Qu.: 0.03141   3rd Qu.: 0.160915   3rd Qu.:0.27167
                        Max.   : 0.317614   Max.   : 0.11078   Max.   : 0.423959   Max.   :0.41394
     Lmyo18a            Lstau             Lunc93a
 Min.   :-0.142457   Min.   :-0.47879   Min.   :-0.348685
 1st Qu.:-0.039341   1st Qu.:-0.29538   1st Qu.:-0.097392
 Median :-0.006496   Median :-0.02735   Median :-0.004646
 Mean   : 0.084436   Mean   :-0.03098   Mean   : 0.025008
 3rd Qu.: 0.271703   3rd Qu.: 0.19182   3rd Qu.: 0.152491
 Max.   : 0.400680   Max.   : 0.56579   Max.   : 0.482297
```

S8.2.4 Two tissues, Dd

```
> by(DSLOGNRQ3, DSLOGNRQ3$Tissue, summary)
DSLOGNRQ3$Tissue: 1cell
 Geno          Tissue       Ldia2 3'          Ldia2 ORF          Ldia1 3'          Lfry
 Dd:5   1cell    :5   Min.   :0.1292   Min.   :-0.063676   Min.   :-1.0596   Min.   :-0.5270
        Ovotestis:0   1st Qu.:0.2196   1st Qu.:-0.056030   1st Qu.:-0.9940   1st Qu.:-0.4544
                      Median :0.2350   Median :-0.047193   Median :-0.9848   Median :-0.4490
                      Mean   :0.2242   Mean   :-0.005429   Mean   :-0.9782   Mean   :-0.4590
                      3rd Qu.:0.2554   3rd Qu.: 0.013846   3rd Qu.:-0.9421   3rd Qu.:-0.4413
                      Max.   :0.2819   Max.   : 0.125908   Max.   :-0.9107   Max.   :-0.4235
     Lfat1            Larp2/3 1a        Larp2/3 3
 Min.   :-0.8691   Min.   :-1.173   Min.   :-0.8284
 1st Qu.:-0.8186   1st Qu.:-1.161   1st Qu.:-0.8089
 Median :-0.8168   Median :-1.128   Median :-0.7785
 Mean   :-0.8118   Mean   :-1.115   Mean   :-0.7580
 3rd Qu.:-0.7990   3rd Qu.:-1.065   3rd Qu.:-0.7045
 Max.   :-0.7558   Max.   :-1.050   Max.   :-0.6696
------------------------------------------------------------------------
DSLOGNRQ3$Tissue: Ovotestis
 Geno          Tissue       Ldia2 3'          Ldia2 ORF          Ldia1 3'          Lfry
 Dd:8   1cell    :0   Min.   :-0.13756   Min.   :-0.14318   Min.   :-0.41418   Min.   :-0.11589
        Ovotestis:8   1st Qu.: 0.01659   1st Qu.: 0.04068   1st Qu.:-0.14871   1st Qu.:-0.05748
                      Median : 0.10326   Median : 0.14344   Median : 0.06840   Median : 0.10680
                      Mean   : 0.08563   Mean   : 0.12971   Mean   :-0.02232   Mean   : 0.08282
                      3rd Qu.: 0.21945   3rd Qu.: 0.21040   3rd Qu.: 0.14270   3rd Qu.: 0.18888
                      Max.   : 0.22777   Max.   : 0.39670   Max.   : 0.22302   Max.   : 0.26220
     Lfat1            Larp2/3 1a        Larp2/3 3
 Min.   :-0.17363   Min.   :0.02028   Min.   :-0.45941
 1st Qu.:-0.02019   1st Qu.:0.23030   1st Qu.:-0.19536
 Median : 0.05595   Median :0.35474   Median : 0.04570
 Mean   : 0.07782   Mean   :0.35335   Mean   :-0.06106
 3rd Qu.: 0.21133   3rd Qu.:0.46117   3rd Qu.: 0.08406
 Max.   : 0.29686   Max.   :0.63870   Max.   : 0.16162  |
```

S8.2.5 Two tissues, dd

