New Views on the *Drosophila* Transcriptome

Jing Wang

A thesis submitted for the degree of Doctor of

Philosophy at the University of Glasgow

Institute of Molecular, Cell and Systems Biology
College of Medical, Veterinary and Life Sciences
University of Glasgow
Glasgow G12 8QQ

January 2014

Abstract

Drosophila is a valuable experimental organism can be used as a reverse genetics model. *Drosophila* Malpighian (renal) tubules are important epithelial tissue in which to study transport mechanisms. RNA-seq has been chosen to investigate *Drosophila* Malpighian (renal) tubules to identify novel genes following a three-way comparison between three popular transcriptome profiling methods. Two types of novel gene have been found in *Drosophila* tubules, coding genes and noncoding genes. Reverse genetics has been applied to identify novel coding gene function in *Drosophila* tubules.

Three-way analysis of *Drosophila* expression microarrays, *Drosophila* tiling micrarrays and *Drosophila* RNA-seq reveal that most gene expression levels are well correlated between the three technologies. *Drosophila* expression microarrays and RNA-seq are correlated better than the correlation between *Drosophila* tiling microarrays and RNA-seq. *Drosophila* expression arrays and *Drosophila* tiling arrays all suffered from cross-hybridization, miss target detection and hybridization background noise, and also have low dynamic range for detecting lowly and highly expressed genes. *Drosophila* tiling microarrays also have a high false-positive detection rate, which may lead to overestimate the transcriptional activities of the genome. RNA-seq has overcome the drawbacks of microarrays and become the leading technology for genome sequencing, transcriptome profiling, novel gene discovery, and novel alternative splicing discovery with wide dynamic range. However, *Drosophila* expression microarrays and tiling microarrays still remain useful. Three-prime expression microarrays offer a means to measure the differential three-prime end processing, and tiling microarrays can be used for novel gene discovery. In this sense, the three technologies complement each other.

Poly(A) selected RNA-seq has been used as a discovery tool for searching novel genes in *Drosophila Malpighian* tubules in this thesis. A TopHat and Cufflinks pipeline has been used as an analytical pipeline for novel gene discovery and differential gene expression analysis between *Drosophila* tubules and whole flies in order to find the tubule-enriched genes.

Reverse genetics has been applied to *Drosophila* to achieve a gene knockdown and overexpression by using the unique Gal4/UAS system to achieve the novel gene knockdown or overexpression in specific tissue and cell types. Novel coding gene CG43968 has been discovered. The location of this gene has been confirmed in tubule main segments, principle cell cytoplasm or apical membrane. The function of this gene has been identified as involvement in tubule secretion, which may relate to calcium transport. Reverse genetics has been confirmed as particularly important for the functional study of novel genes.

Author's Declaration

The research reported within this thesis is my own work except where otherwise stated, and has not been submitted for any other degree.

Jing Wang

Table of Contents

List of Tables

List of Figures

Acknowledgements

Firstly, I would like to thank my supervisors, Professor Julian A.T. Dow and Dr. Pawel Herzyk, for their ideas and guidance throughout the project. My thanks also go to Dr. David F. Wilson and Dr. Venkat R. Chintapalli for all the enthusiasm and time they offered from the research work towards the thesis finish. Next I would like to thank everyone in the Dow/Davies laboratory during the course of my molecular genetics works and everyone in Glasgow Polyomics Group for all the support. Lastly, I would like to thank Tanita and Michael for all the encouragement and support through my PhD.

My thanks and gratitude also goes to my husband and my son for all their love, help and support over the years.

Abbreviations

BACs	bacterial artificial chromosomes
BDGP	the Berkeley *Drosophila* Genome Project
BLAST	Basic Local Alignment Search Tool
CPC	Coding potential calculator
dbEST	EST database
2DE	two-dimensional gel electrophoresis
DGEs	Differential gene expressions
ES	Electrospray
ESTs	Expression sequence tags
FPKM	fragments per kilo base of exon per million mapped reads
GPS domain	G-protein-coupled receptor proteolytic site domain
GCOS	GeneChip Operating software
HPLC	High-performance liquid chromatography
ICC	Immunocytochemistry
IGB	Integrated Genome Brower
IPA	Integrated data analysis
lincRNA	long intergenic non-coding RNA
MALDI	matrix-assisted laser desorption/ionization
MIAME	Minimum information about a microarray experiment
miRNA	microRNA
mlncRNA	mRNA like non-coding RNA
modENCODE	model organism ENCyclopedia Of DNA Elements
MPSS	Massively Parallel Signature Sequencing
NCBI	National Centre for Biotechnology Information
NGS	Next generation sequencing
NHGRI	National Human Genome Research Institute
ORFs	open reading frames

OD	optical density
piRNA	Piwi-interacting RNA
qPCR	quantitative reverse-transcriptase PCR
RIN	The RNA integrity number
RPKM	reads per kilo base of exon per million mapped reads
rRNA	ribosomal RNA
rmRNA-seq	ribo-minus RNA-sequencing
RNA-seq	RNA sequencing
RNAi	RNA interference
RMA	Robust Multiarray Average
SnoRNA	small nucleolar RNA
SNP	single nucleotide polymorphism
SAGE	serial analysis of gene expression
STSs	sequence-tagged sites
tRNA	transfer RNA
TAS	Tiling array analysis software
YACs	Yeast artificial chromosomes

1. **Introduction**

Summary

Since the genome sequence for the model organisms had completed more than ten years ago, great efforts have been made to annotate the genes' structures, predict and search for novel genes. The methods for predicting novel genes include experimental and computational approaches. The experimental approach includes 'open technologies' and 'close technologies'. The 'open technologies' refer to analysis of the transcriptome without *a priori* knowledge of the transcript sequences; technologies such as expression sequencing tags (ESTs), the serial analysis gene expression (SAGE) and next generation sequencing (NGS) are suitable for novel genes discovery. Next generation sequence has shown a great power for discovery of novel splicing and novel genes discovery in the whole genome level. The 'close technologies', such as microarrays, rely on previous sequence knowledge, and are suitable for comparing gene expression in different conditions. The computational gene prediction approaches include 'Extrinsic approaches', '*Ab initio* approaches', 'Combined approaches' and 'Comparative genomic approaches'. In many areas, experimental and computational approaches still provide complementary information. *Drosophila* is a powerful model organism for functionally characterising novel genes. Such novel coding and non-coding RNAs may play important roles in *Drosophila* development and functions. This study presents the tissue specific novel genes by using NGS technology and studies novel gene function by using *Drosophila* as a model organism.

1.1 **Experimental prediction approaches**

1.1.1 ***Expressed sequence tags (ESTs)***

ESTs are historically important, from days when Sanger sequencing was relatively expensive. Rather than fully sequence every clone in a cDNA library, effort was concentrated on sequencing just 5' and 3' ends cDNA using universal primers. ESTs are a single-pass sequence which is created by sequencing the 5' and/or 3' ends of randomly isolated gene transcripts that have been converted into cDNA, ESTs represent partial sequences of cDNA clones, and are typically within the range from 100-700 nucleotides.

1.1.1.1 **ESTs and gene discovery**

ESTs have applications in the discovery of new genes, identification of coding regions in genomic sequences (Adams et al., 1991; Adams et al., 1993a; Adams et al., 1993b) and identification of predicted genes. The first ESTs project was begun in 1991, and found 337 ESTs representing new genes out of 600 randomly-selected human brain cDNA clones (Adams et al., 1991). Since then, the identification of sequence using ESTs has developed rapidly, partly because EST collection is relatively quick and inexpensive by comparison with fully sequencing a given clone. The construction methods for EST libraries were improved gradually to facilitate the novel gene discovery, with random primed libraries or directional clones being most efficient method for discovering novel genes by ESTs (Adams et al., 1993b). Two large public sequence projects, the EST project and the Cancer and Genome Anatomy project (CGAP http://www.ncbi.nlm.nil.gov/ncigap) (Riggins and Strausberg, 2001), have been initiated to rapidly identify or partially identify all expressed genes (Martin and Pardee, 2000). To date, it has accumulated sequences for a total of approximately 74 million different ESTs; these are available in public databases (Genebank dbEST database 01 January 2013) for all species. The rates for novel gene discovery by the EST project were initially high, but declined sharply in recent years. Wang and his colleagues gave examples in whereby 10.4% of human ESTs collected in 1996 were novel sequences (36,000 novel sequences) whilst only 2.7% of ESTs collected in 1998 (638 novel sequences) were novel sequences (Chen et al., 2002b). This result indicates that the identification of novel genes by ESTs in human genome has nearly reached saturation. More methods are needed to identify the rest of the novel genes in the human genome and the genome of other species.

1.1.1.2 **ESTs and phylogenetic analysis**

ESTs are also a tool for phylogenetic analysis. The 5' ESTs, representing the coding sequence of the genes, are more conserved between species, the 3' ESTs, representing the 3' untranslated region, are more specific for the species. The

3' ESTs can help to separate closely related transcripts. Phylogenetic analysis reveals the relations and evolution of the species (Kullberg et al., 2007; Nishiyama et al., 2003). The EST approach allows, at a reasonable cost, a fast extension of data sampling from species outside the genome projects (Kullberg et al., 2007).

1.1.1.3 **ESTs and genome map**

EST information helps to construct the genetic map and physical map, and serves as a foundation for initiating the genome sequencing project. Sequence-tagged sites (STSs) are becoming standard markers for the physical mapping of the human genome. These short sequences from physically-mapped clones represent uniquely identified chrosomal locations (Adams et al., 1991). Yeast artificial chromosomes (YACs), bacterial artificial chromosomes (BACs) and other genomic resources facilitated by the use of STSs and PCR have been employed in building the physical map and gene map for different species (Hong-Bin Zhang, 2001). In 1998, the entire genome sequence of Caenorhabditis *elegans* was reported (The C. elegans Sequence Consortium, 1998). In 2000, the sequences of the euchromatic portion of *the Drosophila melanogaster* genome (Adams et al., 2000), the draft of rice genome sequence have been completed (Pennisi, 2000) and the entire genome of Arabidopsis had been completed by late 2000 (The Arabidopsis Genome Initiative, 2000). Other model species have been sequenced subsequently. The entire human genome project was finished in 2003 (Lander and Doyle, 2001)

All these physical maps of the model organisms provide the basis for the development of expression arrays, also known as DNA chips. Microarray technology emerged after the majority of the genomes of the model organisms being sequenced around 1998, and provided the opportunity to investigate gene expression pattern in specific stages, specific states and specific cell types to find out its biological role.

EST databases

In 1992, a database called dbEST (Boguski et al., 1993) was established to serve as a collection point for ESTs which were then distributed to the scientific community via the EST division of GenBank. The GenBank, which is maintained by the National Centre for Biotechnology Information (NCBI) also contains the CGAP information. The EST division continues to dominate the GenBank, accounting for roughly two-thirds of submissions. GenBank and dbEST sequences are organised into a non-redundant unique gene lists by the UniGene project, (http://www.ncbi.nlm.nih/gov/UniGene), which is considered the most regularly updated source for high quality, non-redundant information on expressed genes. Collections of full-length non-redundant cDNA clones are critical reagents for functional genomics. *Drosophila* Gene Collection release 1 (DGCr1) comprises full-length clones from approximately 40% of the 13,474 genes predicted in D. *melanogaster*. The second release of the DGC (DGCr2) extends the collection to more than 70% of the predicted genes in *Drosophila* (Rubin et al., 2000a; Stapleton et al., 2002). One of the most interesting applications of the EST database (dbEST) is gene discovery. Novel genes can be found by query the dbEST with a protein or DNA sequence (Boguski et al., 1994; Verdun et al., 1998).

