Chapter 3

Microarray analysis of germline perturbations

3.1. Introduction

A classical approach to understanding gene function is to generate loss-of-function phenotypes. Such phenotypes, however, require correct characterisation. Most phenotypes in model organisms have previously been reported at the level of morphology, often requiring many different techniques to measure each parameter. My intention was to develop use of expression microarrays as a single phenotyping methodology to compare genic perturbations resulting in brood-size defects in *C. elegans*. In this chapter I present a detailed rationale behind this project and the utility of the approach as a pathway-specific phenotyping tool with potential future application.

RNA-mediated interference (RNAi) has proved a powerful tool for the generation of loss-of-function phenotypes. In particular the capability of perturbing gene function in *C. elegans* simply by the feeding of bacteria expressing dsRNA has led to the generation of an RNAi library consisting of clones targeting ~86% of annotated coding genes (Fraser *et al*., 2000; Kamath *et al*., 2003; Kamath *et al*., 2001). Whole-genome screens using the RNAi library have revealed loss-of-function phenotypes for many genes under laboratory conditions (Kamath *et al*., 2003). For example, hundreds of genes give brood-size defects by RNAi, indicating a deleterious effect on either germline development or gametogenesis. The observation of a sterile animal at low resolution in an RNAi screen, however, tells us almost nothing about gene function as there are many independent pathways and processes which when perturbed lead to germline defects. It is clear, therefore that a high-resolution phenotyping methodology is required. One possibility is through careful microscopic analysis of the worms themselves along with *in situ* and

immuno-stainings to assess the level, location and combinations of expression of certain key genes, which define the biological state of a tissue. This, however, requires prior knowledge of a number of molecular markers and antibodies against them. It would also require the careful dissection of the germline from many animals, drastically limiting throughput.

An alternative approach is to use microarray expression data to define phenotypes. This has been previously demonstrated in *Saccharomyces cerevisiae* to great effect. The expression profile of mutant strains can be considered as 'molecular phenotypes' — they are read-outs of the expression changes that result from a given mutation. These signatures are high density, since they cover all predicted genes, and quantitative, allowing more criteria to be tested than through staining. In *S. cerevisiae* this allowed genes to be clustered into related functional groupings according to similarities in the expression profiles, even for perturbations that were otherwise sub-phenotypic (Hughes *et al*., 2000). For example, mutations in genes involved in mating yield similar signatures, whereas mutations in genes involved in mitochondrial respiration clustered in a separate cluster. By building a compendium of expression signatures of mutations in genes of known pathways it was then possible to place novel genes into pathways by comparing their signatures with the compendium – for example if a novel gene has a signature that resembles that of the sterol biosynthesis pathway, it suggests that it plays a role in this pathway. This was groundbreaking work by Hughes *et al*. and provided the inspiration for our own study. Whilst yeast and human share many key aspects of eukaryotic life, however, as a single-celled organism yeast is of little use in the study of

cellular signalling and development. An approach such as this would therefore be more relevant to human biology if it were performed in a metazoan. Consequently I set out to validate a similar approach in the nematode *C. elegans*.

As previously discussed, the *C. elegans* germline is a well-studied, largely syncytial tissue with a number of genes and pathways known to control certain processes, for example, the Notch pathway is known to regulate the maintenance of the mitotic stem cell niche, the *gld* genes are known to be involved in the mitosis-meiosis switch and gametogenesis, and Ras/MAPK signalling controls exit from the pachytene stage of meiosis. Broadly the germline goes through two distinct phases – firstly it develops into the complete tissue capable of generating differentiated gametes; secondly it is then continually maintained such that the loss of nuclei to gametogenesis is balanced by proliferation of mitotic nuclei.

Historically, due to the complexities of isolating individual tissues or their RNAs the majority of microarray studies in *C. elegans* have been at the level of the whole animal. Gene expression in any individual tissue has therefore proven difficult to establish. Comparisons of different well-characterised loss-of-function mutants, however, have allowed tissue-specific gene expression to be assessed in the germline. This was aided by the facts that the germline accounts for around half the mass of the adult worm, the great majority of transcripts in the adult, and the expression of ~25% of genes is enriched in this tissue. Consequently changes in gene expression in the germline can be assessed at the level of the whole animal (Jiang *et al*., 2001; Reinke *et al*., 2004; Reinke *et al*., 2000).

For this reason the worm germline is an attractive tissue as the focus of our study. The published expression studies also provide us with an ideal dataset against which to compare our data.

3.2. Outline of Approach

As well as there being many well-studied mutant strains exhibiting brood-size defects, the existence of the *C. elegans* RNAi library permits the generation of loss-of-function animals for almost any gene in the genome. For genes of known function and loss-of-function phenotype, whilst the loss-of-function phenotypes generated by RNAi when visually observed at low resolution do not appear to be as strong as null mutant phenotypes, they nevertheless demonstrate some measure of brood-size defect, as would be expected based on prior knowledge. We therefore have the ability to generate loss-of-function phenotypes for most genes with established roles in germline development.

The stage in germline development at which a defect occurs dictates the extent of development and the mitotic/meiotic character of the germline. I decided to consider four different categories of perturbation in our initial compendium before making comparisons with novel genes. This includes expression profiles of perturbations of genes known to control the three aspects of germline development previously mentioned – maintenance of the mitotic stem cell niche, regulation of the mitosis-meiosis switch, and release from the pachytene stage of meiosis. Thus far, however, all of the genes considered are involved in signalling, transcription and regulation of individual transcripts. Furthermore they appear to have discreet roles in the biology of the animal. In order to provide a contrast to this I chose to perturb components of the basal cellular machinery to see if

they appear distinctly different by array profile. The majority of ribosomal components give completely sterile phenotypes by RNAi. RNAi knockdown of these genes may be expected to give comparable functional defects, reflected in the corresponding microarray expression profiles. Ribosomal knockdowns were therefore added to the study in order to determine whether specific clustering can be achieved and whether the clustering is pathway or strength specific.

To be more clear, the expectations of this study are that the phenotypes of animals deficient for a single component of a signalling pathway will be more similar to that of animals deficient in the same pathway than in another. By using microarrays to generate high-density loss-of-function phenotypes for components of numerous pathways involved in germline development followed by hierarchical clustering, we would expect to rediscover the known pathways as independent branches of the clustering. Novel genes of interest could then be tested against the resulting compendium to provide evidence of their role in a given pathway.

RNA extracted from young adults was used for all experiments in this study. The germline is fully developed by this stage and all of the genes mutated or knocked down in these experiments act before and during the young adult stage.

The two established methods of gene perturbation that could be used in this study are mutation and RNAi. As a long established organism for forward genetics many mutagenesis screens have been performed using ethyl methane sulphonate- (EMS-) or N-

ethyl-N-nitrosourea-(ENU-) induced mutagenesis followed by genetic screening. This has led to a large collection of genetic mutants, which are available to the global *C. elegans* community from the *C. elegans* Genetics Center, USA (http://www.cbs.umn.edu/CGC/). One potential drawback of using such mutants is the possibility of there being some background mutations caused by the mutagenesis, which may not have been removed by out-crossing. Although there are many genetic mutants available there are still many genes pertinent this study for which no genetic mutant is available. RNAi offers an alternative method of genic perturbation, and the RNAi library contains clones allowing the knowdown of the majority of individual coding genes. This therefore necessitates the use of RNAi in this study. RNAi, however, is likely to give less complete perturbation of gene function. I therefore decided to compare RNAi with genetic mutants where possible. The differing level of RNAi knockdown per gene results in a range of brood-size defects. It is also known that there can be a high level of animal-to-animal phenotypic variability on RNAi. The questions that need to be addressed in order to establish the utility of this approach are therefore:

1. Can we rediscover known pathways based on expression profiles (i.e. do different perturbations of the EGF pathway cluster together; do different perturbations of the Notch pathway cluster together and independently of the EGF pathway)?
2. Does RNAi phenocopy mutation (both physiologically and molecularly)?
3. How dependent is molecular phenotype on the strength of the visual phenotype (does strength of phenotype or the pathway that the gene acts in drive clustering)?
4. How dependent is molecular phenotype on the penetrance of a perturbation?