```
> by(SSLOGNRQ2, SSLOGNRQ2$Tissue, summary)
SSLOGNRQ2$Tissue: Foot
 Geno          Tissue       Lmhc             Lmhc nm           Lcol11a 2/1         Lmyo5a
 dd:5   Foot     :5   Min.   :0.5668   Min.   :-0.30272   Min.   :0.6680   Min.   :0.1254
        Ovotestis:0   1st Qu.:0.8287   1st Qu.:-0.10361   1st Qu.:0.7147   1st Qu.:0.3381
                      Median :0.8930   Median :-0.04025   Median :0.8580   Median :0.4541
                      Mean   :0.8495   Mean   :-0.06824   Mean   :0.8331   Mean   :0.4168
                      3rd Qu.:0.9780   3rd Qu.: 0.02157   3rd Qu.:0.9519   3rd Qu.:0.5593
                      Max.   :0.9811   Max.   : 0.08381   Max.   :0.9726   Max.   :0.6071
     Lmyo18a            Lstau             Lunc93a
 Min.   :-0.002971   Min.   :-0.18574   Min.   :-1.0844
 1st Qu.: 0.068737   1st Qu.:-0.08664   1st Qu.:-0.7802
 Median : 0.136355   Median : 0.13474   Median :-0.6172
 Mean   : 0.178349   Mean   : 0.03206   Mean   :-0.7211
 3rd Qu.: 0.323867   3rd Qu.: 0.14892   3rd Qu.:-0.5733
 Max.   : 0.365757   Max.   : 0.14900   Max.   :-0.5504
------------------------------------------------------------------------
SSLOGNRQ2$Tissue: Ovotestis
 Geno          Tissue       Lmhc             Lmhc nm           Lcol11a 2/1         Lmyo5a
 dd:14   Foot     : 0   Min.   :-0.72876   Min.   :-0.16930   Min.   :-0.45291   Min.   :0.00921
         Ovotestis:14   1st Qu.:-0.39252   1st Qu.:-0.10966   1st Qu.:-0.19305   1st Qu.:0.06047
                        Median :-0.26189   Median :-0.01537   Median : 0.04202   Median :0.21782
                        Mean   :-0.20868   Mean   :-0.01116   Mean   : 0.00199   Mean   :0.21060
                        3rd Qu.:-0.06861   3rd Qu.: 0.09488   3rd Qu.: 0.11031   3rd Qu.:0.32824
                        Max.   : 0.57052   Max.   : 0.17271   Max.   : 0.77870   Max.   :0.41710
     Lmyo18a            Lstau             Lunc93a
 Min.   :-0.20621   Min.   :-0.45209   Min.   :-0.86682
 1st Qu.:-0.06782   1st Qu.:-0.05977   1st Qu.:-0.22146
 Median : 0.06264   Median : 0.05972   Median :-0.04816
 Mean   : 0.06956   Mean   : 0.07384   Mean   :-0.11337
 3rd Qu.: 0.20974   3rd Qu.: 0.23134   3rd Qu.: 0.01745
 Max.   : 0.33839   Max.   : 0.49646   Max.   : 0.33100  |
```

S9. QPCR average amplification efficiency

After performing the Cq data analysis, it became apparent that a geometric mean may have been more appropriate for estimating average primer efficiencies because of the non-linear distribution of a percentage value. However there was little difference between the arithmetic mean and the geometric mean and it is assumed this variation would have a negligible effect on the comparative Cq calculations (Table S5).

Table S5 Comparison of calculated arithmetic mean and geometric mean calculated for average percentage amplification efficiency

Endogenous Controls		
Primer Name	**Arithmetic mean**	**Geometric mean**
ACA_11210_F1R1	91.215	91.202
EF1_8940_F1R1	115.544	115.239
HiS_8200_F1R1	94.319	94.317
RPL_2341_F2R2	90.649	90.554
UB_3288_F2R2	92.325	92.322
YWHAZ_562_F1R1	91.825	91.798
Experimental GOIs		
Primer Name	**Arithmetic mean**	**Geometric mean**
ARPI_1-2b	84.705	84.488
ARPII_1-3a	77.477	79.019
COL2A_3-4a	88.992	88.763
MHCI_1-2a	89.181	89.076
MHCII_2-3a	92.437	92.122
MV_F2R2	94.609	94.608
Staufen_3-4a	95.706	95.703
UMVIII_F2R2	91.295	91.295
UNC-93_FR	97.845	97.714
FOR_3'_UTR	91.234	91.179
FOR_ORF	94.846	94.195
PARA_3'_UTR	98.577	98.565
CAD_F1R1	83.789	83.024
FURRY_F1R1	87.614	87.601

S10. QPCR genotype analysis, Ovotestis histograms

Histograms are presented on the following pages for the LOG NRQ values for each of the GOIs within the genotype analysis of the ovotestis (Figure S15 - Figure S28).

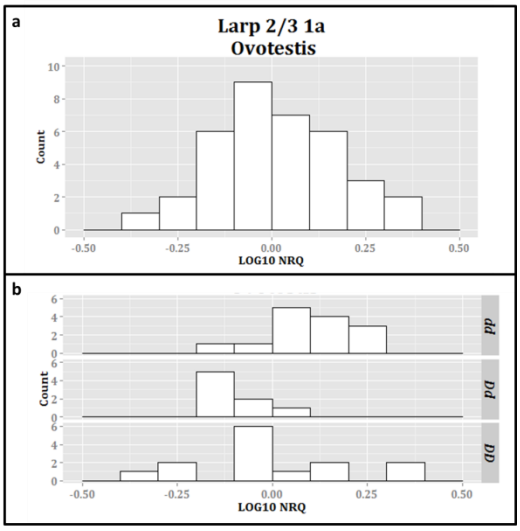


Figure S15 LOGNRQ values of *Larp2/3 1a* in ovotestis (a) and genotype specific ovotestis (b).

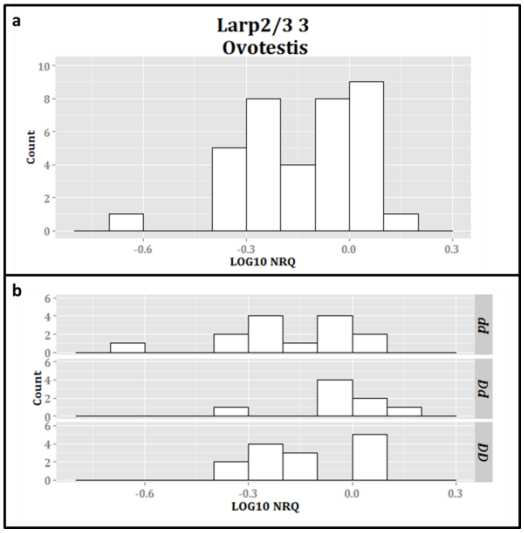


Figure S16 LOGNRQ values of *Larp2/3 3* in ovotestis (a) and genotype specific ovotestis (b).

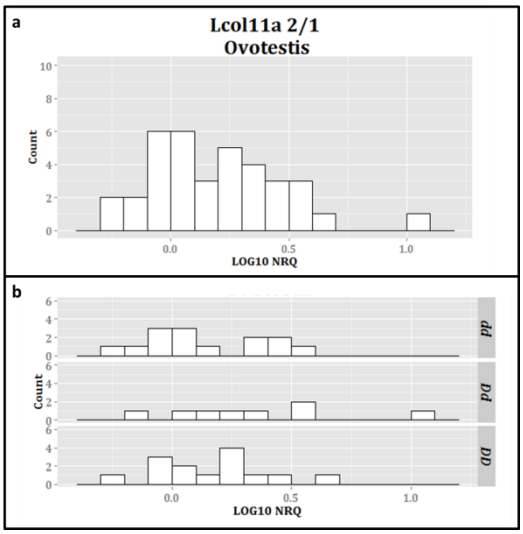


Figure S17 LOGNRQ values of *Lcol11a 2/1* in ovotestis (a) and genotype specific ovotestis (b).

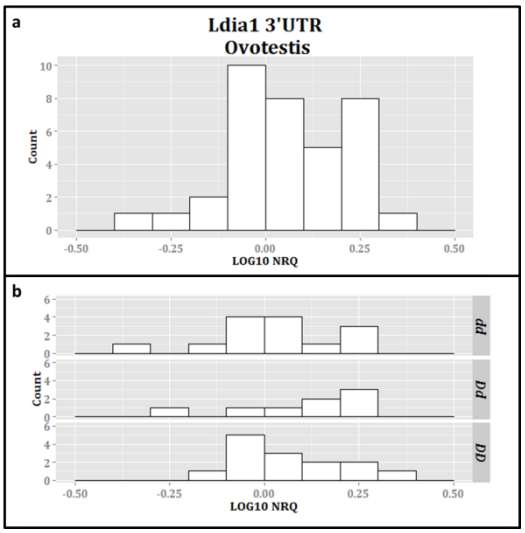


Figure S18 LOGNRQ values of *Ldia1* 3' UTR in ovotestis (a) and genotype specific ovotestis (b).

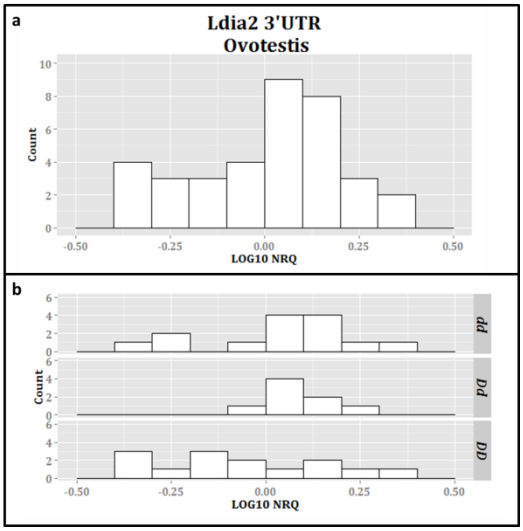


Figure S19 LOGNRQ values of *Ldia2* 3'UTR in ovotestis (a) and genotype specific ovotestis (b).

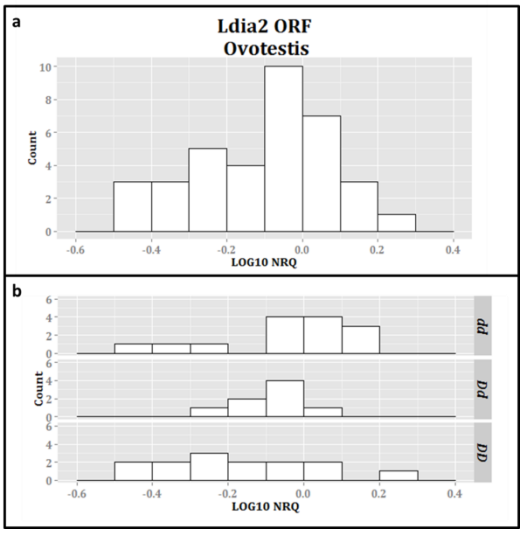


Figure S20 LOGNRQ values of *Ldia2* ORF in ovotestis (a) and genotype specific ovotestis (b).

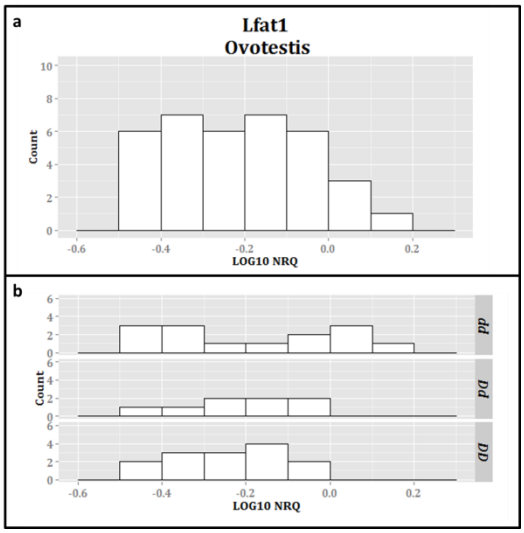


Figure S21 LOGNRQ values of *Lfat1* in ovotestis (a) and genotype specific ovotestis (b).

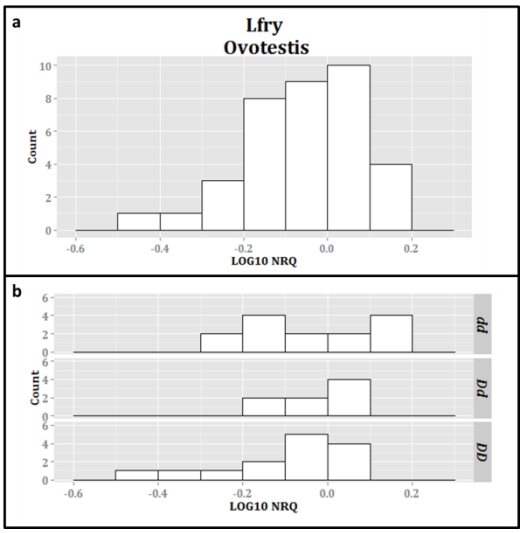


Figure S22 LOGNRQ values of *Lfry* in ovotestis (a) and genotype specific ovotestis (b).

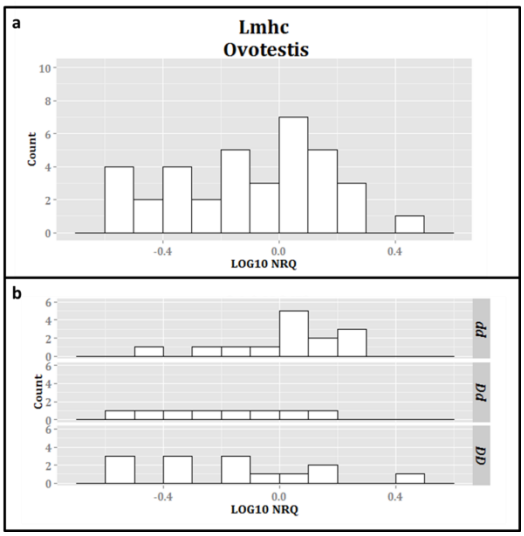


Figure S23 LOGNRQ values of *Lmhc* in ovotestis (a) and genotype specific ovotestis (b).

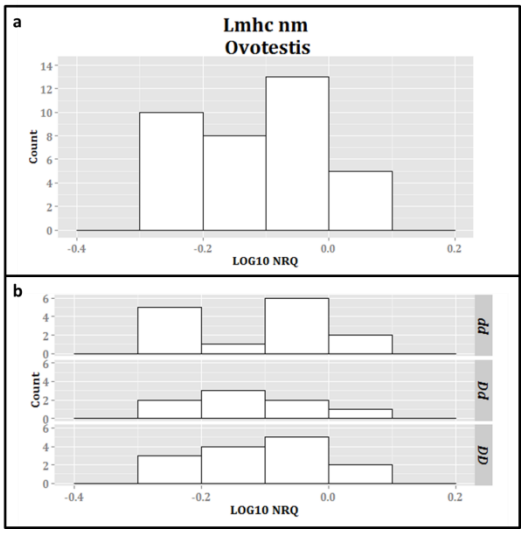


Figure S24 LOGNRQ values of *Lmhc nm* in ovotestis (a) and genotype specific ovotestis (b).

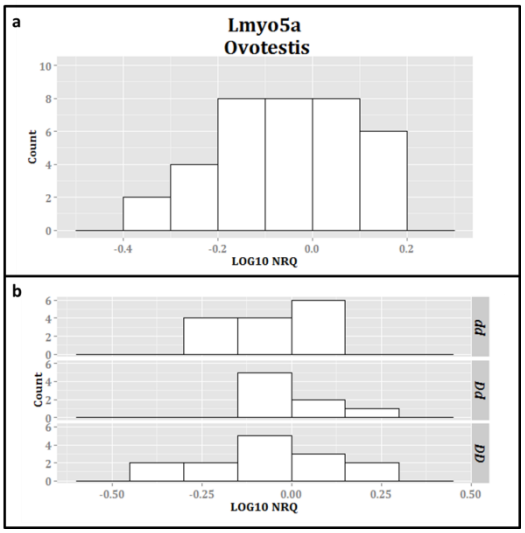


Figure S25 LOGNRQ values of *Lmyo5a* in ovotestis (a) and genotype specific ovotestis (b).

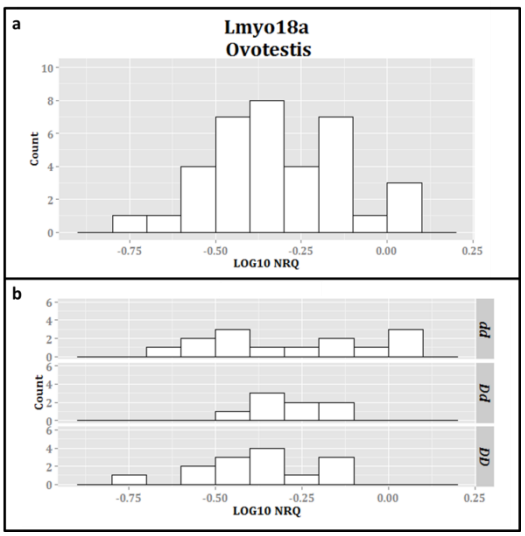


Figure S26 LOGNRQ values of *Lmyo18a* in ovotestis (a) and genotype specific ovotestis (b).

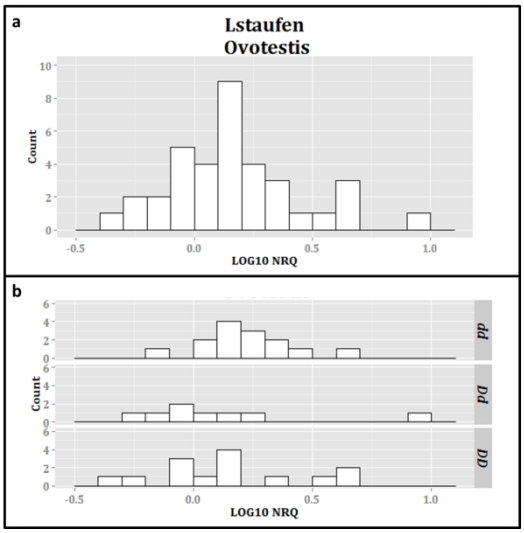


Figure S27 LOGNRQ values of *Lstau* in ovotestis (a) and genotype specific ovotestis (b).

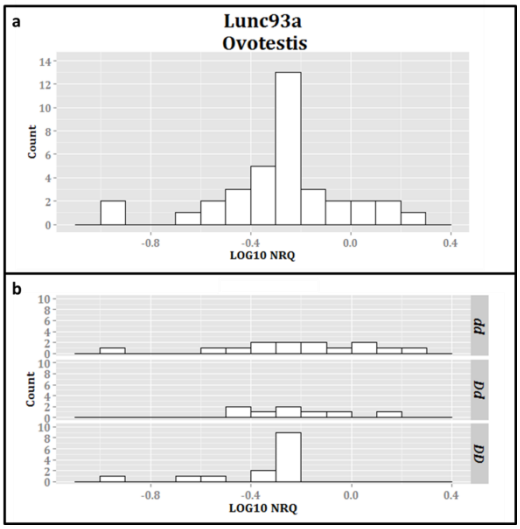


Figure S28 LOGNRQ values of *Lunc93a* in ovotestis (a) and genotype specific ovotestis (b).

S11. eRAD Library preparation

S11.1 mRNA enrichment

Total RNA samples retained a varying proportion of RNA once selected for mRNA ranging from 4.55% - 17.90% (Table S6). This did not appear to be related to sample starting concentration of quality.

Table S6 Nanodrop quantification of TRI Reagent extracted total RNA and sunsequent Poly(A) purified mRNA, Showing Sample descriptive information (Sample Info) including library preparation (Lib.), sample ID (ID) and genotype (Geno). Also provided is a calculated percentage sample retained (% ret.) of the total RNA following mRNA purification.

Sample Info			Total RNA (100µl)				mRNA (10µl)				Ratio
Lib.	ID	Geno	ng/µl	260/280	260/230	total yield µg	ng/µl	260/280	260/230	Total yield µg	% ret
3	8515	*DD*	89.36	1.76	1.00	8.94	94.04	1.71	1.89	0.94	10.52
3	9014	*DD*	169.69	1.81	1.84	16.97	77.00	1.66	1.84	0.77	4.54
3	8544	*Dd*	115.45	1.86	1.30	11.55	150.08	1.68	1.96	1.50	13.00
3	8559	*Dd*	136.98	1.81	1.12	13.70	103.57	1.70	1.92	1.04	7.56
3	8562	*Dd*	140.96	1.81	1.25	14.10	87.88	1.71	1.85	0.88	6.23
3	8869	*Dd*	294.76	1.89	1.93	29.48	142.16	1.84	2.23	1.42	4.82
3	9013	*Dd*	149.34	1.78	1.79	14.93	111.74	1.69	2.03	1.12	7.48
3	8808	*dd*	110.26	1.77	1.48	11.03	116.94	1.69	1.98	1.17	10.61
3	8862	*dd*	211.62	1.90	1.48	21.16	96.23	1.69	1.98	0.96	4.55
3	9007	*dd*	107.33	1.77	1.12	10.73	83.43	1.66	1.93	0.83	7.77
3	9009	*dd*	104.97	1.79	0.61	10.50	91.31	1.69	1.62	0.91	8.70
4	8502	*DD*	55.87	1.59	1.15	5.59	100.02	1.60	1.51	1.00	17.90
4	8582	*DD*	61.42	1.67	1.40	6.14	103.35	1.59	1.43	1.03	16.83
4	8500	*Dd*	96.10	1.70	1.39	9.61	118.02	1.67	1.73	1.18	12.28
4	8522	*Dd*	98.62	1.72	1.52	9.86	125.53	1.63	1.47	1.26	12.73
4	8530	*Dd*	82.78	1.73	1.40	8.28	115.69	1.63	1.42	1.16	13.98
4	8560	*Dd*	78.62	1.82	1.17	7.86	109.62	1.59	1.50	1.10	13.94
4	9001	*Dd*	105.17	1.74	1.41	10.52	115.58	1.64	1.63	1.16	10.99
4	8531	*dd*	83.81	1.67	1.43	8.38	88.99	1.67	1.66	0.89	10.62
4	8587	*dd*	56.78	1.70	0.87	5.68	55.25	1.74	1.91	0.55	9.73
4	8867	*dd*	68.92	1.75	0.51	6.89	86.72	1.59	1.83	0.87	12.58
4	9000	*dd*	78.42	1.72	1.19	7.84	107.10	1.59	1.44	1.07	13.66

S11.2 cDNA yield

The variable cDNA yield was explored by simple correlations with associated mRNA quality measures. The 260/280 showed a potential cause with a positive correlation between 260/280 ratio and resulting total cDNA yield (R^2: 0.6041), however this relationship was reduced following the removal of the one sample of much higher quantity (R^2: 0.3241) (Figure S29).

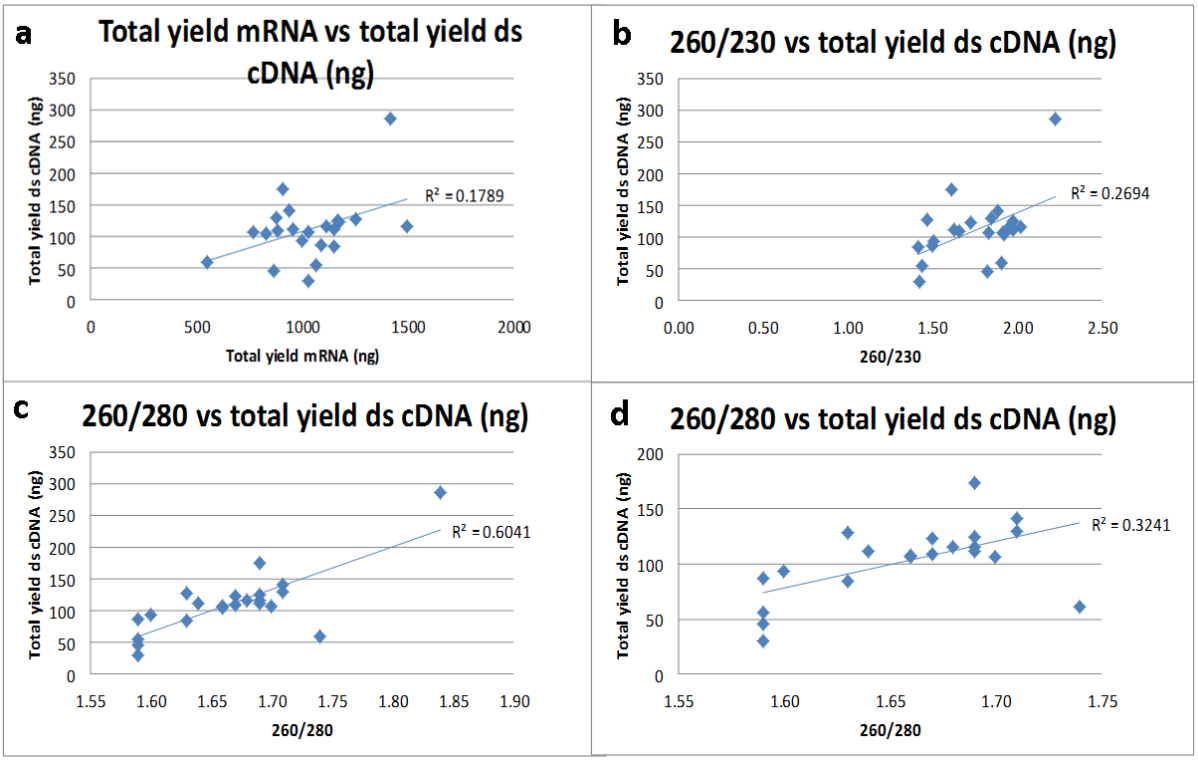


Figure S29 Correlations between mRNA sample quality and resulting cDNA yield, including total yield of mRNA (a), 260/230 ratio (b) and 260/280 ratio including all samples (c) and following the removal of the high quantity individual 8869 (d).

S11.3 Gel extraction

The final step of the eRAD library preparation involves size selection through gel electrophoresis. This step also provides a visual measure of the quality of the sequencing libraries. Library 3/L006 showed a reasonably high level of primer dimer (indicated by the smaller sized distinct band below the main library smear) which is assumed to be removed from the library following extraction (Figure S30, 1b). Library 4/L007 showed a greater problem of primer dimer (Figure S30, 2a). The library was assessed again via gel electrophoresis to ensure the removal of primer dimer following the first size extraction. The lack of primer dimer detectable in the already gel extracted library (Figure S30, 2a), provides support that the size selection through gel electrophoresis is adequate to remove the majority of primer dimer.

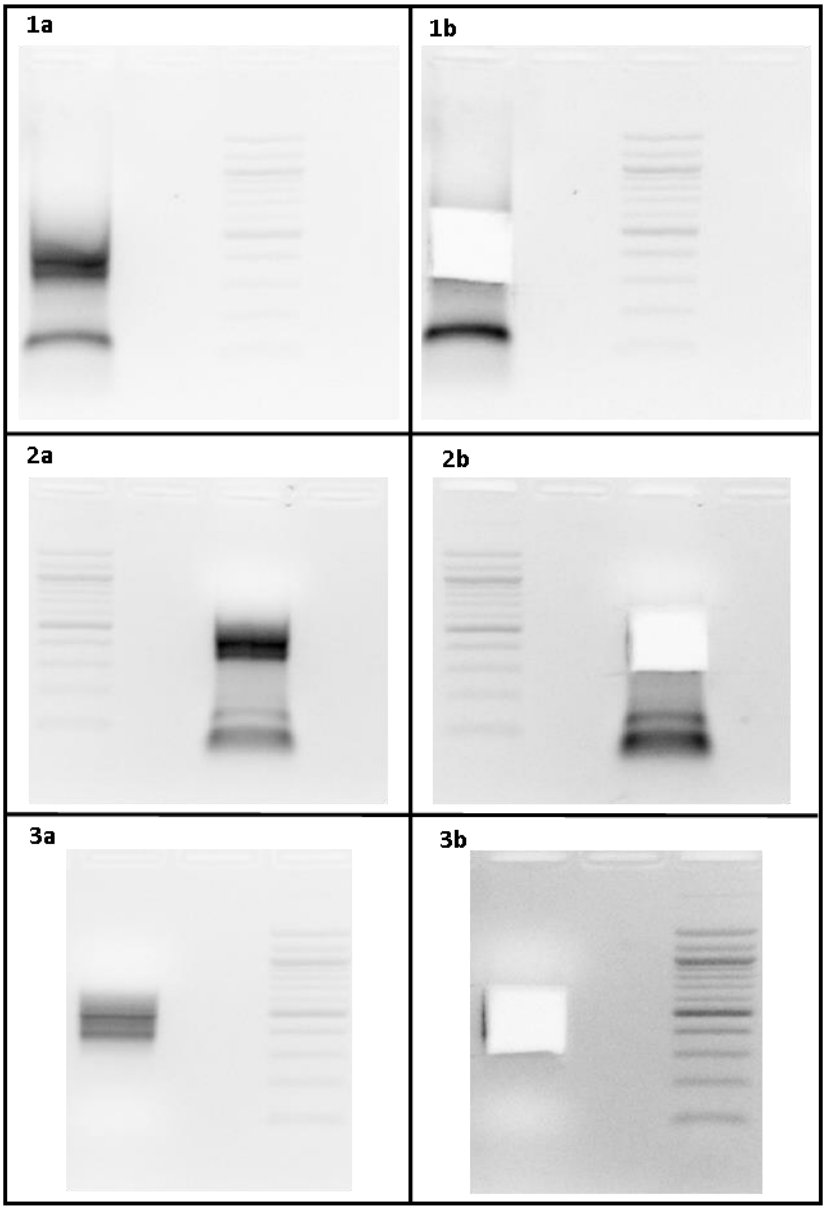


Figure S30 Visualisation of eRAD libraries via agarose gel electrophoresis before (a) and after (b) extraction. Library 3/L006 is shown in 1. The first size selection of Library 4/L007 is shown in 2, and the re-extraction in 3.

S11.4 eRAD Barcode distance calculation

Table S7 shows sequence similarity between each unique barcode within the two eRAD libraries. The highest number of in sequence identical bases was two. Therefore the barcode distance was three.

Table S7 Calculation of sequence distance between the 22 barcodes/MIDs used within the eRAD libraries

Barcodes	ATGCT	CCAAC	AGCTG	CATGA	GCCGG	GAGAT	TGCAA	ACGTA	CGTAT	GTACA	TAATG	TACGT	GTTGT	TGACC	GGTTC	CAGTC	CCTTG	ATTAG	TCTCT	ATCGA	CTTCC	AGAGT
ATGCT	11111																					
CCAAC	xxxxx	11111																				
AGCTG	1xxxx	xxxxx	11111																			
CATGA	xxxxx	1xxxx	xxxxx	11111																		
GCCGG	xxxxx	x1xxx	xx1x1	xxx1x	11111																	
GAGAT	xx1x1	xxx1x	xxxxx	x1xxx	1xxxx	11111																
TGCAA	xxxxx	xxx1x	x11xx	xxxx1	xx1xx	xxx1x	11111															
ACGTA	1x1xx	x1xxx	1xx1x	xxxx1	x1xxx	xx1xx	xxxx1	11111														
CGTAT	xxxx1	1xx1x	x1xxx	1x1xx	xxxxx	xxx11	x1x1x	xxxxx	11111													
GTACA	x1x1x	xx1xx	xxxxx	xxxx1	1xxxx	1xxxx	xxxx1	xxxx1	xxxxx	11111												
TAATG	xxxxx	xx1xx	xxx11	x1xxx	xxxx1	x1xxx	1xxxx	xxx1x	xxxxx	xx1xx	11111											
TACGT	xxxx1	xxxxx	xx1xx	x1x1x	xx11x	x1xx1	1x1xx	xxxxx	xxxx1	xxxxx	11xxx	11111										
GTTGT	x1xx1	xxxxx	xxxxx	xx11x	1xx1x	1xxx1	xxxxx	xxxxx	xx1x1	11xxx	xxxxx	xxx11	11111									