1.1.1.4 **The disadvantages of ESTs**

Expressed sequence tag collection also has limitations when being used for genomic analysis from the accurate representation of genome content, gene sequence, and as windows into the transcriptome activity (Alba et al., 2004). The fact ESTs reflect the actively transcribed genes and represent transcriptome of the certain time and conditions of the tissue. So it is difficult to use EST sequence alone to represent an organism's gene content. Additionally, the fraction of the sequence data is erroneous due to enzyme used to generate the library, the technology for sequence and the analysis algorithms of the sequence data (Bebenek et al., 1989; Metzker, 2005). EST libraries have been shown to be biased towards highly expressed transcripts ; low abundance transcripts are rarely sequenced (Reese et al., 2000). So normalization and subtraction methods had applied in the cDNA library construction facilitate gene discovery in order to identify the lower expressed and specific type genes (Bonaldo et al., 1996; Gu et al., 2011; Verdun et al., 1998). Despite these limitations, it has been shown that EST database can be valid and reliable sources of gene expression and gene discovery data.

1.1.2 *Serial analysis of gene expression (SAGE)*

SAGE is a technique that allows a rapid, detailed analysis of thousands of transcripts (Velculescu et al., 1995). It is designed to provide qualitative and quantitative information on gene expression at the genome level.

In SAGE, short tags of length 9-10 bp obtained from the precise location of the 3' end of the transcripts are concatenated to form long DNA fragments, which can be cloned and sequenced. It allows many genes to be detected in a single lane sequence, and increases the efficiency of the sequence-based transcriptome analysis. Details of the principles underlying SAGE are shown in Figure 1.1.

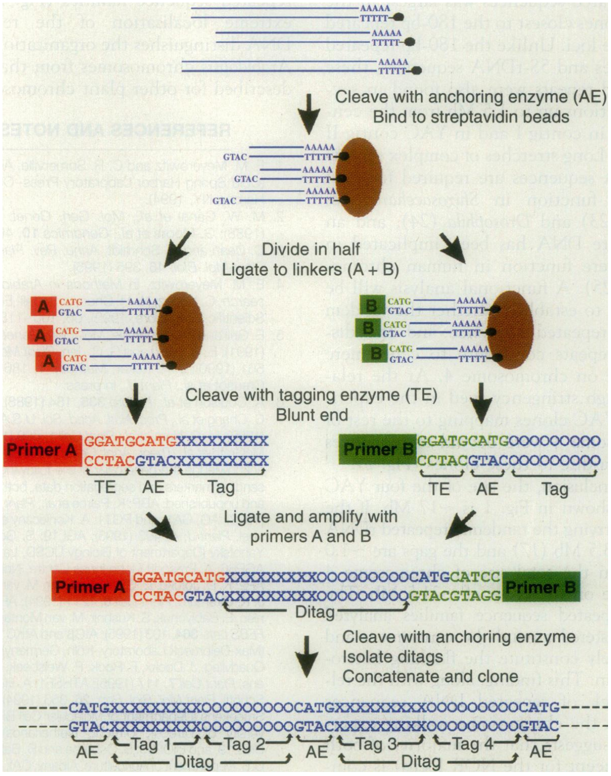


Figure 1-1 Schematic of SAGE.

A short sequence tag (10-14bp) contains sufficient information to uniquely identify a transcript provided that the tag is obtained from a unique position within each transcript; Sequence tags can be linked together to form long serial molecules that can be cloned and sequenced; and quantitation of the number of times a particular tag is observed provides the expression level of the corresponding transcript. Figure adapted from (Velculescu et al., 1995).

1.1.2.1 SAGE and novel gene discovery

The use of ESTs for novel gene discovery had reached saturation in human genome; SAGE as an 'open architecture' system that provided another approach to identify novel genes. One study by Chen *et al*, for example, used SAGE to identify novel genes and transcripts in the human genome. This study found about 70% of the unmapped SAGE tags are derived from the novel transcripts but were difficult to identify by previously available methods (Chen et al., 2002b). Some of these genes were completely novel genes with no matches in any of the expressed gene databases; some of them were from alternative splicing transcripts of known genes.

Another study applied the SAGE technique to identify the novel transcripts and used the SAGE tags as specific polymerase chain primers to amplify the unknown cDNA (van den Berg et al., 1999).

SAGE tags can be converted into cDNA for identifying novel genes to increase its accuracy. GLGI (Generation of long cDNA fragments from SAGE tag for gene identification) can be used for large scale identification of novel genes by converting novel SAGE tags into 3' cDNAs (Chen et al., 2002a). In this way, GLGI can be used as high-throughput procedure to identify the novel SAGE tags.

1.1.2.2 **SAGE and cancer research**

SAGE can use to compare the gene expression patterns in various developmental and disease states, and so has been a valuable approach for the identification of diagnostic and prognostic markers as well as therapeutic markers and transcriptional pathways (Argani et al., 2001; Polyak and Riggins, 2001). Hough et al constructed 10 different SAGE libraries to identify the makers that were up-regulated in ovarian cancer such as MUCI, HE4, Claudin3, Claudin4, SLPI and many more (Hough et al., 2000). An example of pathway analysis by SAGE led to the identification of 216 c-MYC-induced genes and 260 c-MYC-repressed genes that are potential drug targets or cellular markers of transformation (Menssen and Hermeking, 2002).

SAGE can detect gene expression in any cell type or tissue, and can determine the absolute gene expression level. Therefore, SAGE has been selected as the major platform technology for the Cancer Genome Anatomy project. Over 5 million SAGE tags derived from over one hundred human cell types have been assembled and released to the public domain through this project.

1.1.2.3 **SAGE and ESTs**

In comparison to the use of ESTs, SAGE offers a number of advantages. Firstly, SAGE can be performed without *a priori* knowledge of gene sequences, so it is useful for the identification of novel genes or the analysis of poorly characterized transcriptome (Hu and Polyak, 2006; Yamamoto et al., 2001). Secondly, it can quantify gene expression, describe absolute mRNA level and enable comparison of gene expression at different conditions, whereas ESTs can only deliver a single-pass sequence. Thirdly, it allows relatively high-throughput to produce information on genes in a short time; ESTs can only generate a partial gene sequence at each time. Fourthly, SAGE exhibits no bias in the analysis of gene expression especially for low abundance genes, whilst it is difficult to detect low expression genes using ESTs because of the expression bias towards to the high expression genes. Lastly and importantly, the transcript variations from alternative initiation and termination, alternative splicing, trans-splicing and antisense transcription can be revealed by using SAGE.

There are, however, also disadvantages to using SAGE. Firstly, SAGE requires a high amount of input RNA to start with. SAGE cannot be used for the generation of expression profiles when RNA is limited (Datson et al., 1999). Secondly, a 9-10bp tag can unambiguously identify the cDNAs, whereas it is not sufficient to map a gene to a genome precisely (Yamamoto et al., 2001). Thirdly, SAGE is expensive and still time consuming compared to the high-throughput sequencing nowadays due to the need to perform several thousands of PCR and sequence reactions. Lastly, since SAGE tags are cut by a specific enzyme most commonly for example *NlaIII*, any gene which doesn't contain the restriction enzyme cutting site will be missed (Yamamoto et al., 2001).

1.1.3 ***LongSAGE***

As shown in Figure 1-2, LongSAGE is a modified version of SAGE that generates 21bp tags derived from 3' ends of transcripts by using the type IIS restriction endonuclease (*MmeI*), which can rapidly be analysed and matched to genomic sequence data (Saha et al., 2002).

1.1.3.1 LongSAGE and novel gene discovery

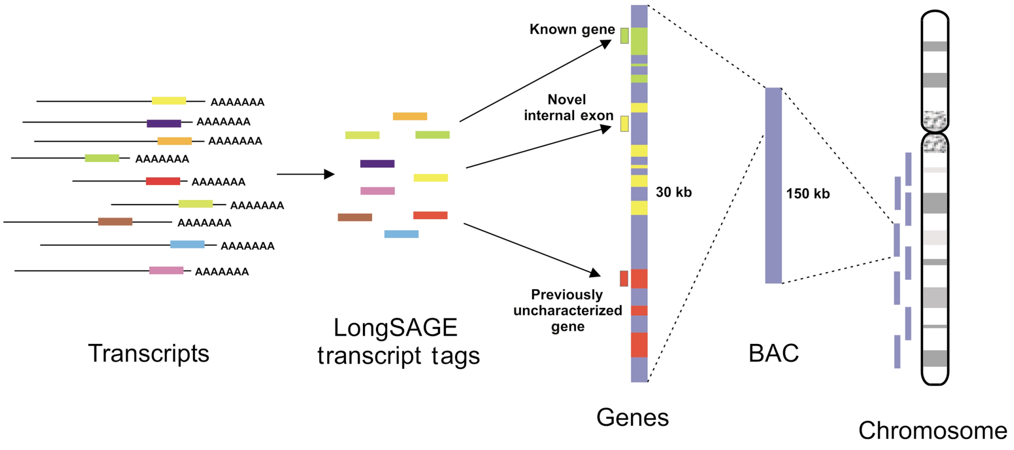


Figure 1-2 Schematic of LongSAGE methods.

LongSAGE is a modified version of SAGE that generates 21bp tags derived from 3' ends of transcripts by using the type IIS restriction endonuclease (*MmeI*), which can rapidly be analysed and matched to genomic. Figure adapted from (Saha et al., 2002).

The Saha group used LongSAGE to find 575 out of approximately 28,000 transcript tags that matched regions within introns of known genes representing either unknown exons of annotated genes or novel genes embedded in the intron of the known genes. They also found 803 out of approximately 28,000 transcript tags that matched regions at least 5KB from the terminal exons of known or predicted genes, representing completely novel genes. This study demonstrates how LongSAGE tags can identify previously unrecognized internal exons and uncharacterized genes (Saha et al., 2002).

LongSAGE tags are much more efficient for the identification of novel genes in the complex genome in comparison with conventional SAGE tags (9-10bp). The first LongSAGE analysis in mouse (Mus Musculus) found 2098 LongSAGE tags that fell into a region containing putative genes predicted by GenScan, providing the experimental evidence for the presence of real genes (Wahl et al., 2005b). The same study of mouse genome by LongSAGE also revealed a large number of novel antisense genes in the mouse genome (Wahl et al., 2005a).

1.1.3.2 LongSAGE and SAGE

LongSAGE and SAGE use the same principle to generate the sequence tags, but LongSAGE uses a different type IIS restriction endonuclease such as *MmeI* and incorporates other modifications to generate 21bp tags whereas SAGE uses type

IIS restriction endonuclease such as Fok I to produce 9-10 bp tags. Due to this increased tag length, LongSAGE can uniquely map the genome and transcriptome whilst SAGE will sometimes map the genome at multiple places. Consequently, LongSAGE is a more precise method for novel gene discovery.

1.1.4 *SuperSAGE*

SuperSAGE is a variant of SAGE technology, which supports transcripts profiling using 26bp tags extracted from cDNA employing the typeIII restriction enzyme EcoP151as a tagging enzyme. This tag length is the longest in use across all the versions of SAGE, and is advantageous for tag-gene annotation, thereby allowing the SuperSAGE technique to be applicable to any eukaryotic organism (Matsumura et al., 2008).

1.1.4.1 **Advantages**

SuperSAGE retains the benefit of using longer sequence tags whilst addressing a technical problem inherent in LongSAGE. LongSAGE improves on SAGE by generating longer tags to map the genome more precisely. However the digestion of a DNA fragment with *MmeI* generates a two-nucleotide recessed 5' terminus, which is difficult to fill in. To solve this technical problem, SuperSAGE uses EcoP151 digestion to generate a two nucleotide recessed the 3' terminus, which is easier to fill in.