In order to answer these questions, for each gene perturbed I used microarrays to expression profile a population of ~10,000 animals in biological triplicate, DAPI stained whole adult animals to broadly assess the quantity of germline present and assessed the fecundity of 12 individual animals by visual phenotyping. Where multiple RNAi clones existed against a gene of interest in the RNAi library they were each used individually in order to compare different strengths of RNAi against the same gene. Each clone may give different levels of observed sterility owing to the fact that they give rise to a different set of siRNAs, giving different efficiencies and levels of transcript knockdown. The genic perturbations (genetic mutants and RNAi) used for this set of experiments are shown in table 3.1. Note that whilst *sem-5* is not confirmed to be required for progression beyond pachytene, it is upstream of *sos-1* in the canonical EGF/ras/MAPK signalling cascade and gives a brood-size defect by RNAi. Consequently it was included in the first round of experiments.

NOTCH PATHWAY	RIBOSOME	RAS/MAPK SIGNALLING	MITOSIS/MEIOSIS SWITCH AND GAMETOGENESIS
glp-1 (or178)	*rps-1 (RNAi)*	*sos-1 (cs41)*	*gld-1 (RNAi)*
lag-2 (q420)	*rps-14 (RNAi)*	*sos-1 (RNAi)* x3	*gld-2 (RNAi)* x3
emb-5 (hc61)	*rpl-20 (RNAi)*	*sem-5 (RNAi)*	
glp-1 (RNAi)	*rpl-21 (RNAi)*	*let-60 (RNAi)*	
lag-2 (RNAi) x2		*mpk-1 (RNAi)*	
emb-5 (RNAi)		*mek-2 (RNAi)*	
lin-12 (RNAi)		*lin-45 (RNAi)*	
lag-1 (RNAi)			

Table 3.1. Genes involved in germline development perturbed in this study. The nature of the perturbation is indicated in parentheses. The column headings indicate pathway or machinery categories into which the below genes fall.

The microarrays chosen for this study were two-colour synthetic oligonucleotide arrays acquired from Washington University in St. Louis, MO, USA. The microarray contains 22,490 70mer genic probes. Detailed specifications can be found here: http://genome.wustl.edu/genome/celegans/microarray/array_spec.cgi. All experimental samples (Cy3) were hybridized against the same mixed-stage reference sample (Cy5). Each perturbation was compared indirectly to wild-type via a mixed-stage reference sample. The wild-type array profile was derived from animals fed on a bacterial strain expressing a non-targeting dsRNA.

It is typical in expression studies using two-colour microarrays that two samples are compared directly by competitive hybridization to the same microarray. Dye swaps are performed in order to correct for the differing efficiencies of incorporation of labelled nucleotides into cDNA by the reverse transcriptase and the different quantum-yields of the two dyes. "Dye swaps" refers to performing a repeat hybridization of the same RNA samples with the fluorescent labels switched. This approach doubles the number of hybridizations that need to be performed which can be financially prohibitive. Comparison of experimental samples via a universal reference sample negates the need for dye swap hybridizations as the experimental sample is always labelled with the same dye. The key requirement of the mixed stage reference sample is that it provides signal above background for the vast majority of spots on the array such that the corresponding genes are included in the analysis. Comparison between any two conditions on different arrays can then easily be inferred via the reference sample as:

(condition A signal/reference signal) ÷ (condition B signal/reference signal) = condition A signal/ condition B signal.

3.3. Initial microarray data processing, normalisation and assessment of data quality

Since the key manner in which two samples are compared on a two-colour microarray is by the measured ratio of signal present per spot, it is necessary that the signal for both samples is sufficiently higher than the measured background such that the ratios can be considered reliable. For this reason low quality spots are filtered out prior to normalization. Further to this, complex experimental platforms such as microarrays are highly prone to experimental and systematic variation, which must be corrected for before accurate measures of expression changes can be drawn between arrays. An example of this is an imbalance of the two dyes on the array, which may result from the laser settings when scanning the array (experimental) but also the position of the spot on the array (systematic). The term "normalization" therefore refers to the correction for experimental and not biological variation between experiments.

There is no general consensus in the scientific community regarding the best method of data normalization. Multiple methods were therefore tested, each a variation on the well-established loess normalization (Yang *et al*., 2002). This can be done in a global way - normalizing all spots together, or in a block-wise way by dividing each microarray into "sub-arrays" and normalizing within the sub-arrays. Global and block-wise loess normalization, both with and without background subtraction was performed using DNMAD (Tarraga *et al*., 2008). Pearson correlation of normalized biological triplicates

was performed. This was to determine the degree of biological and technical reproducibility of experiments. The correlation between each of three independent replicates may allow the identification of an outlying sample, which should be removed. The difference in Pearson correlation between the same samples for different normalization techniques may also indicate which method best corrects for technical variation. Pearson correlation was improved by filtering out spots giving median intensities <150 in either detection channel. The rationale behind this is that lower intensity spots have a higher percentage error in detection, leading to more variability between replicates. This will, however, lead to the loss of good spots and the spots discarded will be different depending on the quality of array and the gain of the lasers on scanning.

Multiple technical replicates were performed of the wild-type sample against the reference sample and the robustness of the system was assessed by the Pearson correlation. This was found to be consistently 0.93-0.96. Assessment of correlations allowed us to compare the performance of normalization methods. All four of the above methods of normalization performed comparably for good arrays. Global loess performed less well for arrays that exhibited marked positional effects, such as the loss of dye intensity near the periphery of arrays.

An alternative normalization method based on a sliding square window surrounding each spot was also tested (Lyne *et al*., 2003). This method outperformed the others, as it uses smaller windows for normalization around the periphery of the array, allowing it to better

account for positional effects. This method also offers an alternative method of filtering out lower quality spots. Spots with < 50% of pixels > 2 SD above median local background signal in one or both channels are flagged absent, unless one channel showed > 95% of pixels > 2 SD above local background. Removal of spots is therefore more consistent and in-line with the quality of the individual arrays. It also retains spots that are highly expressed in one channel and therefore less susceptible to skewing. The script uses only the lower 55% of pixel intensities as this reduces the likelihood of skewing by bright pixels. This script is also more versatile, allowing the default settings to be altered in a graphical user interface. Alternatively large quantities of arrays can be processed at default settings using the command line. This script therefore not only reduces loss of good spots, but is also favourable should we set up a database for automated microarray analysis.

Table 3.2 shows the Pearson correlations between replicates for all arrays for which data is presented in this chapter. It demonstrates that removal of low intensity spots followed by normalization with DNMAD performs well for good quality arrays. The Lyne *et al*. method broadly performs less well for the same good quality arrays but better for the arrays that gave poor correlations using the previous method. The average correlation across all arrays with both methods is identical. The data for each replicate it therefore more likely to be consistent using the Lyne *et al*. normalization script. Critically, the Lyne *et al*. method of filtering poor quality spots permits on average 50% more genes to be considered. The Lyne *et al*. normalization method was therefore chosen for future use.