TGACC	xxx11	xx1x1	xxxxx	xxxxx	xxxxx	xxxxx	11xxx	xxxxx	x1xxx	xx11x	1x1xx	1xxxx	xxxxx	11111								
GGTTC	xxxxx	xxxx1	x1x1x	xx1xx	1xxxx	1xxxx	xxxxx	xxx1x	x11xx	1xxxx	xxx1x	xxxxx	1x1xx	xxxx1	11111							
CAGTC	xx1xx	1xxx1	xxx1x	11xxx	xxxxx	x11xx	xxxxx	xx11x	1xxxx	xxxxx	x1x1x	x1xxx	xxxxx	xxxx1	xxx11	11111						
CCTTG	xxxxx	11xxx	xxx11	1x1xx	x1xxx	xxxxx	xxxxx	x1x1x	1x1xx	xxxxx	xxx11	xxxxx	xx1xx	xx1xx	xx11x	1xx1x	11111					
ATTAG	11xxx	xxx1x	1xxx1	xx1xx	xxxx1	xxx1x	xxx1x	1xxxx	xx11x	x1xxx	xxxx1	xxxxx	x11xx	xxxxx	xx1xx	xxxxx	xx1x1	11111				
TCTCT	xxx11	x1xxx	xxxxx	xx1xx	x1xxx	xxxx1	1xxxx	x1xxx	xx1x1	xxx1x	1xxxx	1xxx1	xx1x1	xxx1x	xx1xx	xxxxx	x11xx	xx1xx	11111			
ATCGA	11xxx	xxxxx	1x1xx	xxx11	xx11x	xxxxx	xx1x1	1xxx1	xxxxx	x1xx1	xxxxx	xx11x	x1x1x	xxxxx	xxxxx	xxxxx	xxxxx	11xxx	xxxxx	11111		
CTTCC	x1x1x	1xxx1	xxxxx	1x1xx	xxxxx	xxxxx	xxxxx	xxxxx	1x1xx	x1x1x	x11xx	xxxxx	x11xx	xxx11	xx1x1	1xxx1	1x1xx	x11xx	xx11x	x1xxx	11111	
AGAGT	1xxx1	xx1xx	11xxx	xxx1x	xxx1x	xxxx1	x1xxx	1xxxx	x1xx1	xx1xx	xx1xx	xxx11	xxx11	x11xx	x1xxx	xxxxx	xxxxx	1xxxx	xxxx1	1xx1x	xxxxx	11111

S12. eRAD library sequencing Fast QC output

A summary of the FastQC reports for each of the raw sequence data files in shown in Figure S31. As can be seen many parameters failed the quality standards. However some of these factors are filtered later in the RAD data analysis. For example, duplicated sequences were removed via the PCR clone filter and per base sequence quality is improved through trimmed reads.

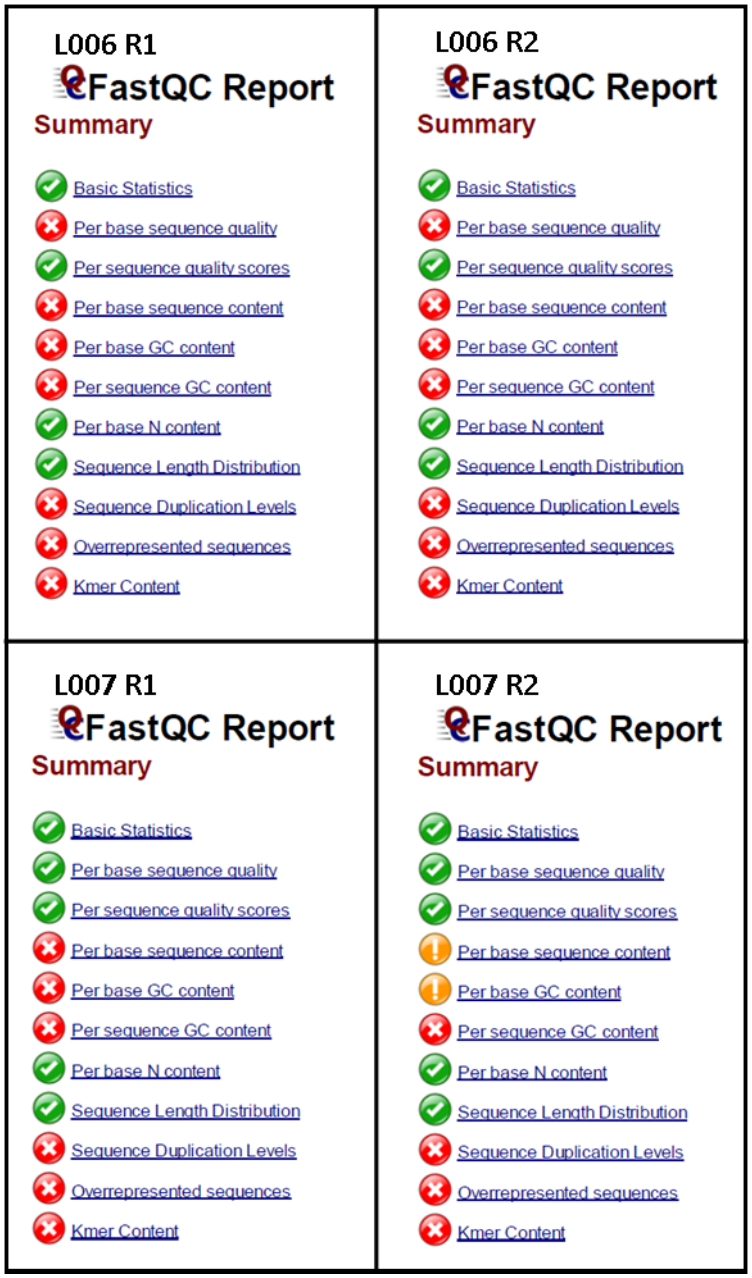


Figure S31 Summary of FastQC report for the eRAD raw sequence data. A green tick represents a pass for quality whereas a red cross indicates data has failed to meet quality standards. An amber exclamation mark advises caution regarding the quality of the data.

S13. Sample representation bias correlations

Potential causes of the sample representation bias were explored by assessing correlation of a number of factors associated with the sample and library preparations by simple scatter plots and R^2 values. No significant correlations were identified (Figure S32, Figure S33).

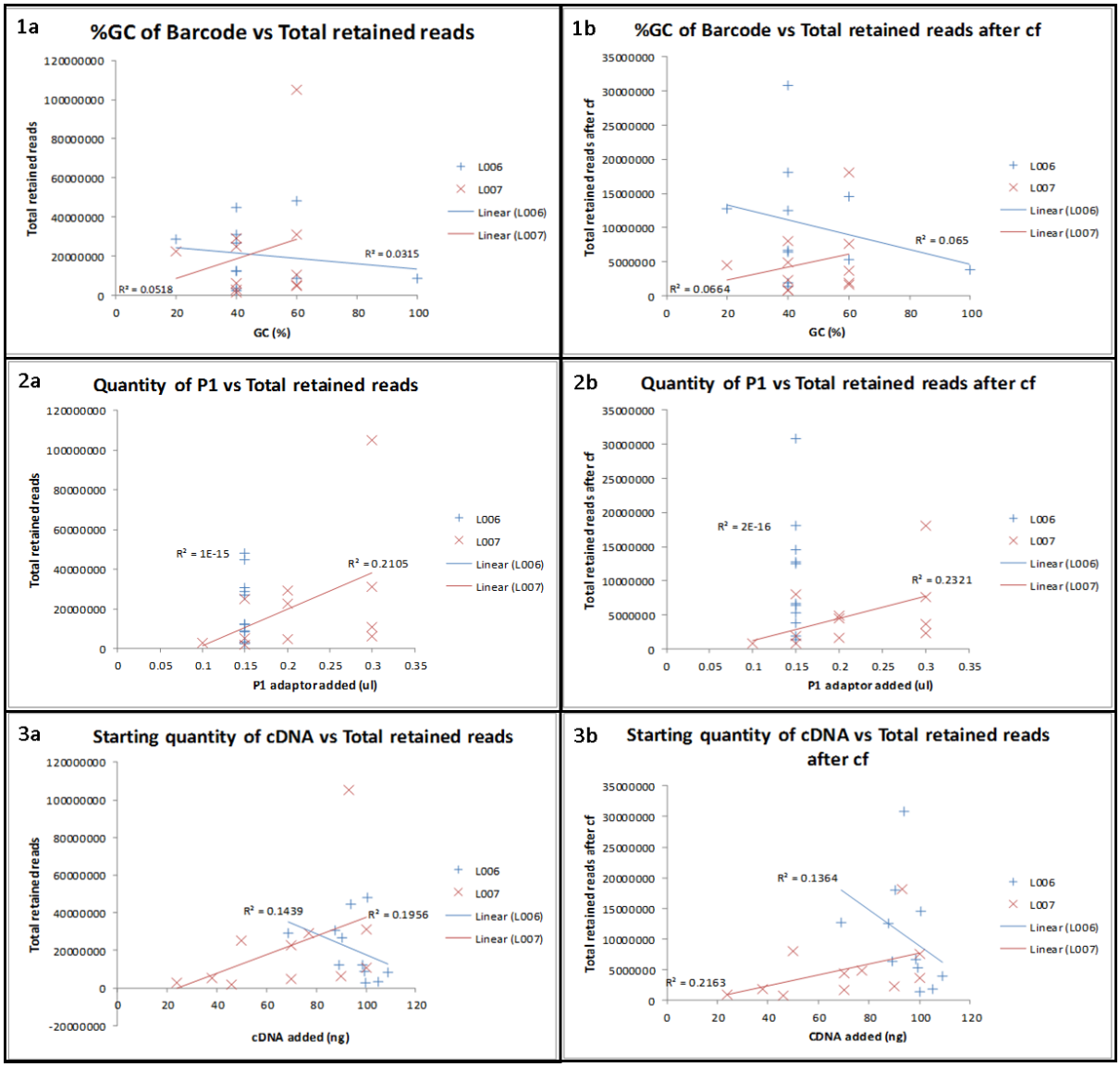


Figure S32 Scatter plots show library specific correlations of library preparation factors and the resulting total retained reads before (a) and after (b) the clone filter. Factors shown include the percentage representation GC bases within the unique sample identifier/barcode (1), the volume (µl) of P1 adaptor added to the sample preparation (2) and the starting quantity (ng) of ds cDNA added to the initial sample digest (3).

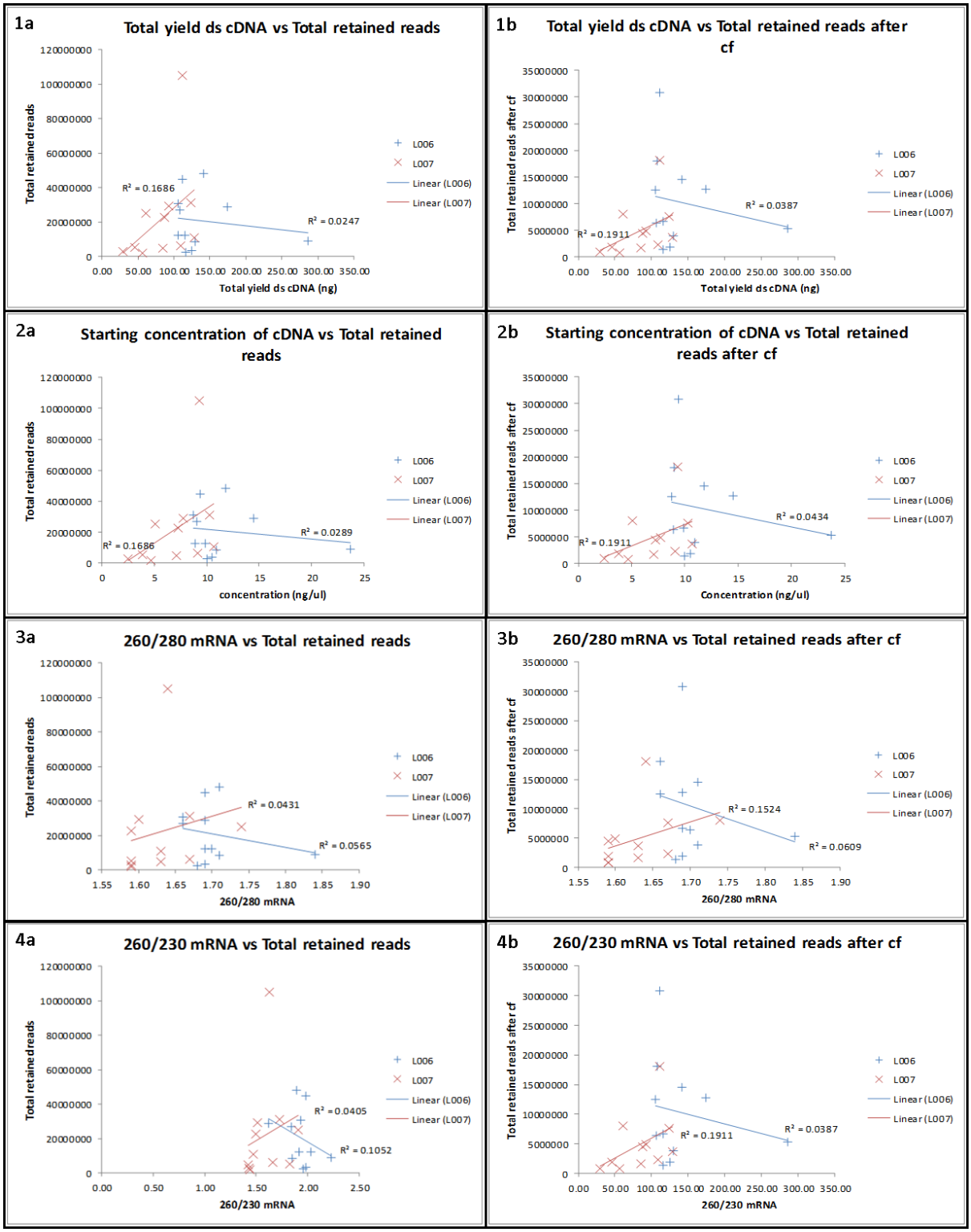


Figure S33 Scatter plots show library specific correlations of cDNA quality factors and the resulting total retained reads before (a) and after (b) the clone filter. Factors shown include the total yield (ng) of the ds cDNA sample (1), the concentration (ng) of the ds cDNA sample (2) and the quality of mRNA prior to cDNA synthesis, inferred from 260/280 ratios (3) and 260/230 ratios (4).

S14. Stacks, denovo parameters trial

Full table description: The minimum sequencing depth to create a stack is quoted per catalogue (m) was increased by factor of three to mimic the minimum stack depth required within an individual (n). The total number of loci/stacks created in the catalogue is shown (Unique stacks), with the number single-nucleotide polymorphisms (SNPs) found. Also shown is the number of 'blacklisted' stacks and subsequent final number of stacks within the catalogue (Total). The total counts of RAD tags specific to a catalogue locus within the Superparent (Parent) and increasing numbers of progeny (prog1 – prog20). Finally the number of tag counts, which contain a minimum of one SNP in one to five progeny are listed (prog1-5).

Table S8 Summary of the six trialled Stacks eRAD 'SUPER' catalogues. Please see main text for the full table description.

	m	**3**	**6**	**9**	**12**	**15**	**33**
SUPER Catalogue summary	**eg. n**	1	2	3	4	5	11
	Unique stacks	220,523	110,418	82,013	68,038	59,259	38,634
	SNPs found	15,436,610	7,729,260	5,740,910	4,762,660	4,148,130	2,704,380
	Blacklisted	11,890	5,737	3,862	2,926	2,360	1,153
	Total	208,633	104,681	78,151	65,112	56,899	37,481
Tag Counts	**Parent**	208,610	104,675	78,151	65,112	56,897	37,479
	prog 1	89,281	80,704	71,658	63,319	56,393	37,418
	prog 2	50,660	55,977	57,194	55,993	53,241	37,355
	prog 3	38,791	43,929	45,673	46,513	46,557	37,076
	prog 4	31,867	36,873	38,548	39,445	39,987	36,299
	prog 5	27,024	31,949	33,570	34,441	34,988	34,639
	prog 6	23,535	28,363	29,968	30,803	31,341	32,142
	prog 7	20,640	25,376	26,952	27,779	28,291	29,398
	prog 8	18,109	22,757	24,309	25,132	25,638	26,783
	prog 9	15,988	20,561	22,093	22,902	23,395	24,521
	prog 10	14,073	18,564	20,070	20,865	21,362	22,470
	prog 20	1,060	3,271	4,247	4,712	4,915	5,061
Tag Counts, SNP = 1	**prog 1**	5,440	5,246	4,504	3,702	3,172	1,706
	prog 2	2,567	2,952	3,155	3,031	2,819	1,673
	prog 3	1,817	2,011	2,170	2,284	2,298	1,622
	prog 4	1,386	1,482	1,570	1,677	1,761	1,541
	prog 5	1,130	1,188	1,239	1,318	1,389	1,425

Table S9 Summary of the six trialled Stacks eRAD 'SUPER6' catalogues. Please see main text for the full table description.

	m	3	6	9	12	15
SUPER6 Catalogue summary	eg. n	1	2	3	4	5
	Unique stacks	163,195	83,180	61,887	51,557	45,295
	SNPs found	11,423,650	5,822,600	4,332,090	3,608,990	3,170,650
	Blacklisted	7,319	3,352	2,230	1,649	1,322
	Total	155,876	79,828	59,657	49,908	43,973
Tag Counts	Parent	155,866	79,828	59,655	49,906	43,971
	prog 1	76,250	67,361	57,477	49,485	43,821
	prog 2	44,460	47,914	48,142	45,958	42,622
	prog 3	33,913	37,198	38,254	38,651	38,214
	prog 4	27,542	30,683	31,662	32,227	32,515
	prog 5	22,911	25,975	26,926	27,458	27,755
	prog 6	19,115	22,083	23,009	23,519	23,807
	prog 7	15,350	18,218	19,113	19,598	19,865
	prog 8	12,032	14,787	15,651	16,102	16,332
	prog 9	8,881	11,461	12,258	12,623	12,758
	prog 10	5,726	7,952	8,558	8,744	8,791
	prog 20	1	-	-	-	-
Tag Counts, SNP = 1	prog 1	4,423	4,195	3,433	2,731	2,298
	prog 2	1,960	2,368	2,495	2,335	2,110
	prog 3	1,323	1,552	1,720	1,789	1,774
	prog 4	979	1,097	1,191	1,284	1,350
	prog 5	753	836	904	963	1,018

Table S10 Summary of the six trialled Stacks eRAD 'SUPER7' catalogues. Please see main text for the full table description.

	m	3	6	9	12	15
SUPER7 Catalogue summary	eg. n	1	2	3	4	5
	Unique stacks	103,657	59,684	47,134	40,275	35,998
	SNPs found	7,255,990	4,177,880	3,299,380	2,819,250	2,519,860
	Blacklisted	7,646	3,732	2,543	1,945	1,543
	Total	96,011	55,952	44,591	38,330	34,455
Tag Counts	Parent	96,005	55,944	44,591	38,330	34,455
	prog 1	59,183	51,840	43,989	38,205	34,407
	prog 2	33,173	36,954	37,262	35,547	33,302
	prog 3	23,879	27,540	28,711	29,192	29,109
	prog 4	18,487	22,088	23,242	23,809	24,194
	prog 5	14,697	18,243	19,358	19,922	20,324
	prog 6	11,714	15,192	16,288	16,839	17,229
	prog 7	9,083	12,441	13,499	14,039	14,416
	prog 8	6,704	9,868	10,877	11,388	11,732
	prog 9	4,398	7,272	8,189	8,614	8,844
	prog 10	1,515	3,518	4,151	4,351	4,420
	prog 20	-	-	-	-	-
Tag Counts, SNP = 1	prog 1	2,524	2,335	1,877	1,577	1,365
	prog 2	1,190	1,429	1,470	1,381	1,280
	prog 3	794	911	1,012	1,047	1,050
	prog 4	575	659	735	768	801
	prog 5	421	481	524	547	583

Table S11 Summary of the six trialled Stacks eRAD 'FULLFAT' catalogues. Please see main text for the full table description.

	m	3	6	9	12	15	33
FULLFAT Catalogue summary	eg. n	1	2	3	4	5	11
	Unique Stacks	301,514	160,363	117,786	95,752	82,475	52,059
	SNPs Found	21,105,980	11,225,410	8,245,020	6,702,640	5,773,250	3,644,130
	Blacklisted	13,257	6,616	4,513	3,426	2,765	1,380
	Total	288,257	153,747	113,273	92,326	79,710	50,679
Tag Counts	Parent	288,233	153,731	113,264	92,324	79,710	50,679
	prog 1	194,660	141,228	110,710	91,621	79,368	50,518
	prog 2	77,130	83,241	82,657	78,481	73,088	50,231
	prog 3	54,263	59,976	62,026	62,768	62,320	49,129
	prog 4	42,518	47,985	49,918	50,967	51,592	46,805
	prog 5	34,913	40,249	42,088	43,100	43,758	43,183
	prog 6	29,422	34,636	36,426	37,378	38,014	38,950
	prog 7	25,182	30,306	32,059	32,991	33,598	34,845
	prog 8	21,709	26,733	28,458	29,378	29,969	31,221
	prog 9	18,990	23,921	25,614	26,525	27,105	38,323
	prog 10	16,601	21,451	23,135	24,033	24,601	25,781
	prog 20	1,397	3,754	4,840	5,351	5,573	5,703
Tag Counts, SNP = 1	SNPs 1 prog 1	10,869	9,500	7,365	5,903	4,968	2,704
	SNPs 1 prog 2	4,662	5,539	5,449	4,944	4,449	2,640
	SNPs 1 prog 3	2,758	3,179	3,515	3,604	3,563	2,497
	SNPs 1 prog 4	2,023	2,225	2,409	2,574	2,703	2,304
	SNPs 1 prog 5	1,615	1,723	1,815	1,929	2,046	2,044

Table S12 Summary of the six trialled Stacks eRAD 'FULLFAT6' catalogues. Please see main text for the full table description.

	m	3	6	9	12	15
FULLFAT6 Catalogue summary	eg. n	1	2	3	4	5
	Unique Stacks	197,518	98,440	73,146	60,930	53,411
	SNPs Found	13,826,260	6,890,800	5,120,220	4,265,100	3,738,770
	Blacklisted	8,053	3,744	2,524	1,906	1,552
	Total	189,465	94,696	70,622	59,024	51,859
Tag Counts	Parent	189,459	94,692	70,616	59,024	51,859
	prog 1	117,450	86,169	69,212	58,632	51,604
	prog 2	55,594	59,506	58,522	54,636	49,976
	prog 3	40,045	43,677	44,847	45,143	44,276
	prog 4	31,504	34,939	36,025	36,637	36,922
	prog 5	25,684	29,005	30,031	30,602	30,943
	prog 6	20,911	24,140	25,134	25,665	25,968
	prog 7	16,613	19,721	20,674	21,178	21,459
	prog 8	12,836	15,797	16,709	17,162	17,383
	prog 9	9,437	12,205	13,042	13,390	13,509
	prog 10	6,066	8,478	9,092	9,263	9,295
	prog 20	2	-	-	-	-
Tag Counts, SNP = 1	SNPs 1 prog 1	6,974	5,850	4,360	3,526	2,980
	SNPs 1 prog 2	2,717	3,336	3,325	3,019	2,708
	SNPs 1 prog 3	1,674	1,965	2,211	2,292	2,232
	SNPs 1 prog 4	1,222	1,356	1,511	1,632	1,689
	SNPs 1 prog 5	957	1,023	1,111	1,192	1,261

Table S13 Summary of the six trialled Stacks eRAD 'FULLFAT7' catalogues. Please see main text for the full table description.

	m	3	6	9	12	15
FULLFAT7 Catalogue summary	eg. n	1	2	3	4	5
	Unique Stacks	173,399	107,893	82,063	67,109	57,927
	SNPs Found	12,137,930	7,552,510	5,744,410	4,697,630	4,054,890
	Blacklisted	8,830	4,558	3,055	2,341	1,846
	Total	164,569	103,335	79,008	64,768	56,081
Tag Counts	Parent	164,561	103,326	79,003	64,766	56,079
	prog 1	144,085	101,996	78,637	64,544	55,916
	prog 2	52,697	56,885	56,112	53,018	49,688
	prog 3	34,123	38,075	39,492	39,966	39,829
	prog 4	24,833	28,675	30,037	30,691	31,156
	prog 5	18,828	22,597	23,922	24,555	25,031
	prog 6	14,509	18,205	19,525	20,137	20,601
	prog 7	11,093	14,677	15,961	16,552	17,001
	prog 8	8,109	11,490	12,718	13,279	13,679
	prog 9	5,331	8,375	9,486	9,951	10,215
	prog 10	2,434	4,646	5,374	5,588	5,680
	prog 20	-	-	-	-	-
Tag Counts, SNP = 1	SNPs 1 prog 1	6,012	4,989	3,922	3,247	2,747
	SNPs 1 prog 2	2,637	3,074	2,906	2,611	2,342
	SNPs 1 prog 3	1,423	1,595	1,737	1,762	1,726
	SNPs 1 prog 4	954	1,032	1,106	1,156	1,190
	SNPs 1 prog 5	664	706	743	772	802

S15. EdgeR: summary statistics

S15.1 SUPER