1.1.4.2 **Applications**

The 26bp tag sequences can be used directly to design PCR primers for amplifying cDNA of corresponding genes; it can thus direct novel gene discovery. The sequencing of the *Nicotiana benthamiana* genome was assisted by this method (Matsumura et al., 2003).

SuperSAGE can be applied to interaction transcriptomes, analysing gene expression during host-pathogen interactions. This is possible since the method allows the simultaneous gene expression analysis of two or more eukaryotic organisms. This approach has, for example, been applied to study the gene expression profiles of both rice plants infected with blast disease and the causative fungus *Magnaporthe grisea* (Matsumura et al., 2003*).*

The SuperSAGE technology has been directly used to make a 26bp oligonucleotide array called “SuperSAGE array”. It combines the advantages of high quantitative SuperSAGE expression analysis with high-throughput microarray technology, allowing precise gene annotation and the monitoring of large-scale gene expression in many samples at a time (Matsumura et al., 2006).

SuperSAGE has also been combined with next generation sequencing to achieve high-throughput, high sensitivity, high reproducibility and accuracy in the analysis of gene expression and interpretation of the genome (Matsumura et al., 2010).

1.1.5 ***Functional Genomics***

Functional genomics is a field of molecular biology that enables exploration of genes, protein functions and interactions on a global scale. The goal of functional genomics is to elucidate function in the context of an organism’s genome (Dow and Davies, 2003). A key characteristic of functional genomics studies is using a genome-wide approach, generally involving high-throughput methods to study the functions of genes. Now that obtaining genome sequence has become routine, assigning function to genes is the current frontier in research (Hawkins et al., 2010). The functions of many genes and proteins are still unknown, or only partially described. Around 40% only of functions of *Drosophila* genes have been determined (Dow and Davies, 2003; Roy et al., 2010).

Microarray technologies have consistently been prominent in functional genomic studies as detailed in section1.2.6. This technology allows the researcher to take a snapshot of the gene expression under certain conditions and identify any change in gene expression between conditions that indicates when a gene is functionally active. The function of genes of interest or their encoded proteins can then be selected and studied using different molecular or physiological techniques. Microarray assays allow massive parallel data acquisition and analysis at a global level (Schena et al., 1998).

Next-generation sequencing is the cutting-edge technology applied in functional genomics as described in section 1.2.7. This technology, which overcomes the disadvantage of microarray, is now predominant in the field of gene expression studies (Shendure, 2008). Advance in DNA-sequencing technology delivers unprecedented insights into the entire collection of a genome's transcribed sequences. In this sense, the technology heralds a new era in the study of gene regulation and genome function (Graveley, 2008). It can quantify gene expression, address how alternative splicing affects the protein function, and determine the function of the non-coding genes, especially novel genes (Graveley et al., 2011). RNA sequencing (RNA-seq), small RNA sequencing (Small RNA-seq) and chromatin immunoprecipitation followed by sequencing (ChIP-seq) are the main methods applied in next-generation sequencing to address the function of coding genes, noncoding genes and also gene regulation in functional genomics (Bellingham et al., 2012; Chen et al., 2013; Mortazavi et al., 2008).

Functional genomics studies can also be undertaken using proteomics technologies. The classic proteomics method uses two-dimensional gel electrophoresis (2DE) to separate proteins. The identities of proteins in individual spots from the gel are then identified by mass spectrometry of their tryptic peptides. There are several different so-called workflows that can be used to characterize a proteome. They all use either 2DE or (High-performance liquid chromatography) HPLC as the separation methods and Electrospray (ES) or matrix-assisted laser desorption/ionization (MALDI) mass spectrometry as the identification methods. This technology can be used to investigate protein expression, protein-protein interaction, post-transcriptional modification, and to determine the protein function.

Functional genomics can also be applied to closing the phenotype gap. Finding all the genes that contribute to the phenotypes will help identify gene functions. Functional genomics draws heavily on reverse genetics (as detailed in section 1.4.3) to elucidate the function of novel genes. The phenotype gap (the mismatch between what a genetic model organism's genome encodes and the reasons that it has historically been studied) emphasizes the need to attract and empower functional biologists (Brown and Peters, 1996; Dow, 2007).

The field of systems biology is especially interested in the interpretation of large post-genomic datasets in a mechanistic context, from transcriptomics, proteomics and metabolomics to the dynamic modelling of systems behaviour and system-level datasets using a diverse collection of computational tools. To understand biology at the system level, we must examine the structure and dynamics of cellular and organismal function, rather than the characteristics of isolated parts of a cell or organism. A system-level understanding of a biological system can be achieved by understanding four key points, system structures, system dynamics, the control method and the design method (Kitano, 2002). This represents another level to achieving the goal of functional genomics.

In the future, integrative genomics and integrative biology will focus on integrating all the high-throughput data from multiple biological techniques to achieve an integrated multi-dimensional view of genomic function; integrative analysis offers the promise of a unified, global view (Hawkins et al., 2010).

1.1.6 ***Microarrays***

Once the whole genome sequences of the major organisms had been completed, searching the function and the structure of the genes is a long term task. This is “functional genomics”. Obtaining an overview of the global gene expression patterns in normal and disease conditions will enable researchers to develop understanding of gene spatio-temporal interactions and regulations. Microarray technology led the transition from studies of the individual biological functions of a few related genes, proteins or pathways towards more global investigations of cellular activity. Microarray technology began in 1989, and was announced to the wider scientific community in a publication by Schena et al that made researcher aware of the potential of array technology (Schena et al., 1995). Schena and colleagues described the high capacity of cDNA microarray technology to monitor the gene expression of 45 *Arabidopsis* genes in parallel. This represented a major advance over Northern blotting, which reported expression level of only one gene at a time. Since then, the use of microarray technologies has been reported for multiple organisms, including yeast (Lashkari et al., 1997), *Drosophila* (White et al., 1999) and human.

1.1.6.1 **Types of microarray**

Microarray technology varies in terms of manufacturing method and detection method. There are two types of arrays in terms of manufacturing method, namely spotted arrays and oligonucleotide arrays. In spotted microarrays, the probes are oligonucleotides (oligos), cDNA or small fragments of PCR products that correspond to the genes that are "spotted" on the glass slide. Oligonucleotide microarrays, typically refers to the specific technique of manufacturing used by companies such as Agilent where the oligos are longer sequences such as 60-mer probes and Affymetrix where the oligos are shorter sequences 25-mer probes; in both cases, the oligos are synthetic in origin, rather than derived from DNA clones.

There are two detection methods: one-colour microarrays or two colour microarrays. In one-colour microarray, one sample is processed, labelled for example with fluorescent dye, and applied to a microarray, such as those available for Affymetrix. In two-colour microarrays, two samples that are to be compared are labelled with different fluorophores and put on one microarray. The relative intensities of each fluorophore are used in ratio-based analysis to identify up-regulated and/or down-regulated genes. The fact that samples share the same background will significantly reduce any background effect and increase the sensitivity of detection (Tang et al., 2007).

Affymetrix microarrays

Affymetrix (www.affymetrix.com) is a company based in United States that manufactures DNA microarrays (also called GeneChips). The company manufactures different types of array; expression arrays, exon arrays, tiling arrays and miRNA arrays of different organisms. The company now designs chip technology aimed towards clinical diagnosis.

Three-prime expression microarrays

The expression arrays are the first generation of the Affymetrix microarrays. The probes are designed to be complementary to the target sequences at the 3'UTR of the annotated, predicted genes and ESTs which called a consensus sequence in Affymetrix parlance (Cui and Loraine, 2009). Each gene is represented by multiple probe pairs (also known as probe sets) which are used to measure the level of transcription of each ORF sequence represented on the Genechip. Each probe set has 25mer probe pairs selected from the target sequence to be perfect match and mismatch oligos. Each probe-pair consists of a perfect-match (PM) and mismatch (MM) probe. The PM probe is a 25-base sequence complementary to the target gene, whilst the MM probe is identical to the PM probe but a single mismatch at 13th base. The sequences on the expression arrays are believed to recognize unique regions of the three-prime of the gene. Figure 1.3 detailed the Genechip design method.

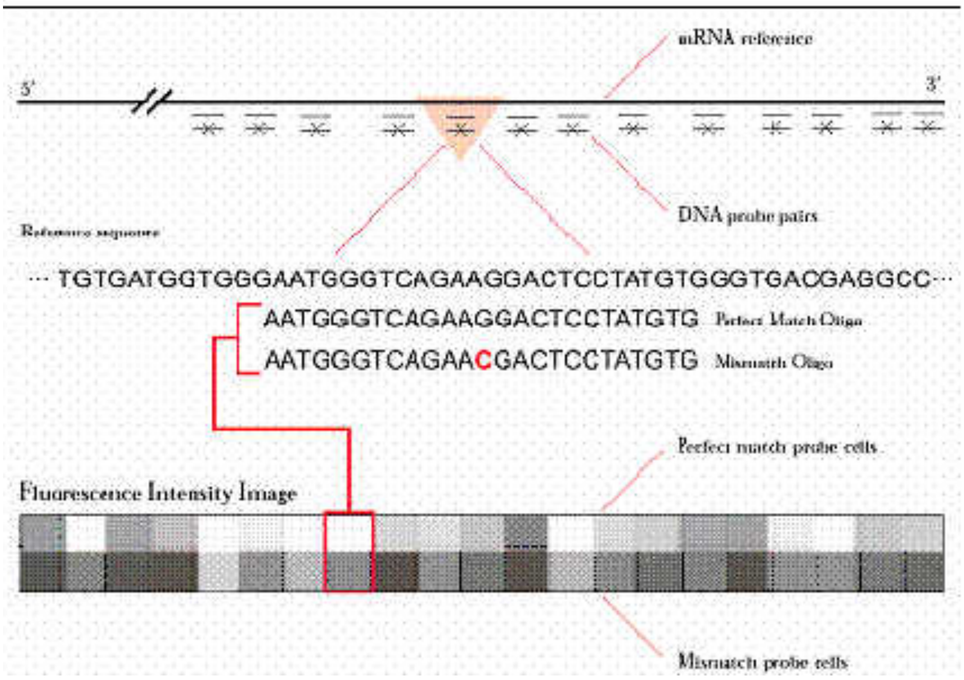

mRNA reference
5'
3'
DNA probe pairs
Reference sequence
··· TGTGATGGTGGGAATGGGTCAGAAGGACTCCTATGTGGGTGACGAGGCC···
AATGGGTCAGAAGGACTCCTATGTG
Perfect Match Oligo
AATGGGTCAGAACGACTCCTATGTG
Mismatch Oligo
Perfect match probe cells
Fluorescence Intensity Image
Mismatch probe cells

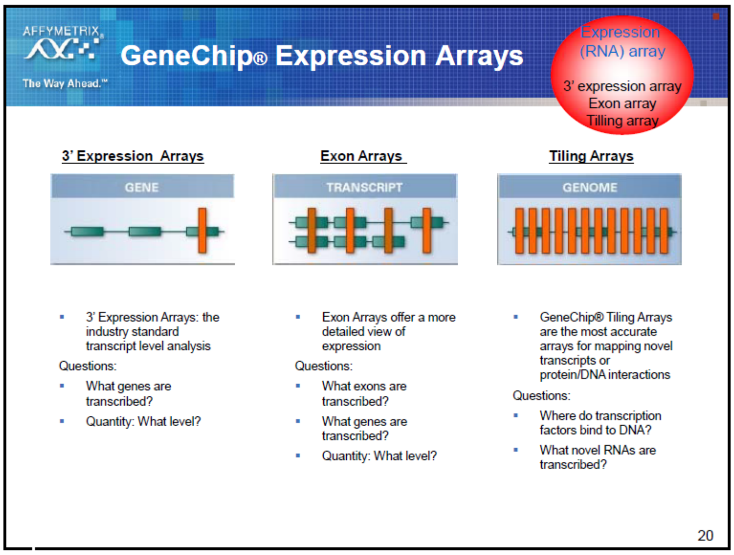
Figure 1-3 A schematic of a Affymetrix probe set.