Arrays compared		Lyne *et al.*	Flagging spots <150 and DNMAD	Arrays compared		Lyne *et al.*	Flagging spots <150 and DNMAD
N2 control 1	N2 control 2	0.94	0.96	*lin-12* 1	*lin-12* 2	0.90	0.95
N2 control 1	N2 control 3	0.91	0.97	*lin-12* 1	*lin-12* 3	0.92	0.95
N2 control 2	N2 control 3	0.94	0.96	*lin-12* 2	*lin-12* 3	0.93	0.95
emb-5 1	*emb-5* 2	0.90	0.92	*lin-3* 1	*lin-3* 2	0.92	0.93
emb-5 1	*emb-5 3*	0.92	0.92	*lin-45* 1	*lin-45* 2	0.86	0.83
emb-5 2	*emb-5* 3	0.89	0.90	*lin-45* 1	*lin-45* 3	0.82	0.83
*emb-5** 1	*emb-5** 2	0.92	0.95	*lin-45* 2	*lin-45* 3	0.88	0.92
*emb-5** 1	*emb-5** 3	0.87	0.93	*mek-1* 1	*mek-1* 2	0.96	0.96
*emb-5** 2	*emb-5** 3	0.92	0.94	*mek-1* 1	*mek-1* 3	0.87	0.79
gld-1 1	*gld-1* 2	0.90	0.91	*mek-1* 2	*mek-1* 3	0.87	0.78
gld-1 1	*gld-1* 3	0.92	0.95	*mpk-1* 1	*mpk-1* 2	0.88	0.89
gld-1 2	*gld-1* 3	0.86	0.90	*mpk-1* 1	*mpk-1* 3	0.84	0.84
gld-2 a 1	*gld-2* a 2	0.90	0.92	*mpk-1* 2	*mpk-1* 3	0.82	0.80
gld-2 a 1	*gld-2* a 3	0.86	0.90	*rpl-20* 1	*rpl-20* 2	0.83	0.88
gld-2 a 2	*gld-2* a 3	0.92	0.90	*rpl-20* 1	*rpl-20* 3	0.84	0.86
gld-2 b 1	*gld-2* b 2	0.91	0.91	*rpl-20* 2	*rpl-20* 3	0.93	0.90
gld-2 b 1	*gld-2* b 3	0.86	0.92	*rpl-21* 1	*rpl-21* 2	0.88	0.88
gld-2 b 2	*gld-2* b 3	0.89	0.86	*rpl-21* 1	*rpl-21* 3	0.84	0.90
gld-2 c 1	*gld-2* c 2	0.92	0.92	*rpl-21* 2	*rpl-21* 3	0.90	0.90
gld-2 c 1	*gld-2* c 3	0.89	0.87	*rps-1* 1	*rps-1* 2	0.86	0.71
gld-2 c 2	*gld-2* c 3	0.83	0.81	*rps-1* 1	*rps-1* 3	0.79	0.68
glp-1 1	*glp-1* 2	0.83	0.79	*rps-1* 2	*rps-1* 3	0.89	0.93
glp-1 1	*glp-1* 3	0.83	0.86	*rps-14* 1	*rps-14* 2	0.91	0.89
glp-1 2	*glp-1* 3	0.93	0.83	*rps-14* 1	*rps-14* 3	0.94	0.92
*glp-1** 1	*glp-1** 2	0.96	0.95	*rps-14* 2	*rps-14* 3	0.92	0.96
*glp-1** 1	*glp-1** 3	0.91	0.92	*sem-5* 1	*sem-5* 2	0.82	0.90
*glp-1** 2	*glp-1** 3	0.92	0.94	*sem-5* 1	*sem-5* 3	0.92	0.93
lag-1 1	*lag-1* 2	0.87	0.86	*sem-5* 2	*sem-5* 3	0.84	0.89
lag-1 1	*lag-1* 3	0.90	0.86	*sos-1* a	*sos-1* a 2	0.77	0.93
lag-1 2	*lag-1* 3	0.94	0.92	*sos-1* a	*sos-1* a 3	0.92	0.96
la g-2 a 1	*la g-2* a 2	0.89	0.93	*sos-1* a	*sos-1* a 3	0.90	0.95
la g-2 a 1	*la g-2* a 3	0.87	0.86	*sos-1* b	*sos-1* b 2	0.84	0.85
la g-2 a 2	*la g-2* a 3	0.91	0.86	*sos-1* c	*sos-1* c 2	0.92	0.86
la g-2 b 1	*la g-2* b 2	0.88	0.89	*sos-1* c	*sos-1* c 3	0.90	0.96
la g-2 b 1	*la g-2* b 3	0.88	0.89	*sos-1* c	*sos-1* c 3	0.87	0.87
la g-2 b 2	*la g-2* b 3	0.89	0.87		Average	0.89	0.89
*lag-2** 1	*lag-2** 2	0.83	0.79	No. spots considered post-filtering		14404.67	9582.69
let-60 1	*let-60* 2	0.83	0.85				
let-60 1	*let-60* 3	0.88	0.89				
let-60 2	*let-60* 3	0.87	0.94				

Table 3.2. Relative Pearson correlations using different normalization methods. Correlations markedly improved by the Lyne *et al.*, method highlighted in yellow. Low quality arrays that were removed from the analysis are indicated in red. Replicate number is indicated after gene name. Letters between gene name and replicate number indicate use of different RNAi clones. An * indicates a genetic mutant rather than RNAi.

The mixed-stage reference sample against which all experimental samples were hybridized was derived from vast quantities of synchronous animals grown in liquid culture. The RNA extracted from the individual cultures before mixing providing us with known quantities of RNA derived from each developmental stage. A key property of the reference sample is that it must represent the vast majority of annotated genes such that the minimum number of spots will be filtered out prior to normalization. For any given microarray > 85% of spots that are filtered as low quality are filtered due to low signal for both dyes. Across all experiments, of the 22,490 genic spots on the array > 20,300 are represented post-normalization by the filtering criteria used. We therefore consider the mixed-stage reference sample to be of suitable quality for the study.

We have idealized our methodology for producing expression data for any given biological condition. We have determined that the materials that we are producing for microarray analysis are adequately consistent and our initial data processing is robust and practical. We will next determine the differential regulation of genes between the conditions for which data have been generated. This gives us a basis for comparison of the different genic perturbations.

3.4. Proof-of-principle experiments

As is clear from table 3.1, I examined the effect of RNAi knockdown for multiple components of different pathways. Where appropriate mutants were available I sought to compare the effects of perturbation by RNAi and mutation. I also used multiple RNAi clones to target certain genes in order to compare the effects of different strengths of

RNAi against the same gene. Further to this, in order to validate our microarray data I sought to compare it with relevant data produced by other labs.

For each biological condition expression-profiled, differentially expressed genes were identified using Student's t-test. All comparisons were to the reference strain N2 fed bacteria expressing non-targeting dsRNA. This provided us with filtered data for each condition, a means of testing how well RNAi phenocopies mutation and a means of benchmarking our data against published data. The number of genes differentially expressed between the wild-type control and each perturbation is shown in table 3.3.

To check that our methods give similar data to other groups I used the comparison of *glp-1(or178)* with reference strain Bristol N2 (wild-type control), which is analogous to the comparison of *glp-4(bn2)* to N2 by Reinke *et al*. (2004). Both mutants lack a germline, however, the molecular identity of *glp-4* is unknown. Genes more highly expressed in N2 relative to either *glp-4(bn2)* or *glp-1(or178)* can be considered to be germline enriched/intrinsic. Reinke *et al*. define 3143 genes thus using Student's t-test (p-value $\leq$ 0.01). We discover 4831 genes by the same method, encompassing 65% of the Reinke set. We consider this to be a very good overlap, given that this is a cross-platform comparison of a 20K PCR product array (Reinke *et al*.) versus our 22.5K synthetic oligo array. Furthermore the inevitable difference in precise timing at which RNA was harvested between the two labs and the fact that the Reinke *et al.* data is derived from worms fed on *Escherichia coli* strain OP50 and ours from animals fed on HT115(DE3) may further explain the discrepancies.

Gene Perturbed	Genes higher than in N2	Genes lower than in N2
emb-5	1258	1232
*emb-5**	3659	6322
glp-1	1846	2005
*glp-1**	3989	6898
lag-1	654	1757
lag-2 a	2123	2607
lag-2 b	2209	1953
*lag-2**	2614	4076
lin-12	1274	1606
gld-1	2212	1243
gld-2 a	1651	2557
gld-2 b	1951	993
gld-2 c	1629	1713
let-60	1571	1465
lin-45	1496	1155
mek-2	818	857
mpk-1	527	771
sem-5	1404	675
sos-1 a	3262	1811
sos-1 b	1798	516
sos-1 c	1723	2327
pkc-1 a	1031	960
pkc-1 b	1888	2602
rpl-20	2649	3111
rpl-21	1671	2159
rps-1	2408	2808
rps-14	1788	1107

Table 3.3. Genes upregulated and downregulated relative to N2 for each condition. The table shows the number of Genes upregulated and downregulated relative to N2 for each genic perturbation, as determined by Students t-test (p-value <0.05). An asterisk indicates a genetic mutant rather than RNAi. A letter after the gene name indicates use of different individual clones used for RNAi knockdown.

Each condition was compared by hierarchical clustering of calculated ratios of perturbation/wild-type control for each gene differentially expressed between the two conditions (p-value ≤0.05), as can be seen in figure 3.1. It is immediately apparent from the clustering achieved that we recapitulate the known biology, with the components of

the Notch, Ras/MAPK and ribosome gene categories each populating their own separate branch of the condition tree. The components of the mitosis-meiosis switch machinery do not form such a clear niche in the clustering however. This is not completely surprising as the complexities of the dual functions of this machinery means different strengths of perturbation are less likely to consistently generate physiologically analogous animals. Furthermore the consideration of only two genes (albeit one of them appearing three times) may not be adequate to resolve the pathway.