```
SUPER_full (SUPER_mono_counts.csv)
> summary(x)
 X7DS_8500.fil     X7DS_8560.fil      X7DS_9001.fil     X7DS_8522.fil     X7DS_8530.fil
 Min.   :  0.00    Min.   :  0.000    Min.   :  0.00    Min.   :  0.00    Min.   :  0.0
 1st Qu.:  0.00    1st Qu.:  0.000    1st Qu.:  0.00    1st Qu.:  0.00    1st Qu.:  0.0
 Median :  0.00    Median :  0.000    Median :  8.00    Median :  0.00    Median :  0.0
 Mean   : 13.13    Mean   :  6.363    Mean   : 26.78    Mean   :  7.47    Mean   :  2.7
 3rd Qu.: 11.00    3rd Qu.:  6.000    3rd Qu.: 27.00    3rd Qu.:  6.00    3rd Qu.:  0.0
 Max.   :457.00    Max.   :300.000    Max.   :490.00    Max.   :255.00    Max.   :173.0

 X7DD_8582.fil      X7DD_8502.fil      X7SS_8531.fil     X7SS_9000.fil     X7SS_8867.fil
 Min.   : 0.0000    Min.   :  0.000    Min.   :  0.00    Min.   : 0.000    Min.   :  0.000
 1st Qu.: 0.0000    1st Qu.:  0.000    1st Qu.:  0.00    1st Qu.: 0.000    1st Qu.:  0.000
 Median : 0.0000    Median :  0.000    Median :  0.00    Median : 0.000    Median :  0.000
 Mean   : 0.6765    Mean   :  9.859    Mean   :  4.53    Mean   : 0.491    Mean   :  3.115
 3rd Qu.: 0.0000    3rd Qu.:  7.000    3rd Qu.:  4.00    3rd Qu.: 0.000    3rd Qu.:  3.000
 Max.   :97.0000    Max.   :385.000    Max.   :193.00    Max.   :96.000    Max.   :182.000