Each gene is represented by multiple probe pairs. Each probe-pair consists of a perfect-match (PM) and mismatch (MM) probe. The PM probe is a 25-base sequence complementary to the target gene, whilst the MM probe is identical to the PM probe but a single mismatch at 13th base Picture taken from www.vsni.co.uk/software/genstat/htmlhelp/marray/AffymetrixChips.htm

The *Drosophila* Genome 2.0 Array was designed with sequence and annotation from FlyBase *Drosophila* Genome draft version 3.1, the Berkeley *Drosophila* Genome Project (BDGP) and additional public content from the *Drosophila* community. The array contains 18,880 probe sets covering over 18,500 transcripts. Fourteen pairs of oligonucleotide probes are used to measure the level of transcription of each ORF sequence represented on the Genechip Drosophila Genome 2.0 Array.

Tiling Microarrays

Tiling arrays are designed with the probes tiled across the whole target genome. The probes for some arrays are partially overlapping such as the S. *cerevisiae* Tiling 1.0R Array, whilst some arrays have non-overlapping probes such as the *Drosophila* tiling 2.0R Array (details also in chapter 3.2.1.2). Tiling arrays can be used for a range of applications including genome mapping, novel gene discovery, DNA-protein interaction (ChIP-chip) and DNA methylation studies. The comparison of different types of microarray design are illustrated in Figure 1.4, the probes of 3'-end expression array are at the 3'-end of the genes, exon array probes are designed in each known exons of the genes and tiling array probes are tiled across the whole genome.



GeneChip® Expression Arrays

Expression (RNA) array

3' expression array
Exon array
Tilling array

3' Expression Arrays

- 3' Expression Arrays: the industry standard transcript level analysis

Questions:

- What genes are transcribed?
- Quantity: What level?

Exon Arrays

- Exon Arrays offer a more detailed view of expression

Questions:

- What exons are transcribed?
- What genes are transcribed?
- Quantity: What level?

Tiling Arrays

- GeneChip® Tiling Arrays are the most accurate arrays for mapping novel transcripts or protein/DNA interactions

Questions:

- Where do transcription factors bind to DNA?
- What novel RNAs are transcribed?

20

Figure 1-4 Diagram of different types of Affymetrix Microarrays.

The picture shows the design strategy of different type arrays. 3' expression arrays' probes at 3' end, exon arrays' probes at major exons and tiling microarrays' probes across the genome. Pictured adapted from Affymetrix Web http//: www. Affymetrix.com.

1.1.6.2 **Microarray and transcriptional profiling**

A transcriptional profile is the main application of microarray that can measure gene expression patterns, gene structure and gene functions at the whole genome level. The whole genome expression array is designed for this purpose. The first whole genome microarray was employed for yeast in 1997; the arrays contained up to 2,479 yeast open reading frames (ORFs). The results of three experiments showed that many genes were differentially expressed under the three environmental conditions (Lashkari et al., 1997). Transcriptional profiling analysis can be used in disease diagnosis, and the analysis of gene expression in cancer including disease pathology, progression, resistance to treatment, response to cellular microenvironments, and may ultimately lead to improve early diagnosis and innovative therapeutic approaches for cancer (DeRisi et al., 1996). Expression analysis using microarray has also been applied in the toxicological research to define how the regulation and expression of genes mediate the toxicological effects associated with exposure to a chemical (Bartosiewicz et al., 2001a; Bartosiewicz et al., 2001b). For *Drosophila*, the Dow/Davies lab created a FlyAtlas website (www.flyatlas.org). This web helps the researchers all over the world to design the correct experiments to look for the gene expression pattern in specific tissues (Chintapalli et al., 2007; Wang et al., 2004).

1.1.6.3 **Microarray and genotyping**

Another main application of microarrays is their use for comparative genomic analysis. The use of microarray technology for genotyping is further advanced than for transcript profiling as illustrated. Single nucleotide polymorphism (SNP) is the most frequent type of variation in the genome. A single nucleotide polymorphism array (SNP array) is a useful tool for studying slight variations between whole genomes. Specific uses of this technology include determining individual genome information (Redon et al., 2006), determining disease susceptibility (Botstein and Risch, 2003) and measuring the efficacy of drug therapies (Martinelli et al., 2009), SNPs can also be used to study genetic abnormalities in cancer (Bacolod et al., 2009).

1.1.6.4 **Microarray and novel gene discovery**

Traditional molecular approaches to identifying genes, including cloning and sequencing large collection of cDNAs (ESTs), have succeeded at identifying tens of thousands genes, they eventually reach a point of greatly diminished returns. Transcripts that are low abundance or expressed in rare cell types or in response to specific stimuli may never be identified by these methods (Mockler et al., 2005). Microarray can be used to solve some of these problems, allowing confirmation of the predicted genes models such as expression arrays as well as a tool for novel gene discovery for example Tiling arrays. Tiling arrays have the probes tiled the whole genome, covering essentially all nonrepetitive regions of the genome, and so enable the discovery of novel genes or novel alternative splicing. Human tiling arrays have been used to interrogate chromosomes 21 and 22 via 25-mer probes spaced on average every 35bp. The human tiling array studies used human cells line and tissue samples. The results indicated that a much larger portion of human genome is transcribed than was previously predicted and also revealed the activity of novel noncoding genes in human genome (Cawley et al., 2004; Kampa et al., 2004; Kapranov et al., 2002). *Drosophila* tiling arrays has been used 25-mer oligonucleotide probes, spaced evenly across the *Drosophila* genome at intervals of approximately 35 base pairs. The studies using tiling arrays of *Drosophila* genome show that 85% of the fly genome is transcribed and processed into mature transcripts, representing 30% of the fly genome and 30% of detected embryonic transcription is unannotated (Manak et al., 2006). Tiling array studies of 25 Drosophila cell lines also revealed more than one thousand novel transcribed regions (Cherbas et al., 2011). *Drosophila* tiling arrays will be discussed further in chapter 3. Custom exon arrays can also be used as a gene discovery tool for detecting novel splice junctions, which can subsequently use to find novel genes.

1.1.6.5 **Genomic DNA mask for probe selection method**

Although oligo-nucleotide arrays are a powerful and widely used tool for large-scale gene-expression profiling, most commercial arrays (Affymetrix arrays) are only available for model species. For example, *Drosophila* expression arrays are only available for *Drosophila melanogaster* but not available for other *Drosophila* species. Hammond and his colleagues developed a method to improve the sensitivity of high-density oligonucleotide arrays when applied to heterologous species by using ‘Genomic DNA based probe selection strategy’ on the available species’ arrays (cross-hybridization) to mask off the heterologous sequences between the two species to improve the sensitivity of the gene expression detection. This is a potential gene discovery method for non-model species (Davey et al., 2009; Hammond et al., 2005).

1.1.6.6 **Microarray data analysis**

Microarray data analysis is the most difficult challenge in microarray development. Microarray results are different from chip to chip, from lab to lab, and even from operator to operator. The issue here is how to normalize the results to make them comparable? The normalization method is the key step. Normalization is a process that adjusts microarray data for effects that arise from the variation in the technology rather than from the biological differences. There are a variety of normalization schemes in use, including total-intensity, ratio-based and both linear and nonlinear regression techniques (Quackenbush, 2001, 2002).

RMA (Robust Multiarray Average) and GC-RMA are a very popular normalization method for microarrays especially for Affymetrix microarrays (Irizarry et al., 2003). Details also referred to Chapter 2, Section 2.7.4.

It is important to deposit microarray data in a format that can be used by others, such as NCBI, Geo and Array express repositories. Minimum information about a microarray experiment (MIAME) is the first successful submission method for microarray data to bring at least some basic standard to a microarray-based assay (Brazma et al., 2001). This standard information makes microarray data more useful and comparable.

There are no definite methods for data analysis but some commercial software and self-made pipelines are applied to the analysis of microarray data, such as Partek (Downey, 2006), Genespring, and Bioconductor which is a major package written in the R statistical language.

1.1.6.7 **Advantages and disadvantages of microarrays**

The microarray is the first technology that allows a global view of the gene expression patterns in the genome. It allows comparative genome analysis to find the SNP, copy number variation, novel genes and alternative splicing.

The disadvantage is that microarray technology uses the hybridization values rather than digital count values to measure genes expression. As a result, microarray doesn’t generate absolute gene expression values and the hybridization values are subjected to background noise. Microarrays require *prior* knowledge of the genome and do not support de novo sequences. For novel gene discovery, the background noise effects inherent in the technology interfere with the identification of gene boundaries so that the information necessary for novel genes is subject to error. Normalization methods are rather difficult to apply to ensure that the microarray analyses can be compared to each other.

1.1.7 *Next-generation sequencing*

In theory, a superior approach would to be sequence every RNA in a sample completely. This has recently become possible. The automated Sanger method for sequences is considered as a 'first-generation' technology and newer methods are referred to as next-generation sequencing (NGS). Unlike the expensive, low throughput, single-pass Sanger sequencing, these newer technologies constitute various strategies that rely on a combination of template preparation, sequencing, imaging, and genome alignment and assembly methods. The major advance offer by NGS is the ability to produce an enormous volume of data cheaply. This will completely change our view in basic, applied and clinical research. There are several platforms used in NGS, based on different principles that have different advantages. RNA-seq is a revolutionary application of NGS for transcriptome profiling and novel gene/novel isoform discovery that will change the outlook for the genome annotation. Third-generation technologies are emerging quickly, for which the reduction in equipment size together with the ability to sequence more cheaply will eventually allow the technology to enter clinical diagnosis. Personalized genomes will benefit everybody.

1.1.7.1 **Commercial platforms currently on the market**

The four dominant commercial platforms currently on the market are the Roche 454 Genome sequencer, the Illumina Genome Analyser (GAI, GAIIx, Hiseq, Miseq), the Life Technologies SOLiD system and the IonTorrent system *(IonProton and PGM)*.

The comparisons of the four technologies are listed in table 1-1.

Table 1-1 Comparison of next-generation sequencing platforms

Platform	Roche454	IonProton	IlluminaGAIIx	Illumina Hiseq2500	ABI SOLiD3
Sequence mechanisms	Pyrosequencing	SBS	SBS	SBS	Ligation and two-base coding
Instrument Cost	$128,000	$243,000	$695,000	$740,000	$595,000
Read length (bases)	up to 1kb	200	75	2x100	50
Gb per run	1 M	60-80 M	18- 35Gb	600 Gb	30-50 Gb
Run time (days)	0.4	0.15	4-9	11	7-14
Pair-end	No	No	Yes	Yes	Yes
Observed Raw Error Rate	1.07%	1%	0.76%	0.26%	0.1%
Reported Accuracy	99.9%	99%	98%	98%	99.94%
Sequence cost per Gb	$310	$16.67	$148	$46	$40
Insert size	700 bases	150 bases	up to 700 bases	up to 700 bases	200-10.000bases
Typical DNA requirement	50-1000ng	100-1000ng	50-1000ng	50-1000ng	10ng-5ug
Advantage	read length, fast	cheap, fast	high throughput	high throughput	accuracy
Disadvantage	error rate with polybase more than 6, high cost, low throughput	short read assembly	short read assembly	short read assembly	short read assembly

1.1.7.2 **Main applications of Next-generation sequencing**

Due to the low cost and high throughput, NGS technologies have a range of application areas. A number of possibilities have arisen due to the fact that it can sequence the genome, provides a digital measure of gene expression, and does not require prior knowledge. The main applications are:

De novo sequencing and assembly

De novo sequencing **is** the initial sequence analysis performed to obtain the primary genetic sequence of a particular organism. Many non-model organisms don't have their genomes sequenced, due to the cost and time involved in determining the sequence. Next generation sequencing makes possible *de novo sequencing* with low cost, less labour and high throughput. By now a number of species have been sequenced. Up to March 2010, 740 sequence projects have been submitted to NCBI, of which 23 have been completed. The other projects are in progress or in draft form (Zhou et al., 2010).