The hierarchical clustering of array profiles is based on a correlation matrix of the differentially expressed genes within all conditions being compared. The standard correlation between all conditions is calculated and each condition arranged in a clustering based on the relative relationship of each condition. This is also performed for each individual expressed gene, leading to a 2-dimensional clustering. For the majority of this chapter I will only discuss one dimension – the clustering achieved between conditions in order to determine the relatedness of perturbations.

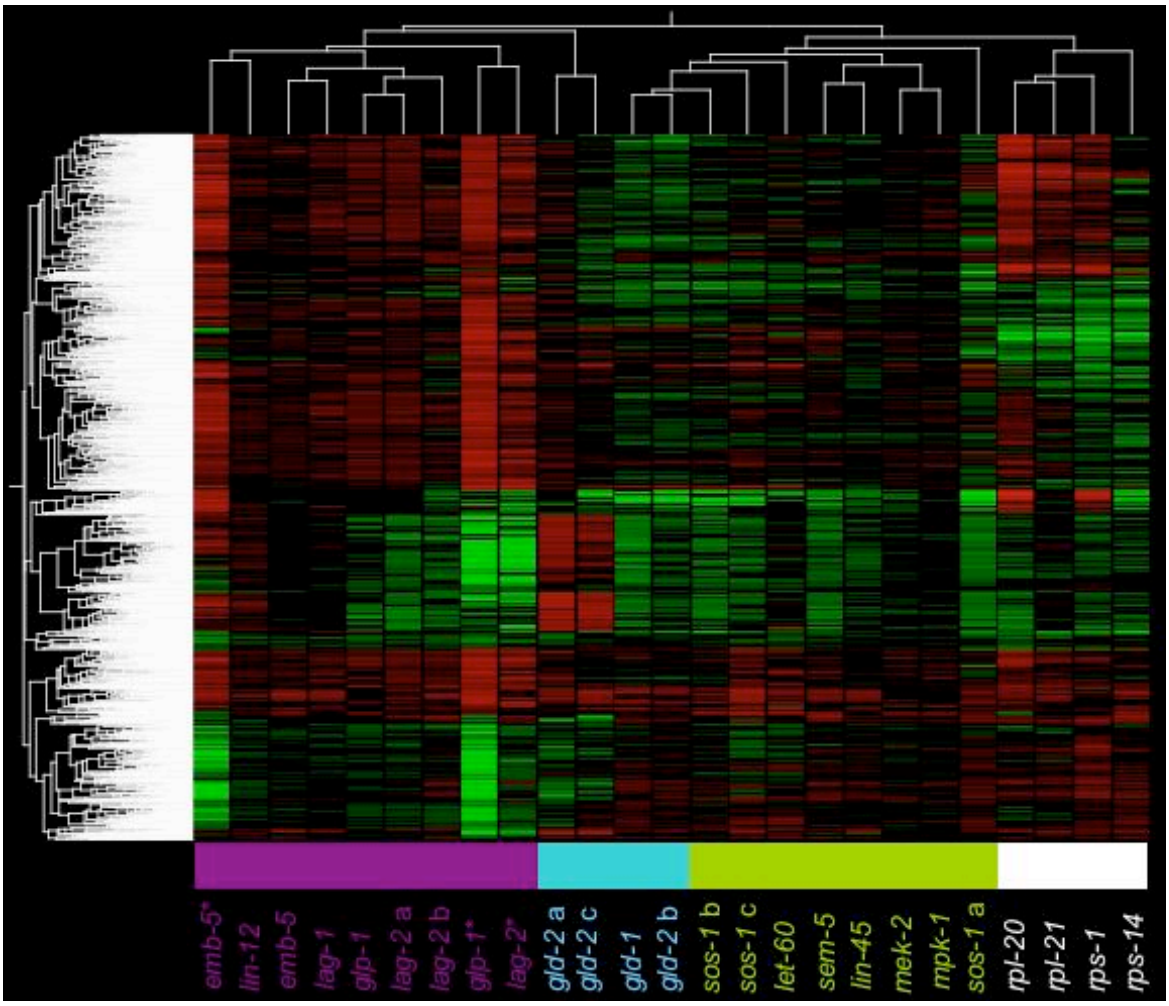

Figure 3.1. Clustering of differentially expressed genes between N2 and each genic perturbation. Calculated ratios of gene signal (perturbation/wild-type) for differentially expressed genes (Student's t-test p-value ≤0.05) were hierarchically clustered. The different pathways and machineries are colour-coded: purple – Notch; blue – mitosis-meiosis switch; green – EGF/ras/MAPK signalling; white – ribosome. Lowercase letters following gene name indicates the use of different individual RNAi clones targeting the same gene. An asterisk indicates a genetic mutant rather than RNAi knockdown. Genes upregulated in each condition relative to wild-type are represented in green and downregulated in red. The intensity of colour is analogous to the magnitude of regulation.

3.5. Low-resolution phenotypic analysis of pathway perturbations

It is necessary to establish that the clustering achieved is not simply indicative of strength of perturbation. Of all the genes discussed as having roles in germline development in chapter 1, the genes in table 3.1 are known to give brood-size defects by RNAi. In parallel with the production of each RNA sample the fecundity of 12 animals was measured relative to wild-type for each RNAi perturbation and mutant (figure 3.2). All of the mutants used in this study are temperature sensitive, having a relatively normal brood size at the permissive temperature and being 100% sterile and lacking a germline at the restrictive temperature. The variability in severity and penetrence of phenotype within pathways for the perturbations shown in figure 3.2 suggest that if pathways can be accurately rediscovered using these array profiles, then it is possible to cluster genes giving mild and variable perturbations into pathways. Figure 3.1 demonstrates that the strength of sterility is not driving the clustering as pathways are reliably rediscovered despite Notch and EGF perturbations giving overlapping ranges of sterility.

Animals representing all perturbations shown in table 3.1 have been DAPI stained and the germline imaged (figure 3.3). We find that whilst *glp-1(or178)* and *glp-1(RNAi)* cluster very closely and appear entirely distinct from *mpk-1(RNAi)* by array profile (as one would predict), by this method of staining they appear distinctly different. At up to 400x magnification all Notch mutants clearly have no germline. Notch perturbations by RNAi, however, are indistinguishable from the other perturbations studied at this magnification, even though their sterility ranges up to ~95%. This is understandable as the Notch mutants studied are temperature sensitive and having been grown from L1 at

the restrictive temperature are expected to almost completely negate gene function whereas RNAi has a cumulative effect over time and is unlikely to give 100% knock-down. This suggests that we may not have been able to recapitulate pathways by comparison of mutants and RNAi by staining alone. We have, however, already demonstrated that RNAi can reliably phenocopy mutation on a molecular level.

In conclusion, the clustering achieved appears to be pathway specific even though the extent and variability of brood-size defects overlaps between pathways for different genic perturbations. Whilst the quantity of germline present in genetic mutants and the equivalent RNAi animals can appear markedly different, on a molecular level the animals appear comparable. We therefore consider the methodology to be validated and ready for comparison with selected candidate genes.