 X7SS_8587.fil     X6DS_9013.fil      X6DS_8559.fil      X6DS_8869.fil     X6DS_8562.fil
 Min.   :  0.00    Min.   :  0.000    Min.   :  0.000    Min.   :  0.00    Min.   :  0.000
 1st Qu.:  0.00    1st Qu.:  0.000    1st Qu.:  0.000    1st Qu.:  0.00    1st Qu.:  0.000
 Median :  3.00    Median :  0.000    Median :  0.000    Median :  0.00    Median :  0.000
 Mean   : 12.99    Mean   :  7.061    Mean   :  7.289    Mean   : 10.41    Mean   :  5.282
 3rd Qu.: 12.00    3rd Qu.:  5.000    3rd Qu.:  7.000    3rd Qu.:  8.00    3rd Qu.:  4.000
 Max.   :415.00    Max.   :477.000    Max.   :362.000    Max.   :463.00    Max.   :537.000

 X6DS_8544.fil     X6DD_9014.fil     X6DD_8515.fil     X6SS_8862.fil     X6SS_8808.fil
 Min.   :  0.00    Min.   :  0.00    Min.   :  0.00    Min.   :  0.00    Min.   :  0.000
 1st Qu.:  0.00    1st Qu.:  0.00    1st Qu.:  0.00    1st Qu.:  0.00    1st Qu.:  0.000
 Median :  0.00    Median :  4.00    Median :  4.00    Median :  4.00    Median :  0.000
 Mean   :  1.94    Mean   : 13.29    Mean   : 16.95    Mean   : 14.41    Mean   :  1.718
 3rd Qu.:  0.00    3rd Qu.: 13.00    3rd Qu.: 14.00    3rd Qu.: 13.00    3rd Qu.:  0.000
 Max.   :444.00    Max.   :553.00    Max.   :546.00    Max.   :657.00    Max.   :114.000

 X6SS_9007.fil     X6SS_9009.fil
 Min.   :  0.00    Min.   :  0.00
 1st Qu.:  0.00    1st Qu.:  0.00
 Median :  4.00    Median :  6.00
 Mean   : 15.75    Mean   : 25.02
 3rd Qu.: 14.00    3rd Qu.: 23.00
 Max.   :570.00    Max.   :674.00
```

S15.2 SUPER_3Q

```
SUPER_3Q (SUPER_mono_2_counts.csv)
> summary(x)
  DS_8500.fil       DS_8560.fil       DS_9001.fil       DS_8522.fil       DD_8502.fil       SS_8531.fil
 Min.   :  0.00    Min.   :  0.000    Min.   :  0.00    Min.   :  0.00    Min.   :  0.000    Min.   :  0.00
 1st Qu.:  0.00    1st Qu.:  0.000    1st Qu.:  0.00    1st Qu.:  0.00    1st Qu.:  0.000    1st Qu.:  0.00
 Median :  0.00    Median :  0.000    Median :  8.00    Median :  0.00    Median :  0.000    Median :  0.00
 Mean   : 13.13    Mean   :  6.363    Mean   : 26.78    Mean   :  7.47    Mean   :  9.859    Mean   :  4.53
 3rd Qu.: 11.00    3rd Qu.:  6.000    3rd Qu.: 27.00    3rd Qu.:  6.00    3rd Qu.:  7.000    3rd Qu.:  4.00
 Max.   :457.00    Max.   :300.000    Max.   :490.00    Max.   :255.00    Max.   :385.000    Max.   :193.00

  SS_8867.fil       SS_8587.fil       DS_9013.fil       DS_8559.fil       DS_8869.fil       DS_8562.fil
 Min.   :  0.000    Min.   :  0.00    Min.   :  0.000    Min.   :  0.000    Min.   :  0.00    Min.   :  0.000
 1st Qu.:  0.000    1st Qu.:  0.00    1st Qu.:  0.000    1st Qu.:  0.000    1st Qu.:  0.00    1st Qu.:  0.000
 Median :  0.000    Median :  3.00    Median :  0.000    Median :  0.000    Median :  0.00    Median :  0.000
 Mean   :  3.115    Mean   : 12.99    Mean   :  7.061    Mean   :  7.289    Mean   : 10.41    Mean   :  5.282
 3rd Qu.:  3.000    3rd Qu.: 12.00    3rd Qu.:  5.000    3rd Qu.:  7.000    3rd Qu.:  8.00    3rd Qu.:  4.000
 Max.   :182.000    Max.   :415.00    Max.   :477.000    Max.   :362.000    Max.   :463.00    Max.   :537.000

  DD_9014.fil       DD_8515.fil       SS_8862.fil       SS_9007.fil       SS_9009.fil
 Min.   :  0.00    Min.   :  0.00    Min.   :  0.00    Min.   :  0.00    Min.   :  0.00
 1st Qu.:  0.00    1st Qu.:  0.00    1st Qu.:  0.00    1st Qu.:  0.00    1st Qu.:  0.00
 Median :  4.00    Median :  4.00    Median :  4.00    Median :  4.00    Median :  6.00
 Mean   : 13.29    Mean   : 16.95    Mean   : 14.41    Mean   : 15.75    Mean   : 25.02
 3rd Qu.: 13.00    3rd Qu.: 14.00    3rd Qu.: 13.00    3rd Qu.: 14.00    3rd Qu.: 23.00
 Max.   :553.00    Max.   :546.00    Max.   :657.00    Max.   :570.00    Max.   :674.00
```

S15.3 SUPER_Bd

```
SUPERBd (SUPER_Bd_mono_counts.csv)
> summary(x)
 X7DS_8500.fil     X7DS_8560.fil      X7DD_8582.fil      X7DD_8502.fil      X7SS_8531.fil
 Min.   :  0.00    Min.   :  0.000    Min.   : 0.0000    Min.   :  0.000    Min.   :  0.00
 1st Qu.:  0.00    1st Qu.:  0.000    1st Qu.: 0.0000    1st Qu.:  0.000    1st Qu.:  0.00
 Median :  0.00    Median :  0.000    Median : 0.0000    Median :  0.000    Median :  0.00
 Mean   : 13.13    Mean   :  6.363    Mean   : 0.6765    Mean   :  9.859    Mean   :  4.53
 3rd Qu.: 11.00    3rd Qu.:  6.000    3rd Qu.: 0.0000    3rd Qu.:  7.000    3rd Qu.:  4.00
 Max.   :457.00    Max.   :300.000    Max.   :97.0000    Max.   :385.000    Max.   :193.00

 X7SS_8587.fil     X6DS_9013.fil      X6DS_8559.fil      X6DD_9014.fil     X6DD_8515.fil
 Min.   :  0.00    Min.   :  0.000    Min.   :  0.000    Min.   :  0.00    Min.   :  0.00
 1st Qu.:  0.00    1st Qu.:  0.000    1st Qu.:  0.000    1st Qu.:  0.00    1st Qu.:  0.00
 Median :  3.00    Median :  0.000    Median :  0.000    Median :  4.00    Median :  4.00
 Mean   : 12.99    Mean   :  7.061    Mean   :  7.289    Mean   : 13.29    Mean   : 16.95
 3rd Qu.: 12.00    3rd Qu.:  5.000    3rd Qu.:  7.000    3rd Qu.: 13.00    3rd Qu.: 14.00
 Max.   :415.00    Max.   :477.000    Max.   :362.000    Max.   :553.00    Max.   :546.00

 X6SS_8862.fil     X6SS_9009.fil
 Min.   :  0.00    Min.   :  0.00
 1st Qu.:  0.00    1st Qu.:  0.00
 Median :  4.00    Median :  6.00
 Mean   : 14.41    Mean   : 25.02
 3rd Qu.: 13.00    3rd Qu.: 23.00
 Max.   :657.00    Max.   :674.00
```

S15.4 SUPER_Rm

```
SUPER_Rm (SUPER_Rm_mono_counts.csv)
> summary(x)
 DS_8500.fil       DS_8560.fil        DD_8502.fil        SS_8531.fil       DS_8559.fil
 Min.   :  0.00    Min.   :  0.000    Min.   :  0.000    Min.   :  0.00    Min.   :  0.000
 1st Qu.:  0.00    1st Qu.:  0.000    1st Qu.:  0.000    1st Qu.:  0.00    1st Qu.:  0.000
 Median :  0.00    Median :  0.000    Median :  0.000    Median :  0.00    Median :  0.000
 Mean   : 13.13    Mean   :  6.363    Mean   :  9.859    Mean   :  4.53    Mean   :  7.289
 3rd Qu.: 11.00    3rd Qu.:  6.000    3rd Qu.:  7.000    3rd Qu.:  4.00    3rd Qu.:  7.000
 Max.   :457.00    Max.   :300.000    Max.   :385.000    Max.   :193.00    Max.   :362.000

 DD_9014.fil       DD_8515.fil        SS_8862.fil        SS_9009.fil
 Min.   :  0.00    Min.   :  0.00     Min.   :  0.00     Min.   :  0.00
 1st Qu.:  0.00    1st Qu.:  0.00     1st Qu.:  0.00     1st Qu.:  0.00
 Median :  4.00    Median :  4.00     Median :  4.00     Median :  6.00
 Mean   : 13.29    Mean   : 16.95     Mean   : 14.41     Mean   : 25.02
 3rd Qu.: 13.00    3rd Qu.: 14.00     3rd Qu.: 13.00     3rd Qu.: 23.00
 Max.   :553.00    Max.   :546.00     Max.   :657.00     Max.   :674.00
```

S15.5 SUPER6

```
SUPER6full (SUPER6_mono_counts.csv)
> summary(x)
  DS_9013.fil       DS_8559.fil       DS_8869.fil       DS_8562.fil       DS_8544.fil
 Min.   :  0.000   Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.000
 1st Qu.:  0.000   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.000
 Median :  0.000   Median :  3.00   Median :  3.00   Median :  0.00   Median :  0.000
 Mean   :  9.377   Mean   : 10.03   Mean   : 13.16   Mean   :  7.11   Mean   :  2.797
 3rd Qu.:  7.000   3rd Qu.:  9.00   3rd Qu.: 11.00   3rd Qu.:  6.00   3rd Qu.:  0.000
 Max.   :508.000   Max.   :478.00   Max.   :493.00   Max.   :537.00   Max.   :444.000

  DD_9014.fil       DD_8515.fil       SS_8862.fil       SS_8808.fil       SS_9007.fil
 Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.000   Min.   :  0.00
 1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.000   1st Qu.:  0.00
 Median :  6.00   Median :  6.00   Median :  6.00   Median :  0.000   Median :  5.00
 Mean   : 17.09   Mean   : 21.28   Mean   : 18.52   Mean   :  2.445   Mean   : 20.36
 3rd Qu.: 17.00   3rd Qu.: 18.00   3rd Qu.: 16.00   3rd Qu.:  0.000   3rd Qu.: 18.00
 Max.   :657.00   Max.   :576.00   Max.   :715.00   Max.   :430.000   Max.   :570.00

  SS_9009.fil
 Min.   :  0.00
 1st Qu.:  3.00
 Median :  9.00
 Mean   : 31.09
 3rd Qu.: 30.00
 Max.   :674.00
```

S15.6 SUPER6_3Q

```
SUPER6_3Q (SUPER6_3Q_mono_counts.csv)
> summary(x)
  DS_9013.fil       DS_8559.fil       DS_8869.fil       DS_8562.fil       DD_9014.fil
 Min.   :  0.000   Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
 1st Qu.:  0.000   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00
 Median :  0.000   Median :  3.00   Median :  3.00   Median :  0.00   Median :  6.00
 Mean   :  9.377   Mean   : 10.03   Mean   : 13.16   Mean   :  7.11   Mean   : 17.09
 3rd Qu.:  7.000   3rd Qu.:  9.00   3rd Qu.: 11.00   3rd Qu.:  6.00   3rd Qu.: 17.00
 Max.   :508.000   Max.   :478.00   Max.   :493.00   Max.   :537.00   Max.   :657.00

  DD_8515.fil       SS_8862.fil       SS_9007.fil       SS_9009.fil
 Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
 1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  3.00
 Median :  6.00   Median :  6.00   Median :  5.00   Median :  9.00
 Mean   : 21.28   Mean   : 18.52   Mean   : 20.36   Mean   : 31.09
 3rd Qu.: 18.00   3rd Qu.: 16.00   3rd Qu.: 18.00   3rd Qu.: 30.00
 Max.   :576.00   Max.   :715.00   Max.   :570.00   Max.   :674.00
```

S15.7 SUPER6_Bd

```
SUPER6_Bd (SUPER6_Bd_mono_counts.csv)
> summary(x)
  DS_9013.fil       DS_8559.fil       DD_9014.fil       DD_8515.fil       SS_8862.fil
 Min.   :  0.000   Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
 1st Qu.:  0.000   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00
 Median :  0.000   Median :  3.00   Median :  6.00   Median :  6.00   Median :  6.00
 Mean   :  9.377   Mean   : 10.03   Mean   : 17.09   Mean   : 21.28   Mean   : 18.52
 3rd Qu.:  7.000   3rd Qu.:  9.00   3rd Qu.: 17.00   3rd Qu.: 18.00   3rd Qu.: 16.00
 Max.   :508.000   Max.   :478.00   Max.   :657.00   Max.   :576.00   Max.   :715.00

  SS_9009.fil
 Min.   :  0.00
 1st Qu.:  3.00
 Median :  9.00
 Mean   : 31.09
 3rd Qu.: 30.00
 Max.   :674.00
```

S15.8 SUPER7

```
SUPER7full (SUPER7_mono_counts.csv)
> summary(x)
  DS_8500.fil      DS_8560.fil       DS_9001.fil      DS_8522.fil      DS_8530.fil
 Min.   :  0.00   Min.   :  0.000   Min.   :  0.00   Min.   :  0.00   Min.   :  0.000
 1st Qu.:  0.00   1st Qu.:  0.000   1st Qu.:  7.00   1st Qu.:  0.00   1st Qu.:  0.000
 Median :  6.00   Median :  3.000   Median : 17.00   Median :  3.00   Median :  0.000
 Mean   : 19.41   Mean   :  9.498   Mean   : 38.63   Mean   : 11.18   Mean   :  3.978
 3rd Qu.: 20.00   3rd Qu.: 10.000   3rd Qu.: 46.00   3rd Qu.: 11.00   3rd Qu.:  4.000
 Max.   :383.00   Max.   :262.000   Max.   :490.00   Max.   :216.00   Max.   :119.000