Whole genome resequencing

Due to the fact that reference genomes are now available for many organism, cataloguing sequence variation and understanding its biological consequences has become a major research aim (Stratton, 2008). Whole genome resequencing for chicken successfully found more than 7,000,000 single nucleotide polymorphisms, almost 1,300 deletions and a number of putative selective sweeps. This information reveals the loci under selection during the domestication of chicken (Rubin et al., 2010). Whole genome resequencing is another genome-wide method for genotyping and copy number analysis that follow on from the use of microarray technology but offers greater accuracy and a more direct means of revealing the genetic variation and risk of disease (Michaelson et al., 2012). In addition to *de novo* mutations, DNA translocation can also be discovered by whole genome resequencing. Whole genome resequencing can also identify viruses, bacteria and other organisms present in complex biological samples by identifying their genome signature, in what is called the subtractive approach (Wilson, 2012).

Transcriptome profiling analysis – RNA-seq

RNA-seq is a recently developed deep sequencing technology for both mapping and quantifying transcriptomes. This technology overcomes the limitation of microarray to become an alternative technology that is able to measure the whole genome expression by measuring the sequenced reads of the transcriptome.

Under the RNA-seq process, the RNA population is fragmented and converted to a cDNA library with adaptors attached to one end or both ends. The DNA with adaptor will become attached to solid flow cells under bridge amplification, and then be sequenced by synthesis in a high-throughput manner to obtain short sequences from either one end (single end) or both ends (pair-end). This is just one method. The other method is strand-specific RNA-seq. Details of these approaches as used in the research conducted for this thesis is given in methods 2.8.

RNA-seq differs from the previous technologies in specific area. Firstly, RNA-seq doesn’t require the prior knowledge to detect the expression of transcriptome.

As a result, RNA-seq is particularly attractive for non-model organisms for which the genomic-sequence has not been determined, a scenario where it would be very difficult to apply to microarray technology.

Secondly, RNA-seq can clearly detect the boundaries of genes, including particularly novel splicing junctions. The technique can also identify noncoding genes and anti-sense RNAs (Young et al., 2012). It is the most advanced technology for analysis the whole genome for novel gene and isoform discovery. The results from RNA-seq suggest the existence of a large number of transcribed regions in every genome surveyed, including *Drosophila* (Graveley et al., 2011), mouse (Mortazavi et al., 2008), Human (Pickrell et al., 2010; Wang et al., 2008), *S.cerevisiae* (Nagalakshmi et al., 2008) and *S. pombe*. (Wilhelm et al., 2008).

Thirdly, RNA-seq is not based on hybridization so RNA-seq is not affected by background noise. RNA-seq can detect abundant expression without saturation; microarrays suffer saturation when the signals are too high.

Detail of comparison of the technologies is summarized in table 1-2, and further discussed in chapter 3 of this thesis.

Table 1-2 Comparison of technologies for transcriptomes analysis

Technologies	EST	SAGE	Microarray	RNA-seq
Principle	Sanger sequence	Sanger sequence	Hybridization	High- throughput sequence
Throughput	Low	Better than EST	High	High
Resolution	Single base	Single base	25-100 bp	Single base
Background noise	NO	NO	YES	NO
Quantitation	NO	YES	Relative	YES
Mapping	Portion	Portion	High	High
De novo sequence	YES	YES	NO	YES
Novel gene detection	Limited	YES	NO	YES
Novel splice junction	YES	NO	Limited	YES
Novel isoform detection	Limited	NO	Limited	YES
Gene structure detection	YES	NO	NO	YES
RNA required	High	High	High	Low
Cost for mapping large genome	High	High	Low	Relative low

Small RNA profiling analysis- miRNA-seq

miRNA-seq is another application of next generation sequencing. MicroRNA, which is normally 19-25bp long in length, modulates protein expression through transcript degradation, inhibition of translation, or sequestering transcripts. miRNA-seq uses size selection from the total RNA, followed by the addition of sequencing adaptors, and RT-PCR amplification and then sequencing. miRNA-seq has been successfully applied to the discovery of novel miRNA (Morin et al., 2008); and the identification of biomarkers for cancer classification, response to therapy, and prognosis (Keller et al., 2011). Further difference expression pattern analysis can identify the regulatory networks of miRNA involved in particular disorders.

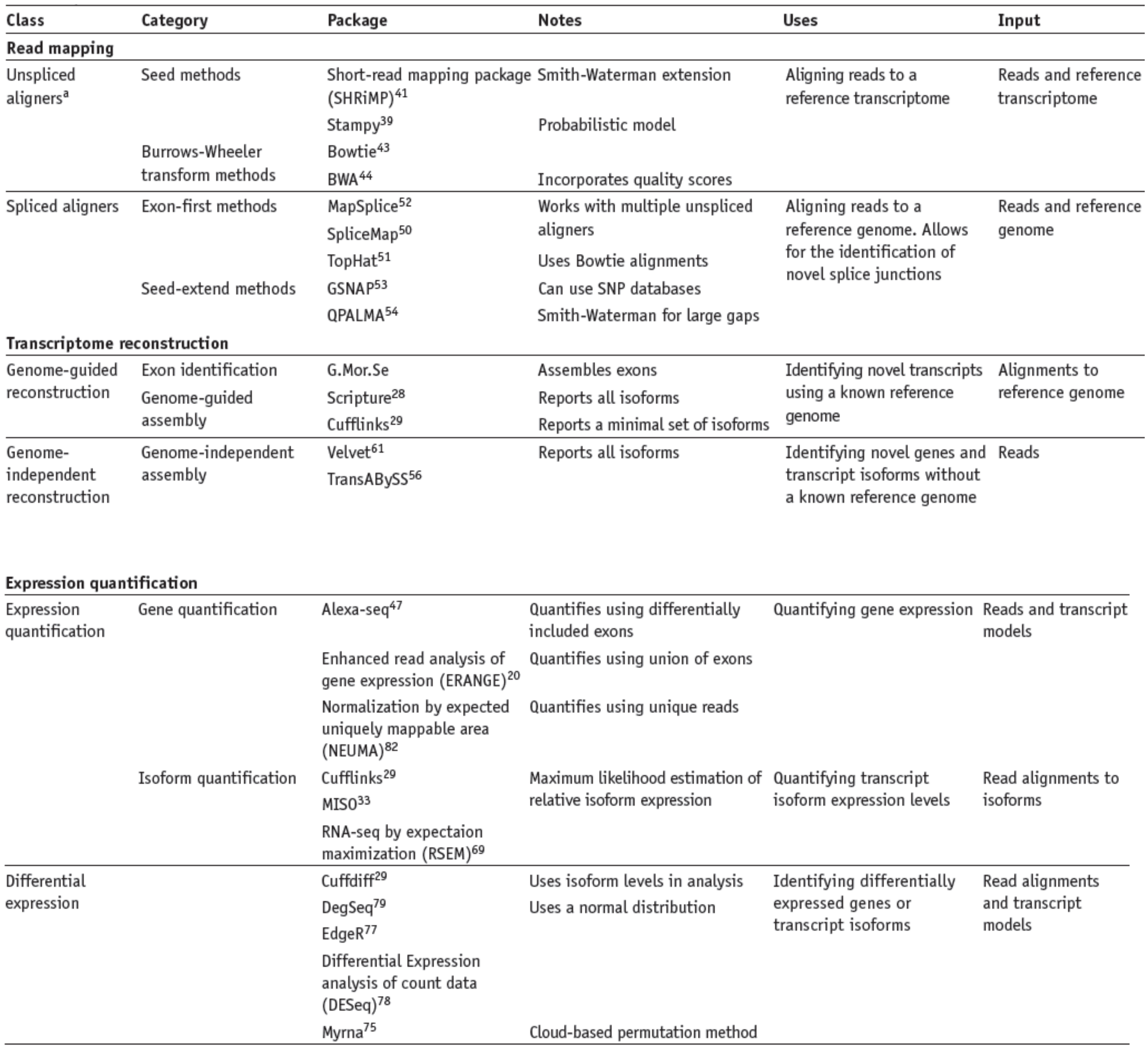
ChIP-seq

ChIP-seq is a method to analyze DNA-binding proteins and DNA interactions. ChIP-seq combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins. Firstly, ChIP applies the immunoprecipitation method using antibodies against transcription factors to pull down the DNA and DNA-binding protein complex. The DNA and protein are then unlinked. Secondly, all the resulting ChIP-DNA fragments are sequenced simultaneously after size selection using a genome sequencer. ChIP can be applied to discover novel noncoding RNAs that regulate the promoters of genes (Guttman et al., 2009). ChIP-seq also has the potential to detect mutations in binding-site sequences, which may directly support any observed changes in protein binding and gene regulation (Northrup and Zhao, 2011).

1.1.7.3 **Main analysis programs applied in next generation sequencing**

Different next generation sequencing analysis programs are listed in table 1-3.

Table 1-3 RNA-seq analysis programs



Class	Category	Package	Notes	Uses	Input
Read mapping					
Unspliced aligners[a]	Seed methods	Short-read mapping package (SHRiMP)[41]	Smith-Waterman extension	Aligning reads to a reference transcriptome	Reads and reference transcriptome
		Stampy[39]	Probabilistic model		
	Burrows-Wheeler transform methods	Bowtie[43]			
		BWA[44]	Incorporates quality scores		
Spliced aligners	Exon-first methods	MapSplice[52]	Works with multiple unspliced aligners	Aligning reads to a reference genome. Allows for the identification of novel splice junctions	Reads and reference genome
		SpliceMap[50]			
		TopHat[51]	Uses Bowtie alignments		
	Seed-extend methods	GSNAP[53]	Can use SNP databases		
		QPALMA[54]	Smith-Waterman for large gaps		
Transcriptome reconstruction					
Genome-guided reconstruction	Exon identification	G.Mor.Se	Assembles exons	Identifying novel transcripts using a known reference genome	Alignments to reference genome
	Genome-guided assembly	Scripture[28]	Reports all isoforms		
		Cufflinks[29]	Reports a minimal set of isoforms		
Genome-independent reconstruction	Genome-independent assembly	Velvet[61]	Reports all isoforms	Identifying novel genes and transcript isoforms without a known reference genome	Reads
		TransABySS[56]			
Expression quantification					
Expression quantification	Gene quantification	Alexa-seq[47]	Quantifies using differentially included exons	Quantifying gene expression	Reads and transcript models
		Enhanced read analysis of gene expression (ERANGE)[20]	Quantifies using union of exons		
		Normalization by expected uniquely mappable area (NEUMA)[82]	Quantifies using unique reads		
	Isoform quantification	Cufflinks[29]	Maximum likelihood estimation of relative isoform expression	Quantifying transcript isoform expression levels	Read alignments to isoforms
		MISO[33]			
		RNA-seq by expectaion maximization (RSEM)[69]			
Differential expression		Cuffdiff[29]	Uses isoform levels in analysis	Identifying differentially expressed genes or transcript isoforms	Read alignments and transcript models
		DegSeq[79]	Uses a normal distribution		
		EdgeR[77]			
		Differential Expression analysis of count data (DESeq)[78]			
		Myrna[75]	Cloud-based permutation method		

Note this table includes the different mapping methods (the popular unspliced aligner is Bowtie, the popular spliced aligner is TopHat); different transcription reconstruction methods (the most popular methods are Cufflinks, Scripture, and Velvet) and different transcription quantification methods (the popular methods are Cufflinks for expression quantification and Cuffdiff for differential expression quantification). Table adapted from (Garber et al., 2011).