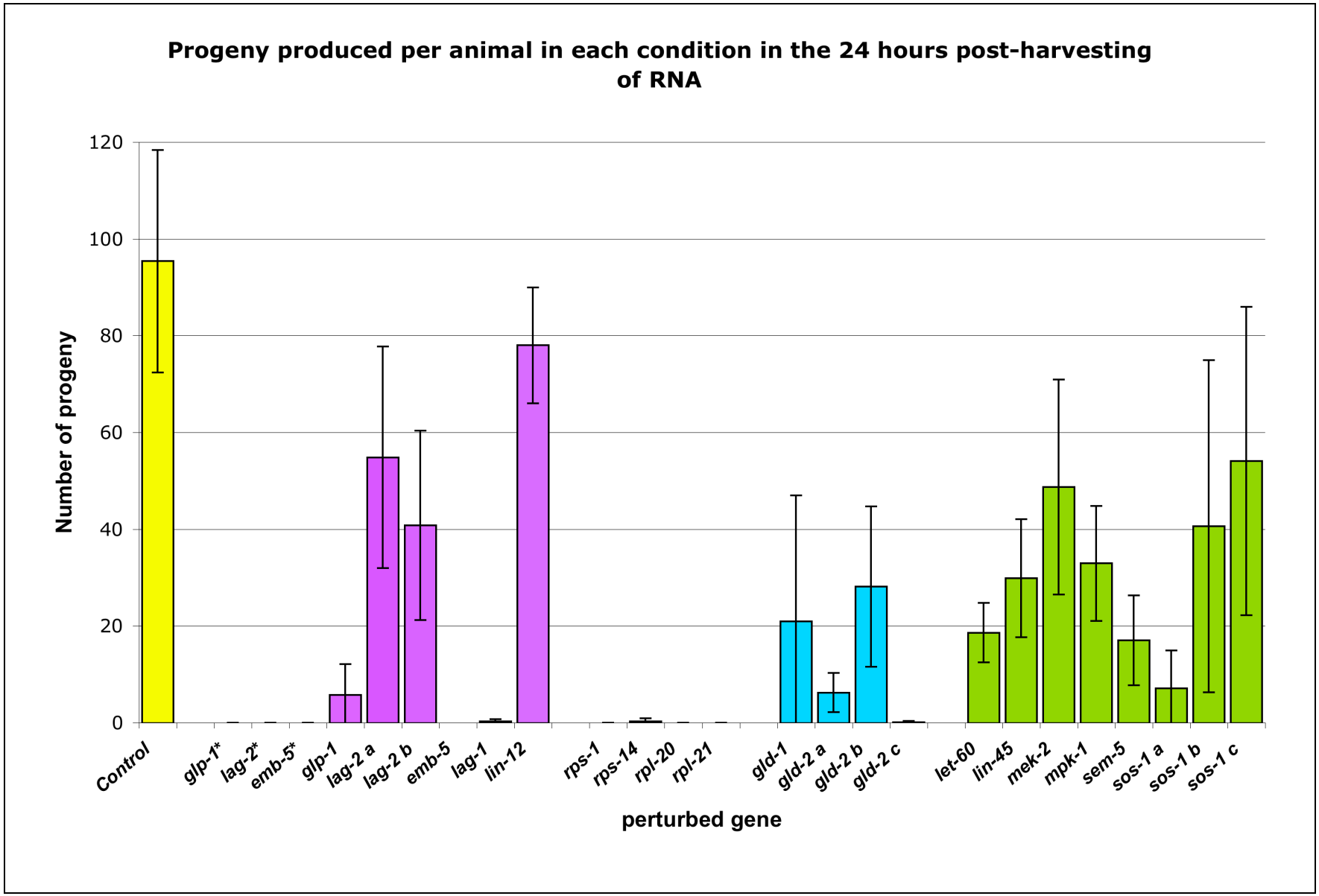
Progeny produced per animal in each condition in the 24 hours post-harvesting of RNA
Number of progeny
120
100
80
60
40
20
0
Control
glp-1*
lag-2*
emb-5*
glp-1
lag-2 a
lag-2 b
emb-5
lag-1
lin-12
rps-1
rps-14
rpl-20
rpl-21
gld-1
gld-2 a
gld-2 b
gld-2 c
let-60
lin-45
mek-2
mpk-1
sem-5
sos-1 a
sos-1 b
sos-1 c
perturbed gene

Figure 3.2. Relative fecundity of germline perturbations. The brood size in the 24 hours after RNA harvesting was assessed for 12 individual animals (3 from each replicate). The graph indicates the number of progeny for each RNAi perturbation, mutant and the wild-type control. Genes are separated and colour-coded according to pathway. Lowercase letters following gene name indicates the use of different individual RNAi clones targeting the same gene. An asterisk indicates a genetic mutant rather than RNAi knockdown.

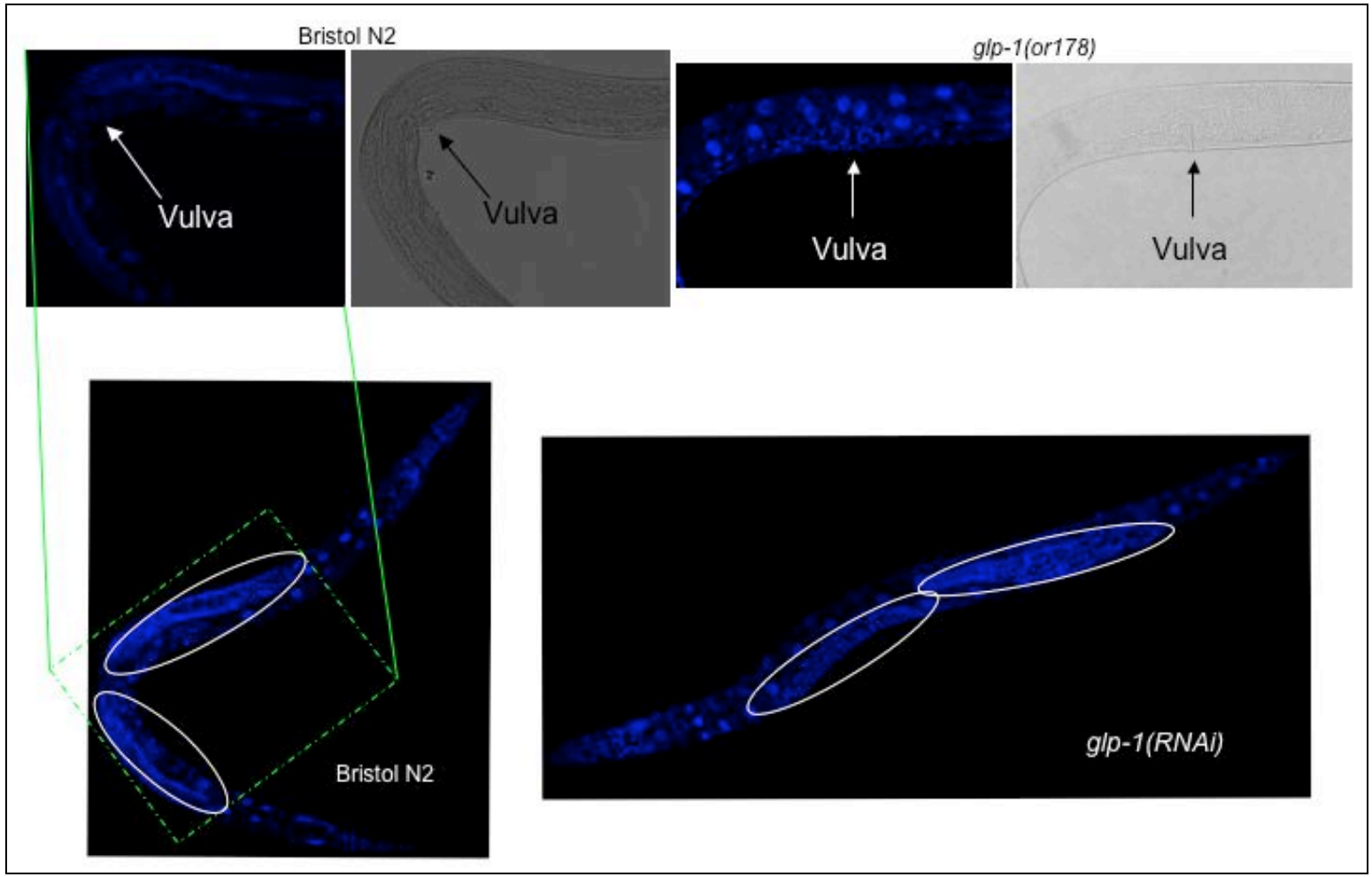

Figure 3.3. DAPI staining of whole animals to assess quantity of germline. This figure shows N2, *glp-1(or178)* and *glp-1(RNAi)* animals as labelled. It is clear that N2 and *glp-1(RNAi)* animals have two clear gonad arms (circled) stretching roughly equidistantly in both directions from the vulva. Higher magnification of this central portion of *glp-1(or178)* reveals no germline.