  DD_8582.fil      DD_8502.fil      SS_8531.fil       SS_9000.fil       SS_8867.fil
 Min.   : 0.000   Min.   :  0.00   Min.   :  0.000   Min.   : 0.0000   Min.   :  0.000
 1st Qu.: 0.000   1st Qu.:  0.00   1st Qu.:  0.000   1st Qu.: 0.0000   1st Qu.:  0.000
 Median : 0.000   Median :  4.00   Median :  0.000   Median : 0.0000   Median :  0.000
 Mean   : 1.011   Mean   : 14.42   Mean   :  6.816   Mean   : 0.7342   Mean   :  4.635
 3rd Qu.: 0.000   3rd Qu.: 14.00   3rd Qu.:  7.000   3rd Qu.: 0.0000   3rd Qu.:  5.000
 Max.   :97.000   Max.   :385.00   Max.   :156.000   Max.   :96.0000   Max.   :156.000

  SS_8587.fil
 Min.   :  0.00
 1st Qu.:  0.00
 Median :  7.00
 Mean   : 18.81
 3rd Qu.: 20.00
 Max.   :415.00
```

S15.9 SUPER7_3Q

```
SUPER7_3Q (SUPER7_3Q_mono_counts.csv)
> summary(x)
  DS_8500.fil      DS_8560.fil       DS_9001.fil      DS_8522.fil      DS_8530.fil
 Min.   :  0.00   Min.   :  0.000   Min.   :  0.00   Min.   :  0.00   Min.   :  0.000
 1st Qu.:  0.00   1st Qu.:  0.000   1st Qu.:  7.00   1st Qu.:  0.00   1st Qu.:  0.000
 Median :  6.00   Median :  3.000   Median : 17.00   Median :  3.00   Median :  0.000
 Mean   : 19.41   Mean   :  9.498   Mean   : 38.63   Mean   : 11.18   Mean   :  3.978
 3rd Qu.: 20.00   3rd Qu.: 10.000   3rd Qu.: 46.00   3rd Qu.: 11.00   3rd Qu.:  4.000
 Max.   :383.00   Max.   :262.000   Max.   :490.00   Max.   :216.00   Max.   :119.000

  DD_8502.fil      SS_8531.fil       SS_8867.fil       SS_8587.fil
 Min.   :  0.00   Min.   :  0.000   Min.   :  0.000   Min.   :  0.00
 1st Qu.:  0.00   1st Qu.:  0.000   1st Qu.:  0.000   1st Qu.:  0.00
 Median :  4.00   Median :  0.000   Median :  0.000   Median :  7.00
 Mean   : 14.42   Mean   :  6.816   Mean   :  4.635   Mean   : 18.81
 3rd Qu.: 14.00   3rd Qu.:  7.000   3rd Qu.:  5.000   3rd Qu.: 20.00
 Max.   :385.00   Max.   :156.000   Max.   :156.000   Max.   :415.00
```

S15.10 SUPER7_Bd

```
SUPER7_Bd (SUPER7_Bd_mono_counts.csv)
> summary(x)
  DS_8500.fil      DS_8560.fil       DD_8582.fil      DD_8502.fil      SS_8531.fil
 Min.   :  0.00   Min.   :  0.000   Min.   : 0.000   Min.   :  0.00   Min.   :  0.000
 1st Qu.:  0.00   1st Qu.:  0.000   1st Qu.: 0.000   1st Qu.:  0.00   1st Qu.:  0.000
 Median :  6.00   Median :  3.000   Median : 0.000   Median :  4.00   Median :  0.000
 Mean   : 19.41   Mean   :  9.498   Mean   : 1.011   Mean   : 14.42   Mean   :  6.816
 3rd Qu.: 20.00   3rd Qu.: 10.000   3rd Qu.: 0.000   3rd Qu.: 14.00   3rd Qu.:  7.000
 Max.   :383.00   Max.   :262.000   Max.   :97.000   Max.   :385.00   Max.   :156.000

  SS_8587.fil
 Min.   :  0.00
 1st Qu.:  0.00
 Median :  7.00
 Mean   : 18.81
 3rd Qu.: 20.00
 Max.   :415.00
```

S15.11 FULLFAT

```
FULLFATfull (Fullfat_mono_counts.csv)
> summary(x)
   X7DS_8500         X7DS_8560         X7DS_9001         X7DS_8522         X7DS_8530
 Min.   :   0.00   Min.   :   0.00   Min.   :   0.00   Min.   :  0.000   Min.   :  0.000
 1st Qu.:   0.00   1st Qu.:   0.00   1st Qu.:   0.00   1st Qu.:  0.000   1st Qu.:  0.000
 Median :   0.00   Median :   0.00   Median :  10.00   Median :  0.000   Median :  0.000
 Mean   :  22.42   Mean   :  13.17   Mean   :  49.28   Mean   :  7.639   Mean   :  3.103
 3rd Qu.:  13.00   3rd Qu.:   8.00   3rd Qu.:  35.00   3rd Qu.:  4.000   3rd Qu.:  0.000
 Max.   :1254.00   Max.   :1205.00   Max.   :1728.00   Max.   :666.000   Max.   :686.000

   X7DD_8582         X7DD_8502         X7SS_8531         X7SS_9000         X7SS_8867
 Min.   :  0.000   Min.   :   0.00   Min.   :  0.00   Min.   :  0.0000   Min.   :  0.000
 1st Qu.:  0.000   1st Qu.:   0.00   1st Qu.:  0.00   1st Qu.:  0.0000   1st Qu.:  0.000
 Median :  0.000   Median :   0.00   Median :  0.00   Median :  0.0000   Median :  0.000
 Mean   :  1.289   Mean   :  20.51   Mean   :  4.32   Mean   :  0.9433   Mean   :  3.601
 3rd Qu.:  0.000   3rd Qu.:   9.00   3rd Qu.:  0.00   3rd Qu.:  0.0000   3rd Qu.:  0.000
 Max.   :686.000   Max.   :1391.00   Max.   :653.00   Max.   :449.0000   Max.   :464.000

   X7SS_8587         X6DS_9013         X6DS_8559         X6DS_8869         X6DS_8562
 Min.   :   0.00   Min.   :   0.000   Min.   :  0.000   Min.   :   0.000   Min.   :  0.000
 1st Qu.:   0.00   1st Qu.:   0.000   1st Qu.:  0.000   1st Qu.:   0.000   1st Qu.:  0.000
 Median :   4.00   Median :   0.000   Median :  0.000   Median :   0.000   Median :  0.000
 Mean   :  25.97   Mean   :   6.748   Mean   :  6.868   Mean   :   9.879   Mean   :  4.899
 3rd Qu.:  18.00   3rd Qu.:   4.000   3rd Qu.:  5.000   3rd Qu.:   5.000   3rd Qu.:  3.000
 Max.   :1587.00   Max.   :1051.000   Max.   :787.000   Max.   :1256.000   Max.   :853.000

   X6DS_8544         X6DD_9014         X6DD_8515         X6SS_8862         X6SS_8808
 Min.   :   0.000   Min.   :  0.00   Min.   :   0.00   Min.   :   0.00   Min.   :  0.000
 1st Qu.:   0.000   1st Qu.:  0.00   1st Qu.:   0.00   1st Qu.:   0.00   1st Qu.:  0.000
 Median :   0.000   Median :  3.00   Median :   0.00   Median :   0.00   Median :  0.000
 Mean   :   1.846   Mean   : 12.88   Mean   :  17.09   Mean   :  13.73   Mean   :  1.519
 3rd Qu.:   0.000   3rd Qu.: 11.00   3rd Qu.:  10.00   3rd Qu.:  10.00   3rd Qu.:  0.000
 Max.   :1289.000   Max.   :758.00   Max.   :1369.00   Max.   :1483.00   Max.   :382.000

   X6SS_9007         X6SS_9009
 Min.   :   0.00   Min.   :   0.00
 1st Qu.:   0.00   1st Qu.:   0.00
 Median :   0.00   Median :   4.00
 Mean   :  15.52   Mean   :  26.01
 3rd Qu.:  10.00   3rd Qu.:  16.00
 Max.   :1263.00   Max.   :1643.00
```

S15.12 FULLFAT_3Q

```
FULLFAT_3Q (Fullfat_3Q_mono_counts.csv)
> summary(x)
   X7DS_8500         X7DS_8560         X7DS_9001         X7DS_8522
 Min.   :   0.00   Min.   :   0.00   Min.   :   0.00   Min.   :  0.000
 1st Qu.:   0.00   1st Qu.:   0.00   1st Qu.:   0.00   1st Qu.:  0.000
 Median :   0.00   Median :   0.00   Median :  10.00   Median :  0.000
 Mean   :  22.42   Mean   :  13.17   Mean   :  49.28   Mean   :  7.639
 3rd Qu.:  13.00   3rd Qu.:   8.00   3rd Qu.:  35.00   3rd Qu.:  4.000
 Max.   :1254.00   Max.   :1205.00   Max.   :1728.00   Max.   :666.000

   X7DD_8502         X7SS_8531         X7SS_8587         X6DS_9013
 Min.   :   0.00   Min.   :  0.00   Min.   :   0.00   Min.   :   0.000
 1st Qu.:   0.00   1st Qu.:  0.00   1st Qu.:   0.00   1st Qu.:   0.000
 Median :   0.00   Median :  0.00   Median :   4.00   Median :   0.000
 Mean   :  20.51   Mean   :  4.32   Mean   :  25.97   Mean   :   6.748
 3rd Qu.:   9.00   3rd Qu.:  0.00   3rd Qu.:  18.00   3rd Qu.:   4.000
 Max.   :1391.00   Max.   :653.00   Max.   :1587.00   Max.   :1051.000

   X6DS_8559         X6DS_8869         X6DS_8562         X6DD_9014
 Min.   :  0.000   Min.   :   0.000   Min.   :  0.000   Min.   :  0.00
 1st Qu.:  0.000   1st Qu.:   0.000   1st Qu.:  0.000   1st Qu.:  0.00
 Median :  0.000   Median :   0.000   Median :  0.000   Median :  3.00
 Mean   :  6.868   Mean   :   9.879   Mean   :  4.899   Mean   : 12.88
 3rd Qu.:  5.000   3rd Qu.:   5.000   3rd Qu.:  3.000   3rd Qu.: 11.00
 Max.   :787.000   Max.   :1256.000   Max.   :853.000   Max.   :758.00

   X6DD_8515         X6SS_8862         X6SS_9007         X6SS_9009
 Min.   :   0.00   Min.   :   0.00   Min.   :   0.00   Min.   :   0.00
 1st Qu.:   0.00   1st Qu.:   0.00   1st Qu.:   0.00   1st Qu.:   0.00
 Median :   0.00   Median :   0.00   Median :   0.00   Median :   4.00
 Mean   :  17.09   Mean   :  13.73   Mean   :  15.52   Mean   :  26.01
 3rd Qu.:  10.00   3rd Qu.:  10.00   3rd Qu.:  10.00   3rd Qu.:  16.00
 Max.   :1369.00   Max.   :1483.00   Max.   :1263.00   Max.   :1643.00
```

S15.13 FULLFAT_Bd

```
FULLFAT_Bd (Fullfat_Bd_mono_counts.csv)
> summary(x)
   X7DS_8500         X7DS_8560         X7DD_8582         X7DD_8502         X7SS_8531
 Min.   :   0.00   Min.   :   0.00   Min.   :  0.000   Min.   :   0.00   Min.   :  0.00
 1st Qu.:   0.00   1st Qu.:   0.00   1st Qu.:  0.000   1st Qu.:   0.00   1st Qu.:  0.00
 Median :   0.00   Median :   0.00   Median :  0.000   Median :   0.00   Median :  0.00
 Mean   :  22.42   Mean   :  13.17   Mean   :  1.289   Mean   :  20.51   Mean   :  4.32
 3rd Qu.:  13.00   3rd Qu.:   8.00   3rd Qu.:  0.000   3rd Qu.:   9.00   3rd Qu.:  0.00
 Max.   :1254.00   Max.   :1205.00   Max.   :686.000   Max.   :1391.00   Max.   :653.00

   X7SS_8587         X6DS_9013          X6DS_8559         X6DD_9014         X6DD_8515
 Min.   :   0.00   Min.   :   0.000   Min.   :  0.000   Min.   :  0.00   Min.   :   0.00
 1st Qu.:   0.00   1st Qu.:   0.000   1st Qu.:  0.000   1st Qu.:  0.00   1st Qu.:   0.00
 Median :   4.00   Median :   0.000   Median :  0.000   Median :  3.00   Median :   0.00
 Mean   :  25.97   Mean   :   6.748   Mean   :  6.868   Mean   : 12.88   Mean   :  17.09
 3rd Qu.:  18.00   3rd Qu.:   4.000   3rd Qu.:  5.000   3rd Qu.: 11.00   3rd Qu.:  10.00
 Max.   :1587.00   Max.   :1051.000   Max.   :787.000   Max.   :758.00   Max.   :1369.00

   X6SS_8862         X6SS_9009
 Min.   :   0.00   Min.   :   0.00
 1st Qu.:   0.00   1st Qu.:   0.00
 Median :   0.00   Median :   4.00
 Mean   :  13.73   Mean   :  26.01
 3rd Qu.:  10.00   3rd Qu.:  16.00
 Max.   :1483.00   Max.   :1643.00
```

S15.14 FULLFAT6

```
FULLFAT6full (Fullfat6_mono_counts.csv)
> summary(x)
   X6DS_9013          X6DS_8559          X6DS_8869          X6DS_8562          X6DS_8544
 Min.   :   0.00   Min.   :  0.00   Min.   :   0.00   Min.   :  0.000   Min.   :   0.00
 1st Qu.:   0.00   1st Qu.:  0.00   1st Qu.:   0.00   1st Qu.:  0.000   1st Qu.:   0.00
 Median :   0.00   Median :  3.00   Median :   0.00   Median :  0.000   Median :   0.00
 Mean   :  10.97   Mean   : 11.55   Mean   :  14.93   Mean   :  8.158   Mean   :   3.26
 3rd Qu.:   7.00   3rd Qu.:  9.00   3rd Qu.:  10.00   3rd Qu.:  5.000   3rd Qu.:   0.00
 Max.   :1051.00   Max.   :787.00   Max.   :1256.00   Max.   :950.000   Max.   :1289.00

   X6DD_9014          X6DD_8515          X6SS_8862          X6SS_8808          X6SS_9007
 Min.   :   0.00   Min.   :   0.00   Min.   :   0.00   Min.   :  0.000   Min.   :   0.00
 1st Qu.:   0.00   1st Qu.:   0.00   1st Qu.:   0.00   1st Qu.:  0.000   1st Qu.:   0.00
 Median :   6.00   Median :   5.00   Median :   5.00   Median :  0.000   Median :   5.00
 Mean   :  19.71   Mean   :  26.03   Mean   :  21.43   Mean   :  2.606   Mean   :  24.49
 3rd Qu.:  17.00   3rd Qu.:  17.00   3rd Qu.:  16.00   3rd Qu.:  0.000   3rd Qu.:  17.00
 Max.   :1097.00   Max.   :1369.00   Max.   :1483.00   Max.   :908.000   Max.   :1290.00

   X6SS_9009
 Min.   :   0.00
 1st Qu.:   0.00
 Median :   8.00
 Mean   :  38.93
 3rd Qu.:  29.00
 Max.   :1643.00
```

S15.15 FULLFAT6_3Q

```
FULLFAT6_3Q (Fullfat6_3Q_mono_counts.csv)
> summary(x)
   X6DS_9013          X6DS_8559          X6DS_8869          X6DS_8562          X6DD_9014
 Min.   :   0.00   Min.   :  0.00   Min.   :   0.00   Min.   :  0.000   Min.   :   0.00
 1st Qu.:   0.00   1st Qu.:  0.00   1st Qu.:   0.00   1st Qu.:  0.000   1st Qu.:   0.00
 Median :   0.00   Median :  3.00   Median :   0.00   Median :  0.000   Median :   6.00
 Mean   :  10.97   Mean   : 11.55   Mean   :  14.93   Mean   :  8.158   Mean   :  19.71
 3rd Qu.:   7.00   3rd Qu.:  9.00   3rd Qu.:  10.00   3rd Qu.:  5.000   3rd Qu.:  17.00
 Max.   :1051.00   Max.   :787.00   Max.   :1256.00   Max.   :950.000   Max.   :1097.00