Commercial software: Typical example of integrated software packages in use for analysis NGS Include CLC bio Genomic Workbench (www.clcbio.com), Partek Genomic Suite (www.partek.com) and Galaxy (www.galaxy.org).

CLC bio Genomic Workbench is commercial NGS software package extends the CLCbio. Main Workbench to provide support for SNP detection, CHIP-seq analysis, browser visualization and other features. This integrated software package offers tools for *de novo* and reference assembly of Sanger, Roche FLX, Illumina, Helicos and SOLid data.

Partek Genomic Suite is next-generation sequencing and microarray analysis software, including support for gene expression and digital gene expression (DGE), exon/alternative splicing and RNA-Seq, copy number and association, ChIP-chip, ChIP-seq, and microRNA in a single software package that allows for analysis of multiple applications in one complete solution.

Galaxy is an open, web-based, easy to setup platform for the analysis of next generation sequencing and genomic data. Galaxy provides tools to manipulate large dataset from RNA-seq and ChIP-seq to genome mapping and annotation.

Open source software : TopHat (Trapnell et al., 2009), Cufflinks (Trapnell et al., 2010), Bowties, Cuffdiff, Cuffmerge and CummeRbund pipeline. Details are also listed in Chapter 2, Section 2.8.6.

1.1.7.4 **Future trends in next-generation sequencing**

Next-generation sequencing has already been used in clinical research will likely enter clinics soon, but there remains a lot of challenges.

With the availability of a multitude of platforms and dramatically lower costs of sequencing, NGS technologies are expected to have a major impact on the way we practice medicine in the near future. It is not far for next generation sequencing to enter the clinical diagnosis if the whole human genome sequences reach to $1000 (Service, 2006). The third generation sequencer Ion Proton is expected to reach this $1000 goal by the end of this year with the P3 chip.

The combination of genomic information along with a detailed molecular analysis of the samples will be important for understanding the onset, progression, and prevalence of disease states (Chen et al., 2012). Building a personalized genome database is important step for health care and drug treatment to be made unique to different patients.

There are still more challenges for NGS to overcome beyond the £1000 limit and the lack of personalized genome information. Firstly, improved sample preparation techniques, PCR-free sequencing will bring down the cost and simplify data analysis by eliminating the major source of artefacts; single molecule sequence will much less than pair-end and mate-end due to its inaccuracy. Secondly, more robust analytical tools that have open source flexibility combined with friendly and efficient user interface and proper data storage software are key issues for next-generation sequencing goes into clinics. Thirdly, NGS data needs to be more accessible through visualization will reduce reliance on specialised bioinformatician. In summary, making NGS ready for clinical use will require personalized genome and technological advances that reduce the cost, complexity of use and equipment size. So it can enter the GP's surgery.

1.2 Computational prediction approaches

There are two approaches for novel gene discovery; one is experimental approach which we had discussed in the previous section. The following section will focus on the analysis and prediction by using computational approach.

1.2.1 *Extrinsic (similarity or evidence-based) gene finding systems*

Extrinsic gene finding systems locate target genes by comparing the RNA or protein sequences under study with all other RNA or protein sequences registered in databases to look for similarity. A high degree of similarity to a known RNA or protein product is strong evidence that a target gene is a protein-coding gene. Approximately 20-50% of newly found genes contain an ancient conserved region that is represented in the database (Fickett, 1996). Basic Local Alignment Tool (blast.ncbi.nlm.nih.gov) is a widely used system designed for this purpose, with several variants. This system provides task specific tools such as blastn (nucleotide blast) for searching a nucleotide sequence database using a nucleotide query; blastp (protein blast) for searching a protein sequence database using protein query; blastx for searching a protein sequence using a translated nucleotide query; tblastn for searching translated nucleotide database using a protein query; tblastx for searching a translated nucleotide database using a translated nucleotide query.

As extensive transcript and protein sequence databases have been generated for human as well as other important model organisms in biology, such as mouse, yeast, *Drosophila* and *C.elegans* so these extrinsic methods are quite popular to use. However, to apply this approach systemically requires extensive sequencing of mRNA and protein products. Although the RefSeq database, Ensembl system and NCBI database contain transcripts and protein sequence for many species, these databases however are both incomplete and contain a number of errors. One specific issue needs to be addressed is the limited availability sequences and protein products for tissue-specific genes, and for some genes that are only expressed at certain times. These limitations mean that the extrinsic evidence for many genes is not yet available.

1.2.2 ***Intrinsic (Ab initio approaches)***

Ab initio gene prediction approaches use statistical and computational methods to detect coding regions, splice sites, and start and stop codes in genomic sequences. These signs can be broadly categorized as either, specific sequences that indicate the presence of a gene nearby termed ‘signals’, or statistical properties of protein-coding sequence itself termed ‘content’. The *Ab initio* approach is the predominant gene prediction approach, due in large part to the fact that it doesn’t depend on sequence similarity and is therefore not limited by the availability of sequence data. Instead, understanding gene structure is the key step to predicting genes.

In the prokaryotic genomes, genes have specific and relatively well-understood promoter sequences (signals), such as the Pribnow box (TATAAT) and transcription binding sites that are easy to identify systematically. Genes that code for proteins comprise open reading frames (ORFs) consisting of a series of codons that specify the amino acid sequence of the protein for which the gene codes. The ORF begins with an initiation codon, usually but not always ATG, and ends with a termination codon that can be TAA, TAG or TGA. Searching for a DNA sequence that begins with an ATG and ends with a termination triplet is a start towards gene annotation. Statistically, one would expect a stop codon approximately every 60-75bp in a random sequence so a much longer stretch without a stop codon is good evidence for an open reading frame. These characteristics make prokaryotic gene finding relatively straightforward and well-designed systems will achieve a high level of accuracy.

Ab initio gene finding in eukaryotes, especially complex organisms like humans, is considerably more challenging for several reasons. Firstly, the promoter and other regulatory signals in these genomes are more complex and less well understood than in prokaryotes. Secondly, the main problem for the human genome and those of other higher eukaryotes is that gene sequences are often split by introns and so do not appear as continuous ORFs. Many ORFs that continue into introns are subject to termination due to the presence of stop codon within introns. Due to the relatively small length of exons compared to introns, simple ORF scanning cannot locate gene sequences. For example, many exons are smaller than 100 codons whilst some are less than 50 codons in length. Thirdly, there is substantially more space between real genes in the human genome and those of higher eukaryotes (70% of human genome is intergenic), increasing the chance of finding spurious ORFs.

Given these issues, three modifications to the basic procedure for ORF scanning have been adopted for eukaryotes. The first of these modifications is codon bias by which not all codons are used equal frequently for particular organism. The second modification is that exon-intron boundaries can be used as a signal to identify genes. The third modification is that upstream control sequences can be used to locate the regions where genes begin. Additional strategies are also possible for specific organisms, such as the identification of CPG islands and binding sites for a poly(A) tails.

GLIMMER and GeneMark software programs are widely used, highly accurate gene finders for prokaryotes (Aggarwal and Ramaswamy, 2002). Eukaryotic *ab initio* gene finders, by comparison, have achieved only limited success, as in the GENSCAN and Geneid programs (Peters et al., 2007). Advanced gene finders for both prokaryotic and eukaryotic genomes typically use complex probabilistic models, such as Hidden Markov models (HMMs), in order to combine information from a variety of different signals and content measurements. Seven *ab intio* programs were evaluated on a nonhomologous mammalian data set by Rogic et al. They reported that among the evaluated programs only GeneScan and HMMs gene were able to predict the precise location of 70-80% coding exons with low false positive rates (Rogic et al., 2001).

1.2.3 *Combined approaches*

Combined approaches bring together extrinsic and *ab initio* approaches by mapping protein and EST data to the genome in order to validate *ab initio* predictions. The *ab initio* approaches have delivered maximum accuracy of 70-80% (Rogic et al., 2001). The similarity search programs are very effective in improving the accuracy of gene prediction. In particular, combining the two methods can improve the overall accuracy by 4-10% (Issac and Raghava, 2004). Usually, *ab initio* gene prediction and similarity searches are run independently with the output from these two approaches being manually integrated for gene annotation. Many automated programs have been developed to combine the two approaches such as GenomeWise, the TwinScan, GenomeScan and EGPred (Issac and Raghava, 2004). The GenomeScan program for gene prediction was developed as an extension of Genescan and incorporates similarity searching for protein detected by BLASTX. GenomeScan is able to predict coding regions missed by using both GeneScan and BLASTX alone, leading to an improvement in the accuracy of gene prediction by 10% (Mathe et al., 2002).

1.2.4 *Comparative genomic approaches*

Comparative genome approaches rely on the sequence similarity to predict genes in a new species by comparison with an already sequenced relative. This approach is based on the principle that nature selection causes genes and other functional elements to undergo mutation at a slower rate than the rest of the genome. This means that the coding regions of genes are more conserved than noncoding regions under evolutionary pressure. Comparison of a few closely related genomes has proved successful for the discovery of protein-coding genes (Kellis et al., 2003). Stark et al. used a comparative analysis of twelve *Drosophila* genomes to predict non-protein-coding RNA genes and structure, and new microRNA (miRNA) genes (Stark et al., 2007). Comparative genomic analysis constitutes a powerful approach for the systematic understanding of any genome.

1.3 *Drosophila* is an ideal model organism for novel gene discovery

1.3.1 *Drosophila as a genetic model organism*

Drosophila has proven to be an excellent model organism for genetics studies. In part, this is due to the organism's small physical size, short development lifecycle and ease of culturing. These facts contribute to *Drosophila* being efficient and cheap to maintain in the lab. More importantly, *Drosophila* provides an excellent balance between genetic power and biomedical similarity to humans; 70% of *Drosophila* genes have clear human homologs (Chien et al., 2002); Genetically, *Drosophila* has a well-defined, fully sequenced, mid-sized genome. The genome encodes approximately 13,600 genes, In comparison to that of *Caenorhabditis elegans*, another widely used model organism, the *Drosophila* genome is longer but encodes somewhat fewer genes with greater functional diversity (Adams et al., 2000). Furthermore, *Drosophila* is a less complicated genome, having a core proteome only twice the size of that of yeast (Rubin et al., 2000b). The relative simplicity and manipulability of the fly genome means we can address some of these biological questions much more readily than in vertebrates (Rubin et al., 2000b). Consequently, *Drosophila* is a very good experimental model.

Drosophila has four pairs of chromosomes, named, X, 2L, 2R, 3L, 3R, 4 in the female, and Y, 2L, 2R, 3L, 3R, 4 in the male. The chromosome X and 4 both have a major left arm and significantly smaller right arm. The *Drosophila* genome is approximately 180Mb in size, which is roughly the size of a single human chromosome. One third of the *Drosophila* genome is centric heterochromatin. The two large autosomes and the X chromosome contain 120Mb of euchromatin; the small fourth chromosome contains only approximately 1Mb of euchromatin (Adams et al., 2000).