3.6. Identification of novel modulators of Ras/MAPK signalling in the germline

Once the compendium of well-characterized genes was established it was necessary to decide how to proceed. There were two clear options – (a) to add to the compendium perturbations of genes giving sterile animals by RNAi or mutation, but with no known link to any of the signalling pathways considered; (b) to query the compendium with candidate modulators of signalling pathways already represented in the compendium. These candidate modulators may either have been discovered in genetic interaction screens for genes that modulate the sterile phenotype of Notch and EGF/ras/MAPK signalling mutants or genes that modulate the multi-vulval (Muv) phenotype in mutants with activated EGF/ras/MAPK signalling. Both of these options appeared viable. The next step chosen was therefore to test candidate modulators revealed in vulval screens against the compendium for reasons discussed below.

As discussed in chapter 1, the *C. elegans* vulva is an extremely well studied tissue, serving as an exemplary model for how different signalling pathways combine to regulate the correct development of an individual tissue. Briefly, a set of vulval precursor cells (VPCs) exists along the ventral axis of the animal. EGF/ras/MAPK signalling to the correct cell leads to a cascade of events and the development of a single 22-cell vulva in the centre of the ventral axis, providing a breach between the uterus and the outside world (figure 1.4). Other cells with the potential to develop into the vulva exist along the ventral axis but do not receive adequate stimulus in wild-type animals, ensuring that only one vulval protrusion forms. Mutations leading to an increase in EGF/ras/MAPK

signalling, however, lead to the development of pseudo-vulvae along the ventral axis of the worm.

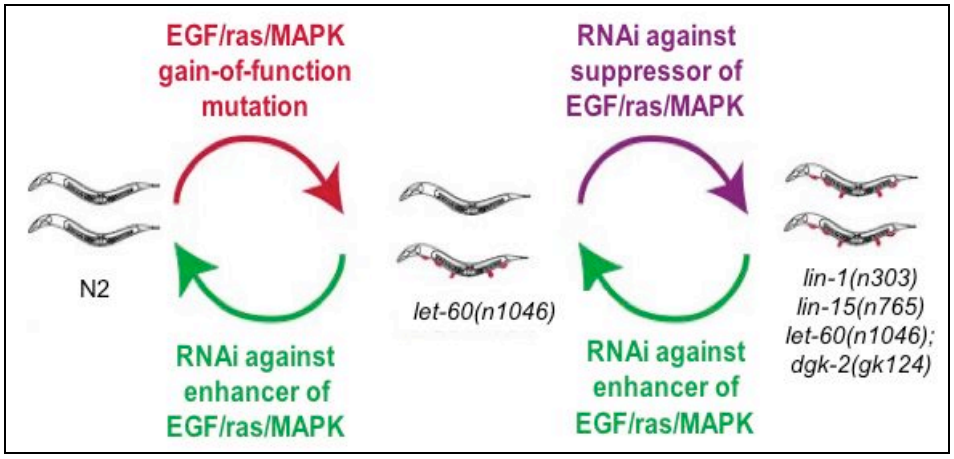

Figure 3.4. Screening for modulators of EGF/ras/MAPK signalling in the vulva. Wild-type animals have a single 22-cell vulva in the centre of their ventral axis. Gain-of-function ras (*let-60*) mutations lead to the formation of pseudo-vulval protrusions (red). RNAi against genes that enhance signalling via ras lead to a decrease in the number of Muv animals i.e. such genes are enhancers of ras signalling. Conversely, RNAi against genes that suppress the consequences of signalling through ras lead to an increase in the number of Muv animals.

In order to identify novel genes that may be involved in EGF/ras/MAPK signalling in *C. elegans*, RNAi screens in mutant animals exhibiting the multi-vulval (Muv) phenotype were performed by Catriona Crombie in the Fraser lab. Specifically, all genes annotated as being signalling (1121), transcription factor (500) or chromatin remodelling (216) genes (Kamath *et al*., 2003) were screened in multiple Muv mutants. Genes that gave a shift in the number of Muv worms by RNAi could be considered candidate modulators of signalling pathways involved in vulval patterning. Genes that when perturbed enhance the Muv phenotype are potential suppressors of EGF/ras/MAPK signalling. Conversely, genes that when perturbed suppress the Muv phenotype are potential enhancers of EGF/ras/MAPK signalling (figure 3.4).

I was specifically interested in genes that are potential enhancers of EGF/ras/MAPK signalling. I therefore selected candidate modulators identified in three different Muv mutants - *lin-1(n303)*, *lin-15(n765)* and *let-60(n1046);dgk-2(gk124)*. As a gain-of-function allele, *let-60(n1046)* gives a Muv phenotype due to increased EGF/ras/MAPK signalling causing more cells along the ventral axis of the worm to adopt 1^{o} VPC fates (see 1.2.2). ~60% of animals carrying this allele exhibit a Muv phenotype 20^{o}C. Genes that enhance or suppress the Muv phenotype can therefore be screened for in this background. A complexity of screening for modulators of the Muv phenotype in the *let-60(n1046)* gain-of-function mutant is that the penetrance of the Muv phenotype is variable, leading to noise in the screens. An unpublished observation made by Andrew Fraser was that crossing of the *let-60(n1046)* gain-of-function allele into a *dgk-2(gk124)* loss-of-function background led to a 100% Muv strain. This suggests that *dgk-2* is a suppressor of EGF/ras/MAPK signalling in the vulva. RNAi screens for suppressors of the Muv phenotype were therefore also performed in *let-60(n1046);dgk-2(gk124)* animals. *lin-1(n303)* and *lin-15(n765)* are both loss-of-function alleles. LIN-1 is a transcription factor and downstream target of EGF/ras/MAPK signalling. Phosphorylation by MPK-1 results in inactivation of LIN-1. *lin-1(n303)* is therefore akin to a EGF/ras/MAPK gain-of-function mutation. 100% of *lin-1(n303)* animals exhibit the Muv phenotype. The *lin-15(n765)* mutation also appears to lead to increased EGF/ras/MAPK signalling, again leading to a 100% Muv population. The *lin-15(n765)* mutation corresponds to loss-of-function of synMuv genes *lin-15A* and *lin-15B*. This may lead to a increase in *lin-3* signalling to the VPCs from neighbouring hypodermal

cells. As a consequence the VPCs that adopt a 3^o fate in wild-type animal adopt 1^o fates leading to pseudo-vulval protrusions.

GENE NAME	% MUV ANIMALS	GENETIC BACKGROUND	GENE FUNCTION
M01B12.5	20	*let-60(n1046);dgk-2(gk124)*	putative RIO kinase
R10D12.10	20	*let-60(n1046);dgk-2(gk124)*	Serine/threonine kinase
pkc-1 a	28	*let-60(n1046);dgk-2(gk124)*	Serine/threonine kinase
pkc-1 b	31	*let-60(n1046);dgk-2(gk124)*	Serine/threonine kinase
D2096.12	41	*let-60(n1046);dgk-2(gk124)*	Protein kinase
D2096.8	72	*let-60(n1046);dgk-2(gk124)*	Nucleosome assembly protein
K08F11.5	79	*let-60(n1046);dgk-2(gk124)*	Predicted Ras related/Rac-GTP binding protein
F27E5.2	53, 17, 15	*lin-1(n303), lin-15(n765), let-60(n1046);dgk-2(gk124)*	PAX transcription factor

Table 3.4. Selected genes suppressing the Muv phenotype in RNAi screens in 100% Muv mutants. Indicated are the genes against which RNAi was performed, the average % Muv animals across the three screens, and the mutant backgrounds in which the hits were observed. A letter following the gene name indicates multiple individual clones used to independently target the same gene.

A total of 24 novel genes were identified as consistently suppressing the Muv phenotype in three independent screens. All of these genes could potentially be tested against the compendium of expression profiles. A set of 7 genes (table 3.4) were initially selected for testing. RNAi against all of these genes except one gave severe morphological defects in the animals. This was problematic for two reasons – firstly it made the animals extremely difficult to stage accurately; secondly, it made it likely that there would be considerable changes in expression as a result of somatic defects. Consequently these

genes were discarded. The one selected candidate modulator of EGF/ras/MAPK signalling that yielded a seemingly wild-type phenotype with slight brood-size defects on RNAi in N2 was *pkc-1*. Two RNAi clones targeting *pkc-1* exist in the library, both of which reduce the severity of the Muv phonotype in the *let-60(n1046);dgk-2(gk124)* mutant. This implies that *pkc-1* may be an enhancer of EGF/ras/MAPK signalling. When RNAi against *pkc-1* using the two different clones was tested against our compendium of array profiles *pkc-1* clustered with the known EGF/ras/MAPK pathway in both cases (figure 3.5). The Muv screens and expression profiling of *pkc-1(RNAi)* therefore provide two independent forms of evidence that *pkc-1* is involved in EGF/ras/MAPK signalling in *C. elegans*.