   X6DD_8515          X6SS_8862          X6SS_9007          X6SS_9009
 Min.   :   0.00   Min.   :   0.00   Min.   :   0.00   Min.   :   0.00
 1st Qu.:   0.00   1st Qu.:   0.00   1st Qu.:   0.00   1st Qu.:   0.00
 Median :   5.00   Median :   5.00   Median :   5.00   Median :   8.00
 Mean   :  26.03   Mean   :  21.43   Mean   :  24.49   Mean   :  38.93
 3rd Qu.:  17.00   3rd Qu.:  16.00   3rd Qu.:  17.00   3rd Qu.:  29.00
 Max.   :1369.00   Max.   :1483.00   Max.   :1290.00   Max.   :1643.00
```

S15.16 FULLFAT6_Bd

```
FULLFAT6_Bd (Fullfat6_Bd_mono_counts.csv)
> summary(x)
   X6DS_9013          X6DS_8559          X6DD_9014          X6DD_8515          X6SS_8862
 Min.   :   0.00   Min.   :  0.00   Min.   :   0.00   Min.   :   0.00   Min.   :   0.00
 1st Qu.:   0.00   1st Qu.:  0.00   1st Qu.:   0.00   1st Qu.:   0.00   1st Qu.:   0.00
 Median :   0.00   Median :  3.00   Median :   6.00   Median :   5.00   Median :   5.00
 Mean   :  10.97   Mean   : 11.55   Mean   :  19.71   Mean   :  26.03   Mean   :  21.43
 3rd Qu.:   7.00   3rd Qu.:  9.00   3rd Qu.:  17.00   3rd Qu.:  17.00   3rd Qu.:  16.00
 Max.   :1051.00   Max.   :787.00   Max.   :1097.00   Max.   :1369.00   Max.   :1483.00

   X6SS_9009
 Min.   :   0.00
 1st Qu.:   0.00
 Median :   8.00
 Mean   :  38.93
 3rd Qu.:  29.00
 Max.   :1643.00
```

S15.17 FULLFAT7

```
FULLFAT7full (Fullfat7_mono_counts.csv)
> summary(x)
   X7DS_8500         X7DS_8560        X7DS_9001         X7DS_8522         X7DS_8530
 Min.   :   0.00   Min.   :  0.00   Min.   :   0.00   Min.   :  0.000   Min.   :  0.000
 1st Qu.:   0.00   1st Qu.:  0.00   1st Qu.:   5.00   1st Qu.:  0.000   1st Qu.:  0.000
 Median :   4.00   Median :  0.00   Median :  15.00   Median :  0.000   Median :  0.000
 Mean   :  27.45   Mean   : 16.15   Mean   :  60.05   Mean   :  9.303   Mean   :  3.715
 3rd Qu.:  19.00   3rd Qu.: 12.00   3rd Qu.:  49.00   3rd Qu.:  6.000   3rd Qu.:  0.000
 Max.   :1179.00   Max.   :865.00   Max.   :1728.00   Max.   :503.000   Max.   :947.000

   X7DD_8582         X7DD_8502         X7SS_8531        X7SS_9000         X7SS_8867
 Min.   :  0.000   Min.   :   0.00   Min.   :  0.00   Min.   :  0.000   Min.   :  0.000
 1st Qu.:  0.000   1st Qu.:   0.00   1st Qu.:  0.00   1st Qu.:  0.000   1st Qu.:  0.000
 Median :  0.000   Median :   0.00   Median :  0.00   Median :  0.000   Median :  0.000
 Mean   :  1.583   Mean   :  24.79   Mean   :  5.28   Mean   :  1.157   Mean   :  4.367
 3rd Qu.:  0.000   3rd Qu.:  14.00   3rd Qu.:  4.00   3rd Qu.:  0.000   3rd Qu.:  4.000
 Max.   :385.000   Max.   :1589.00   Max.   :647.00   Max.   :449.000   Max.   :728.000

   X7SS_8587
 Min.   :   0.00
 1st Qu.:   0.00
 Median :   8.00
 Mean   :  31.31
 3rd Qu.:  24.00
 Max.   :1587.00
```

S15.18 FULLFAT7_3Q

```
FULLFAT7_3Q (Fullfat7_3Q_mono_counts.csv)
> summary(x)
   X7DS_8500         X7DS_8560        X7DS_9001         X7DS_8522
 Min.   :   0.00   Min.   :  0.00   Min.   :   0.00   Min.   :  0.000
 1st Qu.:   0.00   1st Qu.:  0.00   1st Qu.:   5.00   1st Qu.:  0.000
 Median :   4.00   Median :  0.00   Median :  15.00   Median :  0.000
 Mean   :  27.45   Mean   : 16.15   Mean   :  60.05   Mean   :  9.303
 3rd Qu.:  19.00   3rd Qu.: 12.00   3rd Qu.:  49.00   3rd Qu.:  6.000
 Max.   :1179.00   Max.   :865.00   Max.   :1728.00   Max.   :503.000

   X7DD_8502         X7SS_8531        X7SS_8867         X7SS_8587
 Min.   :   0.00   Min.   :  0.00   Min.   :  0.000   Min.   :   0.00
 1st Qu.:   0.00   1st Qu.:  0.00   1st Qu.:  0.000   1st Qu.:   0.00
 Median :   0.00   Median :  0.00   Median :  0.000   Median :   8.00
 Mean   :  24.79   Mean   :  5.28   Mean   :  4.367   Mean   :  31.31
 3rd Qu.:  14.00   3rd Qu.:  4.00   3rd Qu.:  4.000   3rd Qu.:  24.00
 Max.   :1589.00   Max.   :647.00   Max.   :728.000   Max.   :1587.00
```

S15.19 FULLFAT7_Bd

```
FULLFAT7_Bd (Fullfat7_Bd_mono_counts.csv)
> summary(x)
   X7DS_8500         X7DS_8560        X7DD_8582         X7DD_8502         X7SS_8531
 Min.   :   0.00   Min.   :  0.00   Min.   :  0.000   Min.   :   0.00   Min.   :  0.00
 1st Qu.:   0.00   1st Qu.:  0.00   1st Qu.:  0.000   1st Qu.:   0.00   1st Qu.:  0.00
 Median :   4.00   Median :  0.00   Median :  0.000   Median :   0.00   Median :  0.00
 Mean   :  27.45   Mean   : 16.15   Mean   :  1.583   Mean   :  24.79   Mean   :  5.28
 3rd Qu.:  19.00   3rd Qu.: 12.00   3rd Qu.:  0.000   3rd Qu.:  14.00   3rd Qu.:  4.00
 Max.   :1179.00   Max.   :865.00   Max.   :385.000   Max.   :1589.00   Max.   :647.00