Drosophila has a fully sequenced genome (so far 12 different species have been sequenced) and detailed annotations. In addition, a number of online resources have been created for *Drosophila* that offers a wealth of genetic and physiological information. Examples include (flybase.org) and FlyAtlas (flyatlas.org). Fly genetics is enriched with powerful genetic markers and balancer chromosomes that facilitate marking genes of interest (the knockdown or overexpress) and tracing the lineages over the generations. The *Drosophila* stock centres maintain 40,000 genetically characterized lines, some of which are human disease models, together with RNAi stocks for all *Drosophila* genes. In addition, classic mutants and P-element insertional alleles, tissue-specific and cell type-specific drivers enable the generation of millions of transgenic flies for research.

1.3.2 *The Drosophila malpighian tubules as a model for epithelial fluid transport*

For functional genomics, it is important to be able to study function in a specific tissue. For many genes, Malpighian tubule is ideal. This section is detailed in Chapter 5, Section 5.1.6.

1.3.3 *Drosophila and reverse genetics*

Reverse genetics is an approach to discovering the function of a gene by analyzing the phenotypic effects of specific gene sequences obtained by DNA sequencing (Adams and Sekelsky, 2002). This has been proposed as a quick and promising way of inferring function for a novel gene (Dow and Davies, 2003) . This approach is illustrated in Figure 1-5. The success of the reverse genetics approach depends on the model organisms, and gene homology and phenotypes available. A variety of model organisms with appropriate genetic power, full genomic sequence and defined physiological knowledge can be chosen to serve as reverse genetic models. *Drosophila* is an excellent model organism for creating a gene knock-down because of its unique GAL4/UAS system (details in Chapter 5, Section 5.1.1), and its use in this context benefits from the higher level of homology to higher organisms. There are various RNAi lines for all genes available in stock centres; In particular, the UAS/GAL4 system is particularly advantages for creating gene knock-downs in specific cell types.

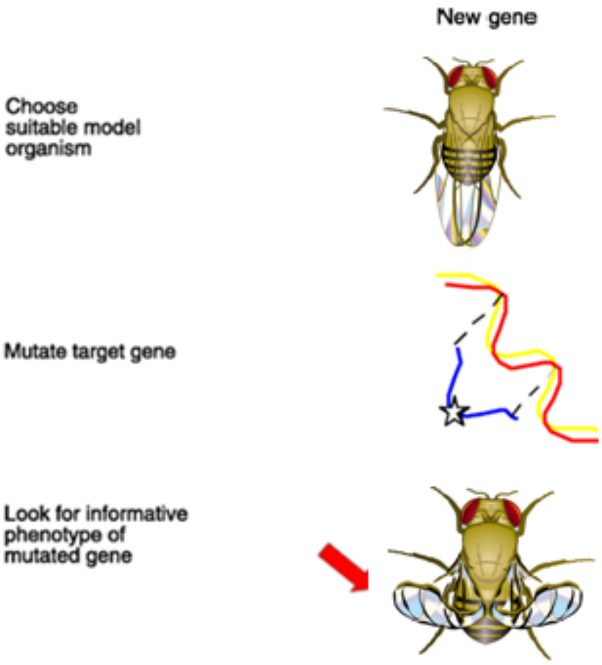

New gene
Choose suitable model organism
Mutate target gene
Look for informative phenotype of mutated gene

Figure 1-5 Diagram of reverse genetics.

Reverse genetics is a popular method for investigating novel gene function. The reverse genetic method firstly requires choosing a suitable model organism (*Drosophila* is an ideal model organism). Thereafter, the predicted novel gene is mutated (by gene knock-down or gene overexpression), so that potential changes in the phenotype and the function of the novel gene can be investigated. Picture adapted from (Dow and Davies, 2003), modified by Jing Wang.

There are two main methods in reverse genetics, gene knock-down (RNA interference, RNAi) and gene overexpression. There are two ways to introduce RNAi into *Drosophila* in search of associated phenotypes: one is through microinjection into the embryo. The other method is expressing RNA with a long inverted repeat that can fold back on itself to become double-stranded (Lam and Thummel, 2000). Different vectors have been developed to introduce RNAi into the cells or embryos. In several *Drosophila* RNAi vectors, a functional intron used as the linker sequence increases the effectiveness of RNAi (Lee and Carthew, 2003). One such novel vector that has been developed for RNAi is pRISE (Figure 1-8a). The pRISE vector contains a characteristic repeat of the Gateway recombination cassette *attR1-cm*^{1}*-ccdB-attR2* enabling insertion of the same target sequence in both orientations using Gateway Technology (Invitrogen). This involves cloning a trigger sequence into an appropriate entry vector, such as pENTR/D-TOPO (Invitrogen), and placing it between the *attL1* and *attL2* recombination sequences. An inverted repeat sequence between an intron can be transferred easily to pRISE by an *in vitro* reaction mediated by LR, which also has a pentameric GAL4 binding sequence for conditional expression (Kondo et al., 2006) Details can been see in Figure 1-6. This method is especially suitable for investigating novel gene function in *Drosophila*.

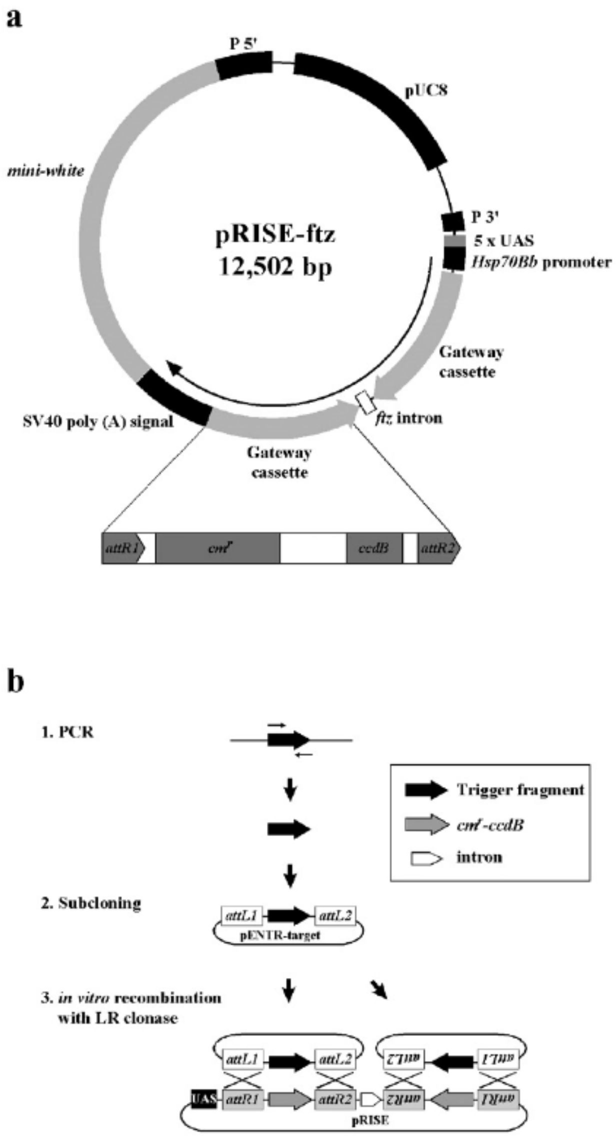


Figure 1-6 Rapid construction of RNAi transgene by pRISE.

a. **Physical map of pRISE. b. Schematic representation of RNAi transgene construction using pRISE**. Clone a trigger sequence into an appropriate entry vector, placing it between the attL1 and attL2 recombination sequences. An inverted repeat sequence in both orientations between an intron can be transferred easily to pRISE by an in vitro reaction mediated by LR, which also has a pentameric GAL4 binding sequence for conditional expression. Picture adapted from (Kondo et al., 2006).

Other ready-to-go vectors for generating overexpress (PTW) transgene lines by using Gateway technology which attached the downstream of UAS can combine with the GAL4 line create a powerful tool for functional analysis of the novel genes.

Tagged fusion proteins are priceless tools for monitoring the activities of biomolecules in living cells. PTWV is the destination Entry/Gateway® vectors for expressing fluorescent fusion proteins in *Drosophila melanogaster* (Akbari et al., 2009).

1.3.4 ***Drosophila and transgenesis***

Transgenesis in general can be defined as a group of technologies that allow DNA to be introduced into an organism of choice. The main goal of transgenesis is to integrate a foreign piece of DNA (a transgene) into an organism's genome to result in germline transmission (Venken and Bellen, 2007). In order to identify a novel gene, the *P* element construct or *P* element mediated RNAi vector with the novel gene (reporter gene) will be introduced into the organism's genome by germline transformation. In subsequent generations produced after crossing with the RNAi driver line, the novel gene will be identified by a specific expression pattern or specific tissue expression.

- ***P*** **element-mediated transgenesis**

The classic method for fly transgenesis is *P* element-mediated transgenesis, which has been one of the most important breakthroughs in germline transgenesis in *Drosophila*. *P* elements are transposable elements, or transposons, that were originally identified within the fly's own genome; these can cause gene mutation by 'random jumping' around the genes. The *P* element-mediated fly germline transformation is also detailed in Chapter 5, Section 5.1.2. Enhancer trapping which involves generating a *P* element construct that carries a reporter gene is widely used in *Drosophila* for generation of cell type markers that are often exquisitely specific. This enables identification of novel genes on the basis of expression pattern. Specification of expression patterns will indicate the location and function of the gene. The first generation of enhancer trapping used P-element-mediated trangenesis to detect tissue-specific genes and reveal regional specification in *Drosophila* tubules (Sozen et al., 1997). However, The *P* element-mediated transgenesis has two major drawbacks: the size of the DNA that can be integrated is limited and the insertion sites cannot be controlled (Venken and Bellen, 2007).

- **GAL4/UAS system**

The second generation of the enhancer trap is the UAS/GAL4 system, which also used the *P* element-mediated method.

Transposon-mediated transgenesis enable the development of the GAL4/UAS binary system of adapted *P* elements, for tissue-specific expression of introduced DNA sequences. Target gene expression in a temporal and spatial fashion has proven to be one of the most powerful techniques to address gene function *in vivo* (Duffy, 2002). The GAL4/UAS system was first developed for targeted gene expression in *Drosophila* (Brand and Perrimon, 1993). At present, it is a technique specific for *Drosophila*. Details are listed in Chapter 5, Section 5.1.1.

- RNA interference (RNAi)

Due to the several disadvantages of *P* element-mediated transgenesis, various improvements in fly transgenic techniques have been made. The most popular method used with fly is RNA interference (RNAi). Details about how the RNAi works are detailed in Chapter 1, Section 1.3.3.

1.4 Categories of RNA in *Drosophila* cells

As summarized in Table 1-4, multiple categories of RNA can be found in *Drosophila* cells. Understanding the categories of RNA is important to determine the novel genes categories which were found in RNA-seq technology later in this thesis.

Table 1-4 Types of Total RNA

Types	Name	Size	Location	Number in *Drosophila*	Functions	Poly-A tail
Coding RNA	mRNA		cytoplasm	13,6000	Coding Protein	Yes
Noncoding RNA	tRNA		cytoplasm	297	deliver amino acid	No
	rRNA		cytoplasm	101	place to make protein	No
	Short ncRNAs					
	miRNAs	19–24 bp	widespread locations	78	Targeting of mRNAs and many others	Yes
	piRNAs	26–31bp	nuclei and cytoplasm	50	Transposon repression, DNA methylation	No
	Mid-size ncRNAs					
	snoRNAs	60–300 bp	Nuclei and cytoplasm	255	rRNA modifications	No
	Long ncRNAs					
	lincRNAs	>200 bp	widespread loci	1,00	Examples include scaffold DNA–chromatin complexes	No
	mlncRNA	>200	Cytoplasm And nuclei	100		Yes

tRNA (transfer RNA), rRNA (ribosomal RNA), microRNA (miRNA), piRNA (Piwi-interacting RNA), snoRNA (small nucleolar RNA), lincRNA (lonng intergenic non-coding RNA), mlnRNA (mRNA-like non-coding RNA). Location (position in cell)

The coding RNAs of *Drosophila* are very well studied. The noncoding RNAs have been extensively studied in recent years especially following the development of high-throughput sequencing technology. As a result, more and more noncoding RNAs have been discovered (Table 1-5) such as mRNA-like noncoding RNA (mlncRNA).