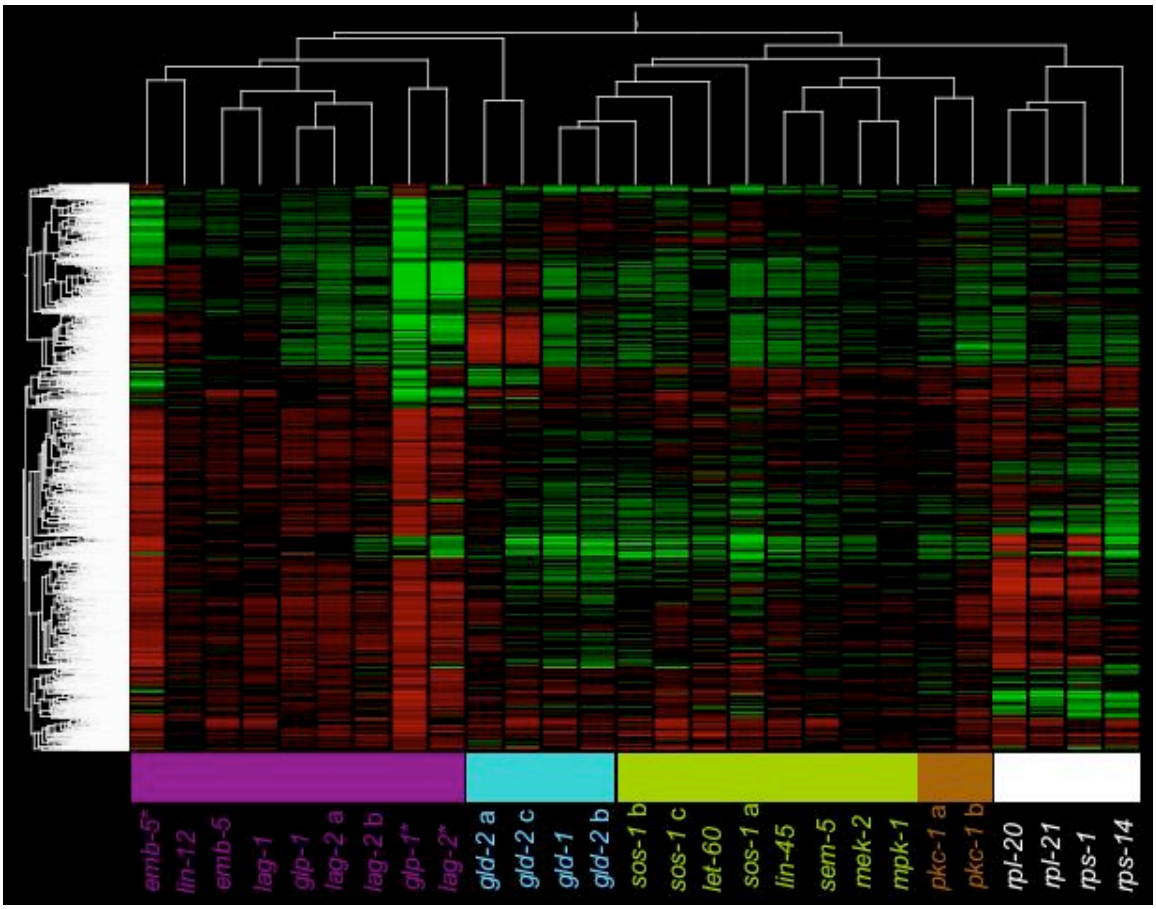

Figure 3.5. *pkc-1* clusters with the EGF/ras/MAPK signalling pathway. Genes and colour scheme are as in figure 3.1. The two RNAi experiments followed by expression profiling of *pkc-1* are labelled in brown. Lowercase letters following gene name indicates the use of different individual RNAi clones targeting the same gene. An asterisk indicates a genetic mutant rather than RNAi knockdown.

There is a well-established and conserved functional relationship between *pkc-1* and *dgk-2* (reviewed in Mellor and Parker, 1998; Merida *et al*., 2008; Nishizuka, 1984). *pkc-1* is an orthologue of mammalian protein kinase C, which is a diacylglycerol (DAG) dependent protein kinase. *dgk-2* is an orthologue of mammalian DAG-kinase, which phosphorylates DAG, converting it to phosphatidic acid. In this way it removes an essential factor for *pkc-1* activity (figure 3.6). Loss-of-function *dgk-2* therefore leads to increased *pkc-1* activity. That *dgk-2* loss-of-function increases the Muv phenotype in *let-60(n1046)* animals and *pkc-1(RNAi)* decreases it in *let-60(n1046);dgk-2(gk124)* is further

evidence that DAG signalling and EGF/ras/MAPK signalling are functionally related. Functional links between PKC and EGF/ras/MAPK signalling have previously been identified in mammalian and avian species (e.g. Banan *et al*., 2001; Crotty *et al*., 2006; Heo and Han, 2006; Lee *et al*., 2006a; Lee *et al*., 2006b; Sriraman *et al*., 2008).

The clustering of *pkc-1(RNAi)* as predicted amongst the other conditions in the compendium demonstrates our ability to provide further evidence of the signalling modulation indicated by the RNAi screens of the Muv phenotype. Our identification of *pkc-1* in this way represents a firm hit and will likely lead to further comparisons of screening-detected signalling modulators against our compendium.