   X7SS_8587
 Min.   :   0.00
 1st Qu.:   0.00
 Median :   8.00
 Mean   :  31.31
 3rd Qu.:  24.00
 Max.   :1587.00
```

S16. EdgeR: Additional differential expression analyses

The removal of individuals from datasets generally did not reduce the amount of variation within the datasets (Figure S34, Figure S35).It did however result in a reduction of individuals within the dataset and in most cases, an increased number of presence/absence relationships in gene expression (Figure S37, Figure S38). The balanced datasets provide interesting insight into the clusters of presence/absence relationships as they should be equally likely in each direction. Clusters were seen indicating a greater tendency for one genotypic group to have absence of expression than the other. However these clusters were observed in both directions and therefore are unlikely to be caused by genuine biological differences (Figure S39, Figure S40).

The additional tags found to be significantly DE in the filtered datasets were all presence/absence relationships. These are assumed to have arisen from the removal of individuals, especially in the DD group as opposed to a genuine biological signal and are therefore not discussed further.

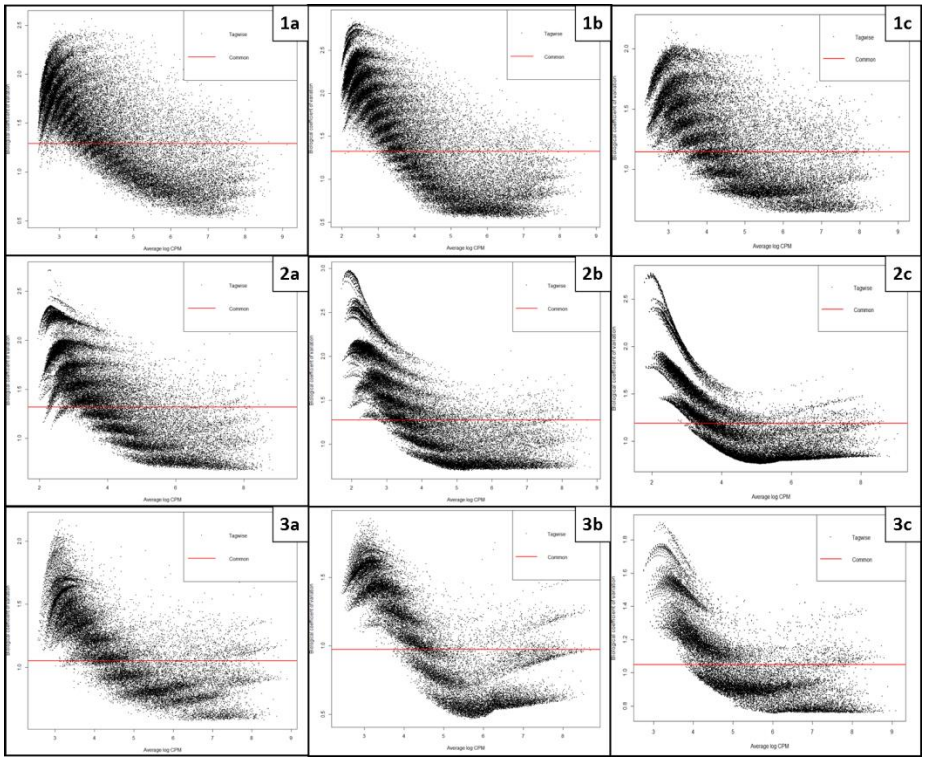


Figure S34 Biological coefficient of variation plots for SUPER (1), SUPER6 (2) and SUPER7 (3), the full dataset (a), the filtered '3Q' (b) and 'Bd' (c) datasets. Each black dot represents the variation for each tag (tagwise variation) and the red line indicates the overall level of variation (common).

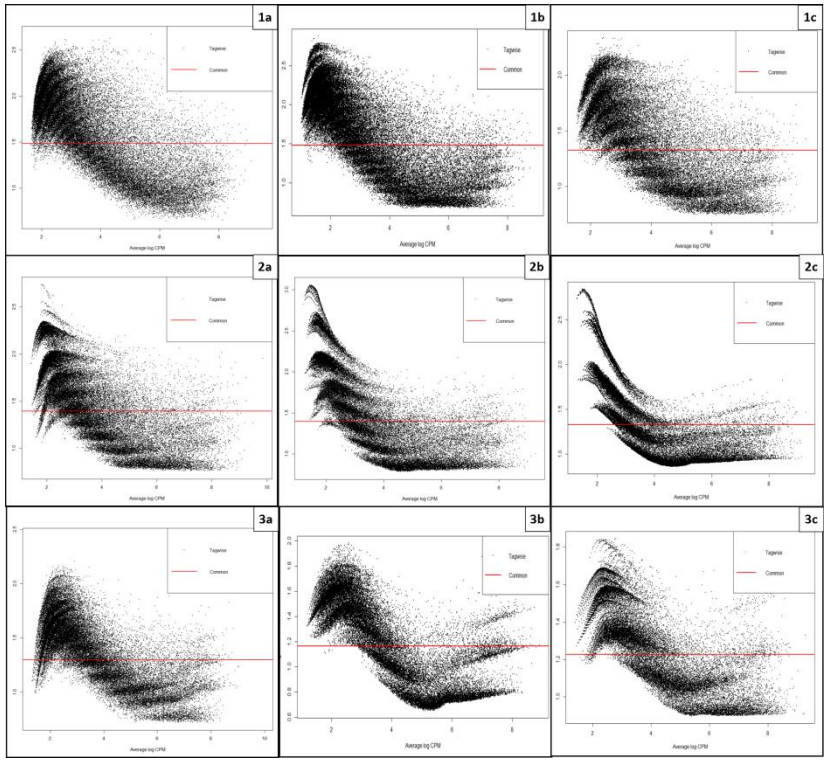


Figure S35 Biological coefficient of variation plots for FULLFAT (1), FULLFAT6 (2) and FULLFAT7 (3), the full dataset (a), the filtered '3Q' (b) and 'Bd' (c) datasets. Each black dot represents the variation for each tag (tagwise variation) and the red line indicates the overall level of variation (common).

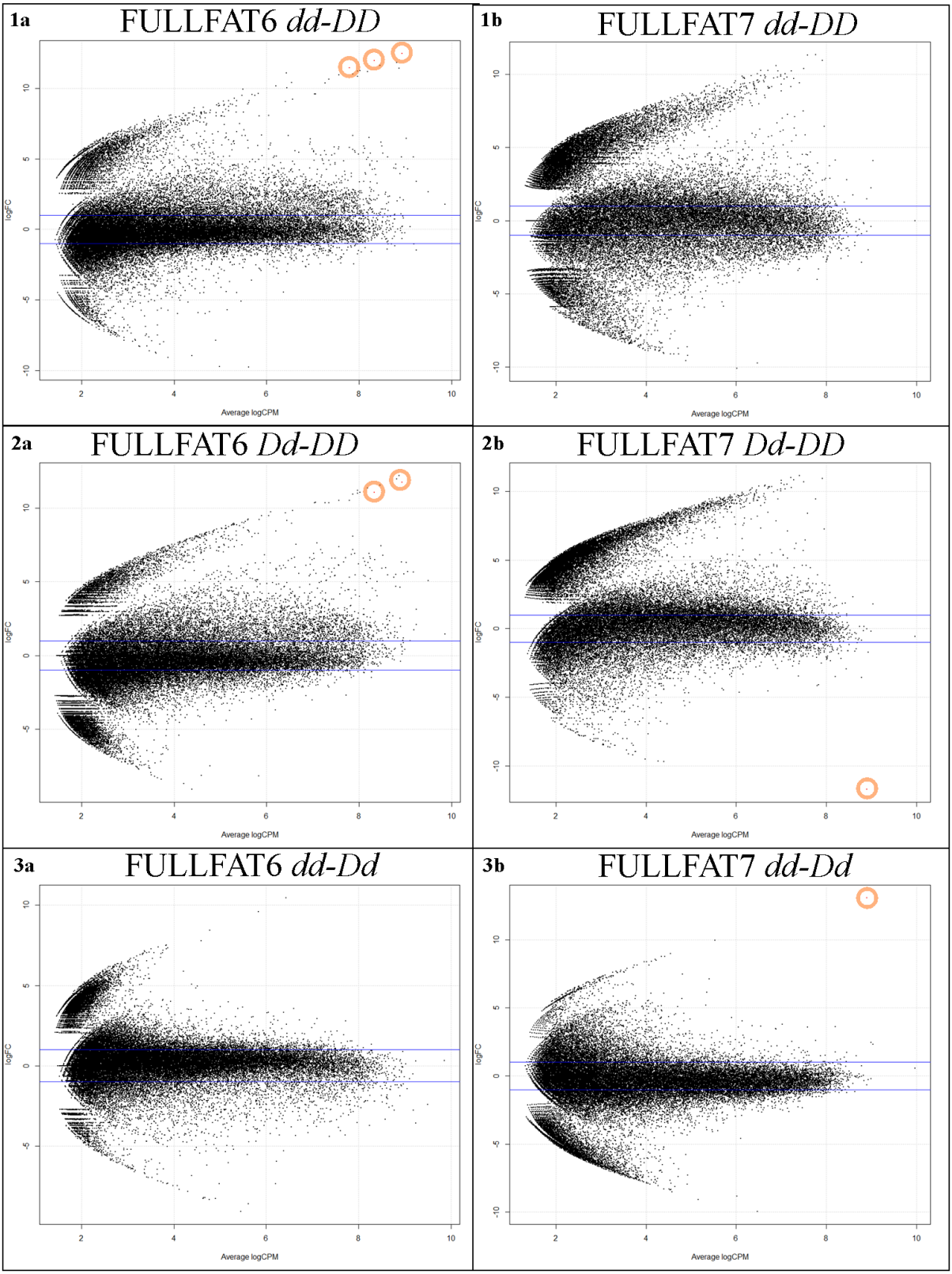


Figure S36 Log fold change in eRAD tag sequence counts between genotypes DD & dd (1); DD & Dd (2) and Dd & dd (3) in datasets FULLFAT6 (a) and FULLFAT7 (b).The direction of relative expression is indicated in the title of each plot. Each data point (shown in black) represents an eRAD tag. Significantly differentially expressed tags are shown in red and emphasized by a circle. The blue lines indicate a two-fold difference in expression.

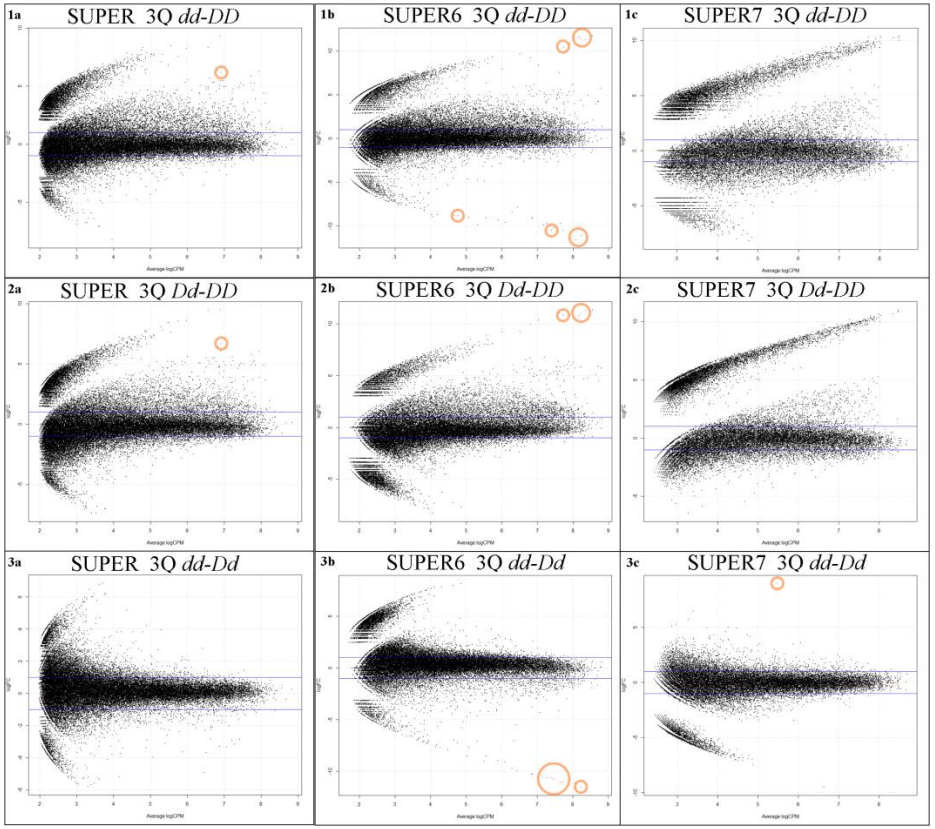


Figure S37 Log fold change in eRAD tag sequence counts between genotypes DD & dd (1); DD & Dd (2) and Dd & dd (3) in datasets SUPER (a), SUPER6 (b) and SUPER7 (c).The direction of relative expression is indicated in the title of each plot. Each data point (shown in black) represents an eRAD tag. Significantly differentially expressed tags are shown in red and emphasized by a circle. The blue lines indicate a two-fold difference in expression. The pairwise comparisons to SUPER7_3Q DD group were not statistically valid however are presented to depict presence/absence relationship.

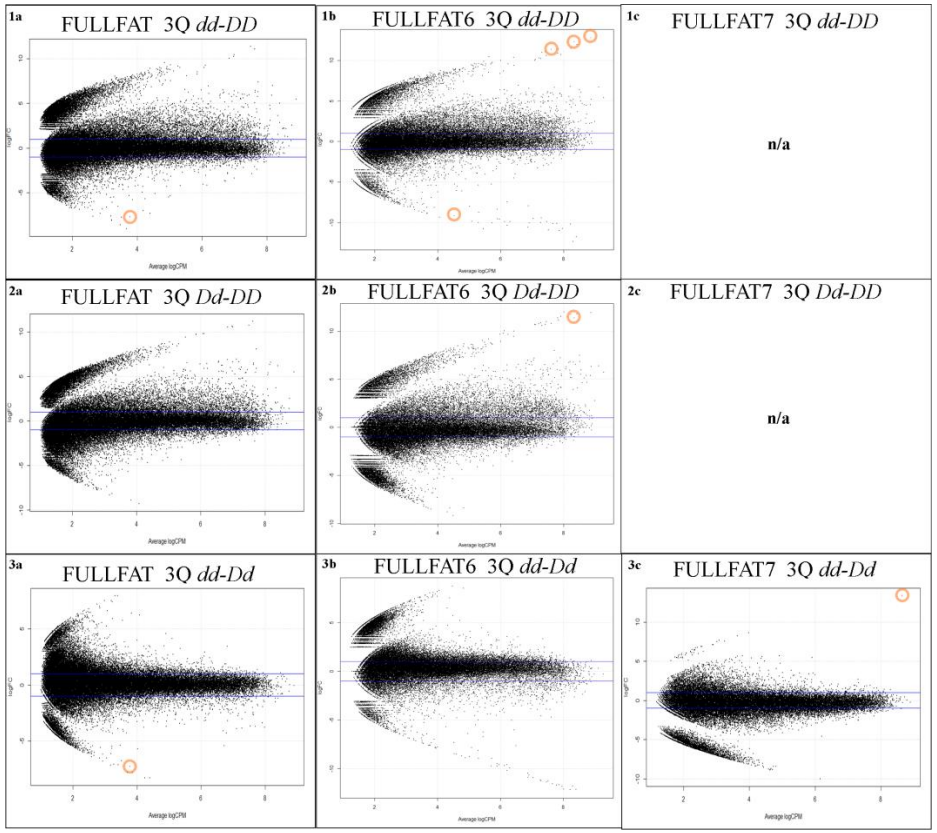


Figure S38 Log fold change in eRAD tag sequence counts between genotypes DD & dd (1); DD & Dd (2) and Dd & dd (3) in datasets FULLFAT (a), FULLFAT 6 (b) and FULLFAT 7 (c).The direction of relative expression is indicated in the title of each plot. Each data point (shown in black) represents an eRAD tag. Significantly differentially expressed tags are shown in red and emphasized by a circle. The blue lines indicate a two-fold difference in expression. The pairwise comparisons to FULLFAT7_3Q DD group were not statistically valid and subsequently not presented (n/a).

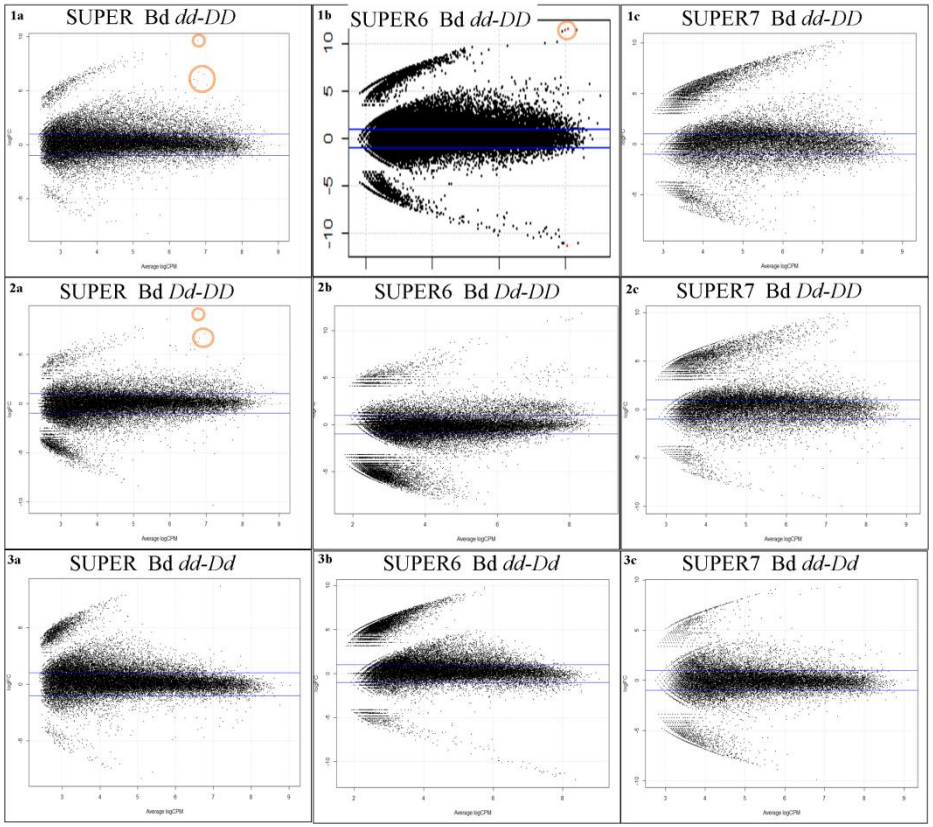


Figure S39 Log fold change in eRAD tag sequence counts between genotypes DD & dd (1); DD & Dd (2) and Dd & dd (3) in datasets SUPER_Bd (a), SUPER6_Bd (b) and SUPER7_Bd (c).The direction of relative expression is indicated in the title of each plot. Each data point (shown in black) represents an eRAD tag. Significantly differentially expressed tags are shown in red and emphasized by a circle. The blue lines indicate a two-fold difference in expression.

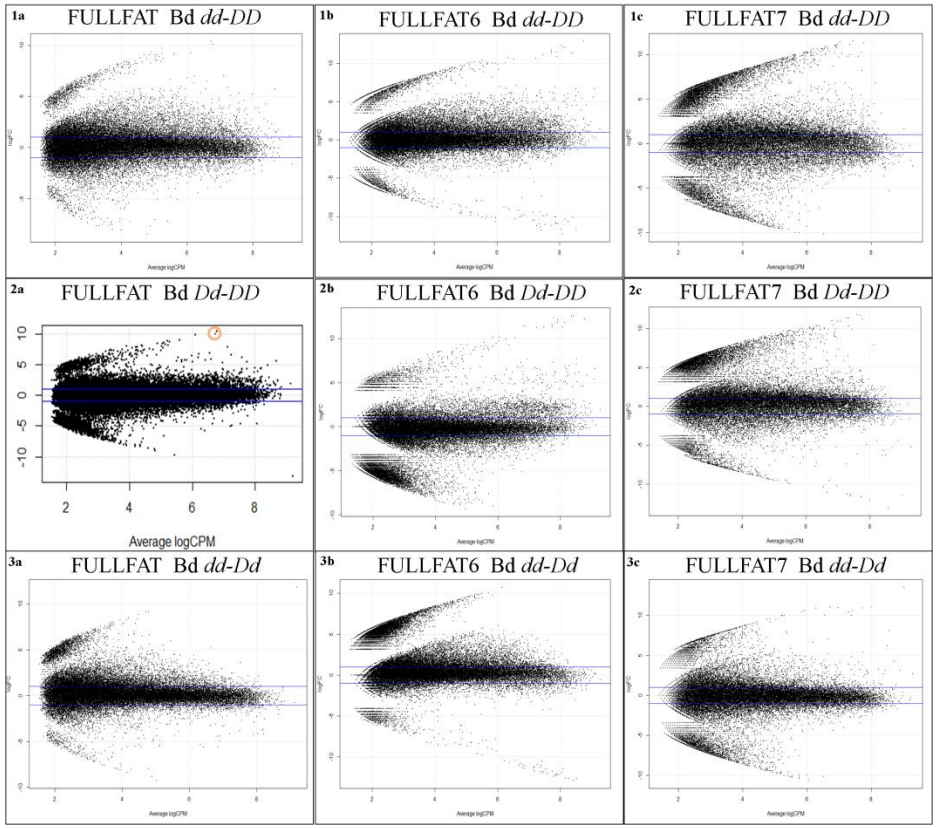


Figure S40 Log fold change in eRAD tag sequence counts between genotypes DD & dd (1); DD & Dd (2) and Dd & dd (3) in datasets FULLFAT_Bd (a), FULLFAT 6_Bd (b) and FULLFAT 7_Bd (c).The direction of relative expression is indicated in the title of each plot. Each data point (shown in black) represents an eRAD tag. Significantly differentially expressed tags are shown in red and emphasized by a circle. The blue lines indicate a two-fold difference in expression.

S17. Blast2GO quantitative values

A summary of the quantitative descriptions of each of the three assemblies generated through Blast2GO is presented in Table S14. One tag was labelled 'without analysis'. Interestingly this same sequence was not analysed in the SUPER and SUPER7 assemblies.

Table S14 Quantitative summary of the Blast2 GO assessed de novo transcriptome assemblies

Assembly	SUPER	SUPER6	SUPER7
Total contigs	35,696	30,438	25,654
without analysis	1	0	1
With Blast no hits	22,169	18,390	15,447
With Blast hits	3,110	2,682	2,275
With Blast2GO Annotation	10,416	9,366	7,931
Average length (all contigs)	313	306	310
Average length (annotated contigs)	352	341	351
Percentage Annoted (%)	29.18	30.77	30.92

References

AppliedBiosystems (2010). "TRI Reagent Solution Protocol."

Barcia, R., J. M. Lopez-Garcia and J. I. Ramos-Martinez (1997). "The 28S fraction of rRNA in molluscs displays electrophoretic behaviour different from that of mammal cells." Biochemistry and Molecular Biology International **42**(6): 1089-1092.

Casey, B. and B. P. Hackett (2000). "Left–right axis malformations in man and mouse." Current opinion in genetics & development **10**(3): 257-261.

Eisenberg, E. and E. Y. Levanon (2013). "Human housekeeping genes, revisited." Trends in Genetics **29**(10): 569-574.

Liu, M. M., J. W. Davey, R. Banerjee, J. Han, F. Yang, A. Aboobaker, M. L. Blaxter and A. Davison (2013). "Fine Mapping of the Pond Snail Left-Right Asymmetry (Chirality) Locus Using RAD-Seq and Fibre-FISH." Plos One **8**(8).

Liu, M. M., J. W. Davey, D. J. Jackson, M. L. Blaxter and A. Davison (2014). "A conserved set of maternal genes? Insights from a molluscan transcriptome." International Journal of Developmental Biology **58**(6-8): 501-511.

Qiagen (2007). "RNeasy micro handbook, 2nd Edition." **Dec. 2007**.

ThermoScientific (2010). "Thermo Scientific NanoDrop Spectrophotometers, Nucleic Acid.": 14.

List of Supplementary Tables

List of Supplementary Figures