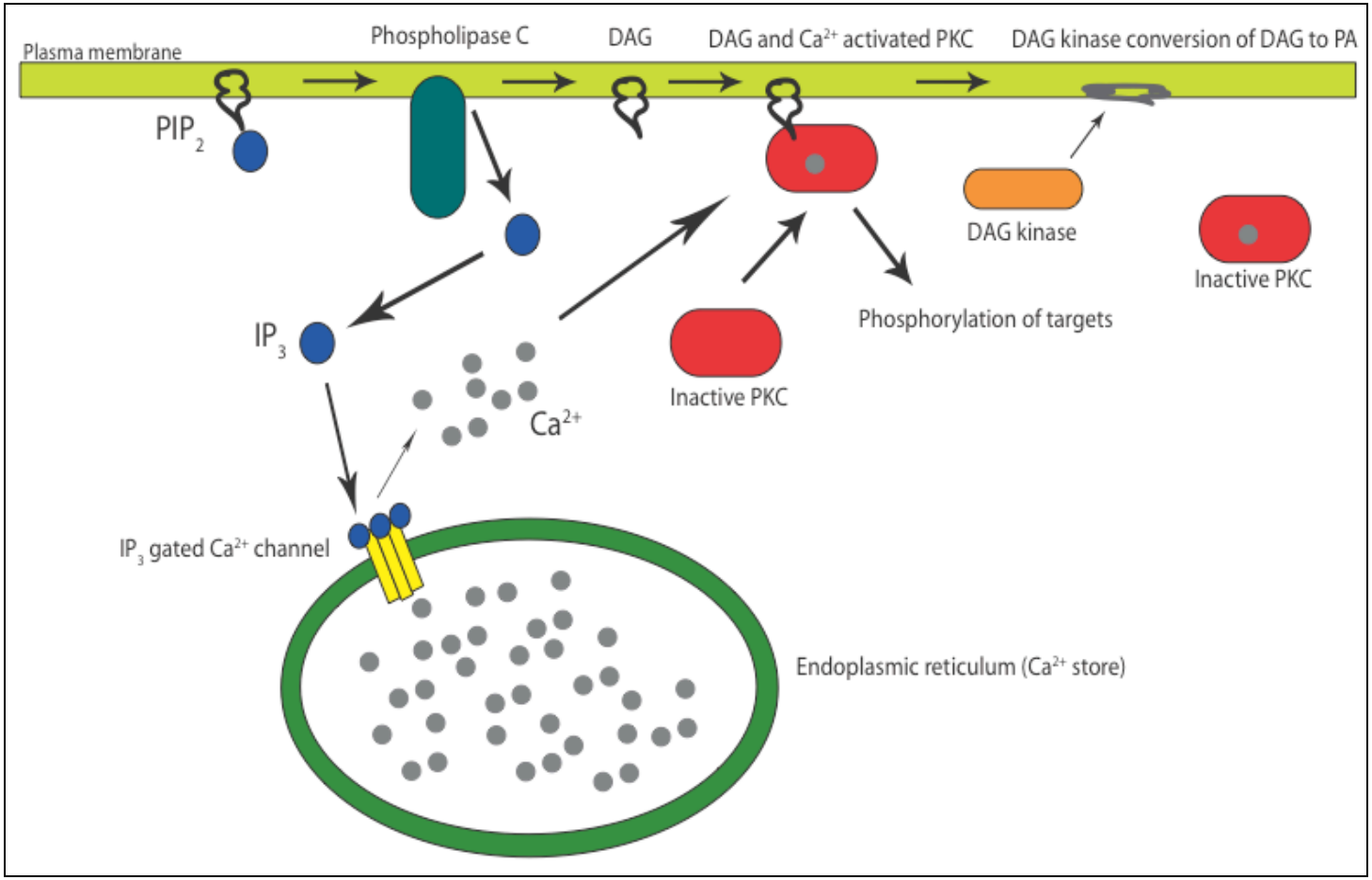

Figure 3.6. The activity of PKC is modulated by the activities of PLC and DGK. Phospholipase C (PLC) converts phosphatidylinositol bisphosphate (PIP_2) to inositol trisphosphate (IP_3) and diacylglycerol (DAG). Increased cellular IP_3 leads to the opening of IP_3 gated Ca^{2+} channels in the endoplasmic reticulum. Protein kinase C (PKC) is then activated by Ca^{2+} binding and tethering to the plasma membrane by DAG. DAG is converted to phosphatidic acid (PA) by DAG kinase (DGK). This results in PKC being released from the plasma membrane and inactivation.

3.7. The differentially expressed genes

It would be a missed opportunity to consider this data set only in terms of our ability to distinguish functional relationships between perturbed conditions. Rather, the genes that change in expression are likely to be of some interest in themselves. A number of papers from the Reinke and Kim labs over the years have used comparative expression profiling of mutant animals to identify genes enriched in the germline, gametes and both male and hermaphrodite soma (Jiang *et al*., 2001; Reinke, 2002; Reinke *et al*., 2004; Reinke *et al*., 2000). Our knowledge of the physiological changes caused as a result of perturbing these genes means that we know which parts of the germline should be enriched for each set of perturbations. We also have a number of perturbations in each category meaning that the number of times we see the same gene change in each can be a measure of our confidence that the expression of these genes is enriched in those regions. Specifically, genes upregulated in animals with Ras/MAPK signalling perturbations may be highly expressed in meiotic prophase. Conversely, the genes downregulated are likely to act after meiotic prophase, such as in gametogenesis. Genes downregulated on Notch perturbation are likely to be generally germline enriched genes. Upregulated genes may be enriched in the soma. Such lists of genes can be limited to genes specifically regulated only in certain conditions. For example, genes downregulated for every Notch perturbation but not downregulated for any other perturbation are highly likely to be mitotic-enriched genes. Genes upregulated for every Ras/MAPK perturbation and no other condition are more likely to be meiosis-enriched genes without contamination of soma-enriched genes. Genes up- or downregulated on either Notch or Ras/MAPK

perturbation and not the other or ribosomal perturbation are listed in appendix 1 (data CD), along with the number of perturbations of that class for which that regulation is seen.

As to the different general properties of the genes that fall into these classes, interpretation has proven difficult. Firstly, as is apparent from the clustering, the number of genes changing for any perturbation ranges from many hundreds to many thousands. Too much is changing for individual processes to be singled out. There is little functional information assigned to many genes and that which is, is often derived from their differential expression patterns observed in microarray experiments (e.g. sperm enriched genes). The identification of such genes being under-represented in a compendium of germline perturbations is not novel and of little biological value. An obvious analysis would be to see if any of our resulting gene lists are significantly enriched for any Gene Ontology (GO) terms – a set of definitions relating to gene properties or function. In *C. elegans* this is a fruitless endeavour as there are insufficient GO terms assigned to genes such that any statistical inference can be made. This is not to say that there is no value in this differential expression information beyond its ability to drive clustering of conditions. Numerous recent studies have applied the knowledge of common expression patterns amongst comparable conditions as the source data for biological network construction (Beer and Tavazoie, 2004; Freeman *et al*., 2007). This dataset may be ideally suited to such analysis, a possibility that is worth pursuing in future.

3.8. Discussion

Considering the progress made to this point it seems sensible to compare the approach relative to a more conventional staining approach. Array profiling is a powerful methodology and offers potential advantages over a staining approach for a number of reasons. Firstly, previous work as well as the data presented here has shown that the animal-to-animal variability of RNAi means that methodologies considering populations rather than individuals are more clean and powerful. Each RNA sample used in this study is derived from ~10,000 worms, many more than could be analysed post-staining for mitotic/meiotic markers. Secondly, microarrays offer an established technological platform that can test vastly more parameters than maximally 4-colour histological staining. It also lends itself to straightforward statistical analysis, which is preferable to counting large numbers of nuclei and attempting to categorise perturbations based on morphology and staining. The wealth of signalling components that lead to sterility may indicate hitherto unrecognized pathways and machineries involved in germline development. Our ultimate goal was to categorize such genes, which could potentially be beyond the capacity of current histological staining methods. A potential defect of this methodology, however, is that it is likely to be insensitive to physiological changes affecting only a few cells. Such changes are more likely to be identified by a detailed staining approach.

The rediscovery of the known biological machineries by clustering of the array profiles is firm evidence of our ability to place genes in pathways based on biological function. Since the clustering is inevitably very plastic and subject to change depending upon the

array profiles added to it, a method of testing the robustness of the clusters should perhaps be applied. An example of this would be a bootstrap approach. This would involve multiple rounds of removing random sets of genes and reclustering. The ability of the pathways to remain together in isolation within the clustering under these circumstances may act as an indicator of how strong the associations are within the clustering. It may also identify the key genes, which drive the clustering.

The obvious next step is the querying of more genes against the compendium. As previously stated, the list of candidates is vast including all signalling and transcription factor genes giving sterile phenotypes for as yet undetermined reasons. This list could be limited to genes that give sterile genetic interactions with components of the Notch or Ras/MAPK pathways i.e. genes that increase the brood-size defect of Notch or Ras/MAPK mutants by RNAi. A complexity of this is that genes identified in genetic interactions screens often interact with components of both pathways and others (Lehner *et al*., 2006). This hints at the complexities of interpreting genetic interactions but perhaps this expression approach represents an opportune system to study this.

It is clear that any such inference of gene function via a compendium such as this requires additional forms of evidence before inference can be considered confirmed. An obvious way in which this could be done is detailed dissection and staining of germlines. A number of markers have been suggested for immuno-staining of germlines (Crittenden and Kimble, 2008). These markers can be used to determine the relative quantities of each region of the germline. For example, GLP-1, FBF-1, FBF-2 or CEP-1 could be used

to mark the mitotic region. HIM-3 could be used as a marker of meiotic prophase, whilst RME-2 and SP56 mark the oocytes and sperm respectively. Staining of *pkc-1(RNAi)* germlines represents an obvious candidate for such staining. In this case we would expect to see an increase in the HIM-3 stained regions and decrease in RME-2 and SP56 stained regions relative to wild-type. That said, it is the complexities and limited resolution of this that was the motivation for this project in the first place. The limited brood-size defect for some of the conditions that appear in the compendium may indicate that germline staining may be inconclusive. The reality, however, appears to be that in order to assign genes to pathways at least a subset of novel genes added to the compendium would have to be evaluated in this way. For a subset of genes to exist many more genes would have to be tested against the compendium. Whilst obvious candidate genes exist, it was necessary to weigh the value of pursuing this project further against the potential of other projects to bear fruit. The project detailed in the following chapters was running concurrently with this in order to provide a fall-back position should this project have proven unworkable. Although this project appears far from unworkable it was not pursued further as it was deemed of lower potential than that which follows.