As is the case for mRNA, polyadenylation also plays a role in mRNA-like noncoding RNAs. In particular, mlncRNAs that are like mRNAs are spliced capped, and polyadenylated just like protein-coding genes, and so may play important role in cellular processes. The introns of these RNAs are conserved in species. They lack open reading frames so they are unlikely to make protein, but these RNAs have important functions in cellular processes (Hiller et al., 2009; Jiang et al., 2011). Exons and introns of mlncRNA can be processed into microRNAs (miRNA) or small nucleolar RNAs (snoRNAs) (Kapranov et al., 2007). Other mlncRNAs exert their function as large RNAs such as the *hsrw* RNA which is key to the heat shock response in *Drosphila* (Arya et al., 2007*)*. None of the mlncRNAs have been well studied across *Drosophila* species.

1.5 **Aims of the project**

1.5.1 ***Primary aim/Searching for novel genes in Drosophila tubules***

The *Drosophila* genome project is essentially complete and effort is being made to annotate the 13,600 genes discovered to date. There may still be a number of novel genes that are not yet discovered in *Drosophila* genome. New technology developments have led to the discovery of novel genes and novel transfragments (transcribed fragments of the genome) in *Drosophila* as well as other organisms. These are novels genes and novel transfragments that have been found to be transcribed but have not yet been annotated. These novel genes and novel transfragments may play an important role in the *Drosophila* genome.

The project presented in this thesis draws on the wealth of physiological and genomic data available for the *Drosophila* tubule, and particularly on the extensive experience and knowledge gained within Dow/Davies lab from working with this tissue over a period of 30 years (Wang et al., 2004).

The primary aim of this project is to look for novel genes of *Drosophila* tubules. The project will start with comparing the three technologies, *Drosophila* expression microarrays, *Drosophila* tiling microarrays and the recently developed next-generation sequencing technology RNA-seq in order to evaluate which technology is best for profiling the *Drosophila* genome in terms of novel genes and novel splicing, particularly in relation to *Drosophila* tubules based on the measurement of expression level of the annotated *Drosophila* genes.

The reverse genetics technology will be applied to the tissue-specific novel genes (completely new transcribed units in tubules) which are generated by this work in order to search for the function. This will involve validation using reverse transcript polymerase chain reaction (RT-PCR) and transgenic resources such as the GAL4/UAS system. Data generated during the project will be deposited in FlyBase, FlyAtlas, Ensembl and other resources to benefit the fly research community.

Other new features which generated within this work, such as features which change coding sequences, regions transcribed from both strands, and novel splicing forms will be studied as well.

1.5.2 *Secondary aim/Application of standard array technology to investigate gene expression of different Drosophila species*

The secondary aim of the work presented in this thesis is to evaluate the use of genomic DNA-based probe selection method as means of improving the sensitivity for gene detection when applying the standard microarrays to heterologous species.

Standard Affymetrix three-prime expression microarrays will be applied to different *Drosophila* species that are related-to *Drosophila melanogaster* by varying distances. Both a closely related *Drosophila* species *D. simulans* and a medium distanced *Drosophila* species *D. pseudoobscura will be assessed* by applying both *D. simulans* and *D. pseudoobscura* genomic masks. In this way, a similar sequence of *Drosophila* probes will be chosen after assessment by genomic hybridization from the related species in order to evaluate the sensitivity of the genomic selection technology.

2. **Materials and Methods**

2.1 **Fly stocks**

The original stocks used in this study are presented in Table 2-1. Each fly stock is characterized in terms of fly line identifier, genotype, description, and source of origin for this project.

Table 2-1 Fly stocks used in this study.

Flies were either obtained from the stock centres or made in house. Each fly stock is characterized in terms of fly identification; genotype; description explaining what the fly stock is; reference, where it was obtained from (if it is so). The genotype of a chromosome is indicated only if there is a mutation or some other kind of variant on it. Chromosomes are listed in order: *X/Y*; *2; 3; 4*, where semi-colons separate the genotype symbols for each different chromosome*.* w^+*,* indicates the wild type allele of white gene on sex chromosome. w^- , indicates no allele of white gene on sex chromosome. A chromosomal genotype written on a single line indicates that the stock is homozygous for that genotype; heterzygosity is denoted by a two-line genotype. + indicates wild type. TM3, CYO indicates the balancer chromosome.

Fly ID	**Genotype**	**Description**	**Reference**
Canton S	*w*+; +; +; wild type	Wild type - *Drosophila melanogaster*	http://flybase.org/reports/FBst0000001.html
c42-GAL4	*w*-; +; *c42-GAL4*/*c42-GAL4*	GAL4 enhancer trap specific to the tubule principal cells.	(Sozen et al., 1997); Dow/Davies Lab
c724-GAL4	*w*-; +; *c724-GAL4*/*c724-GAL4*	GAL4 enhancer trap specific to the tubule stellate cells.	(Sozen et al., 1997); Dow/Davies Lab
Tubulin-GAL4 UAS Dicer/Tm3Sb	w-; +; *Tubulin-GAL4 UAS Dicer/TM3Sb*	Universal driver has GAL4 transcription factor	Dow/Davies Lab
Actin-GAL4- cyo	*w-; actin-GAL4/cyo;* +	Universal driver has GAL4 transcription factor	Dow/Davies Lab
Simulans	w^+; +;+	Wild type-*Drosophila simulans*	Dow/Davies Lab
Pseudoobscura	w^+;+;+	Wild type-*Drosophila pseudoobscura*	Steven Goodwins Lab
OregonR	w+; +;+ Wild type	Wild type - *Drosophila melanogaster*	Dow/Davies Lab
UAS-3L1a-RNAi	w^-; +; 3L1a	3L11a double stranded *RNA*	BestGene Jing Wang

Fly ID	Genotype	Description	Reference
	RNAi/Tm3	(*dsRNA*) fusion with upstream UAS	Dow/Davies Lab
UAS-3L3a-RNAi	w^{-} ; +; 3L3a RNAi/TM3,sb	3L13a double stranded *RNA* (*dsRNA*) fusion with upstream UAS	BestGene Jing Wang Dow/Davies Lab
UAS-3L4a-RNAi	w^{-}; 3L4aRNAi/cyo;+	3L14a double stranded *RNA* (*dsRNA*) fusion with upstream UAS	BestGene Jing Wang Dow/Davies Lab
UAS-3L5a-RNAi	w^{-}; +; 3L5a RNAi/TM3	3L15a double stranded *RNA* (*dsRNA*) fusion with upstream UAS	BestGene Jing Wang Dow/Davies Lab
UAS-3L6a-PTW	w^{-}; 3L6aPTW/cyo;+	3L6a PTW cDNA fusion with upstream UAS promoter	BestGene Jing Wang Dow/Davies Lab
UAS-3L7a-PTWV-YFP	w^{-}; 3L7aPTWV/cyo;+	3L7a PTWV cDNA fusion with upstream UAS promoter and downstream yellow fluorescent protein (YFP)	BestGene Jing Wang Dow/Davies Lab

2.2 Normal fly husbandry

Drosophila melanogaster (Canton S strain, RNAi and overexpressed flies before cross, c42 and c724 driver lines et al) flies were normally raised on standard medium (Appendix I) on a 12:12 h L:D cycle, at 23°C, and at 55% room humidity. The flies were transferred to new vials after two days cross and lay eggs. Adults' flies emerged in ten days normally; they were subsequently transferred to fresh vials on the day of emergence, and used seven days later.

2.3 Transgenesis fly husbandry

RNAi and overexpressed crossed flies were raised on standard medium on a 12:12 L:D cycle, in a 26°C SANYO incubator, and at 60-70% relative humidity. The flies were transferred to new vials every two days after flies crossed for three times. The GAL4/RNAi or GAL4/overexpressed flies merged in eight days after cross, then the new emerged flies were transferred to a fresh vials. The flies had transferred to fresh food every two days. The flies were used for seven days after they were emerged.

2.4 Tissue dissections

Flies were anaesthetised briefly by chilling on ice, then immediately dissected for tissues in *Drosophila* Schneider's medium (Invitrogen UK).

Table 2-2 Different tissue amount for RNA-seq, tiling microarrays, species arrays and cDNA.

Tissue	**Definition**	**Number per sample**	**Total amount of RNA obtained (ng)**
Adult Head	Severed at the neck. Includes brain, eyes, cuticle and some fat body.	100-150	2-9µg
Adult Tubule	Both anterior and posterior tubules with their common ureters, severed at the junction with the gut.	20 each separately total 20-80	2-9µg
Adult Testis	Testis excluding the accessory glands.	20 each separately total 50-60	2-9µg
Whole fly	Whole animal.	15-20	2-9µg

Equal number of males and females contributed to each RNA sample. Sufficient tissues were dissected in Schneider's medium to obtain 2-9 µg total RNA. As this involves significant pooling for such tiny tissues, tissues were collected immediately after each dissection into RLT buffer for RNA extraction (Section 2.5). This procedure was repeated 3 times for each tissue; that is, each RNA-seq sample corresponds to an independent biological replicate. For whole fly RNA extraction, at least 20 flies were used for each sample.

2.5 Total RNA extraction

2.5.1 *Extraction method*

RNA extraction was carried out in a nuclease-free environment using RNeasy Mini kit (for whole flies, heads), or RNAeasy Micro kit (for tubules, testis) according to the manufacturer's protocol (Qiagen UK). After the dissections and/or collection of whole flies in RLT buffer, homogenizations were performed manually using a small blue rod/pestle for whole flies and heads, or an ultrasonic cell disruptor (Misonix, Inc., USA) for tubules and testis. The tissues were immediately frozen at -80°C then processed later or processed immediately.

Then the homogenate was centrifuged for 3 min and supernatant was collected into a fresh 1.5 ml micro centrifuge tube. The rest of the protocol was according to Qiagen kit. An on column DNA digestion step (using Qiagen DNAase kit) was included for reducing genomic DNA contamination. RNA was eluted using 14-30 µl of nuclease-free water from the column and it was stored at -80°C until further use.

2.5.2 *Quantification and quality check*

Three quality controls were performed on the isolated RNA to check quantity, purity and integrity.

The quantity and purity of RNA were tested by Nanodrop (ND-1000 V3.7.1). 1µl RNA is required to load the machine. The optical density (OD) of the sample at 260nm was used to determine the concentration of the RNA in a solution. The ratio of OD of the sample measured at 260nm and 280nm was used to determine the purity of RNA. A 260/280 ratio in the range 1.8-2.0 inclusive indicated a good level of purity, a 260/280 ratio less than 1.8 or greater than 2.0 indicated the contamination of the RNA. A 260/230 ratio helped to determine the purity of RNA as well. The 260/230 ratio in the range 1.8-2.0 inclusive indicated the good purity of RNA, 260/230 less than 1.8 or greater than 2.0 indicated the contamination of the RNA.

The integrity of RNA was checked using Agilent 2100 Bioanalyzer. An example RNA profile for RNA extracted from a sample of *Drosophila* tubules is shown in Figure 2-1. The RNA integrity number (RIN) software algorithm allows classification of total RNA based on a number system from 1-10. Higher RIN indicates greater integrity. RIN values about 7-8 were considered to indicate an acceptable level of RNA purity for experiment work. RIN less than 7 was taken to indicate degradation of the total RNA. Microarray and mRNA-seq require the RIN of total RNA to be above 8.

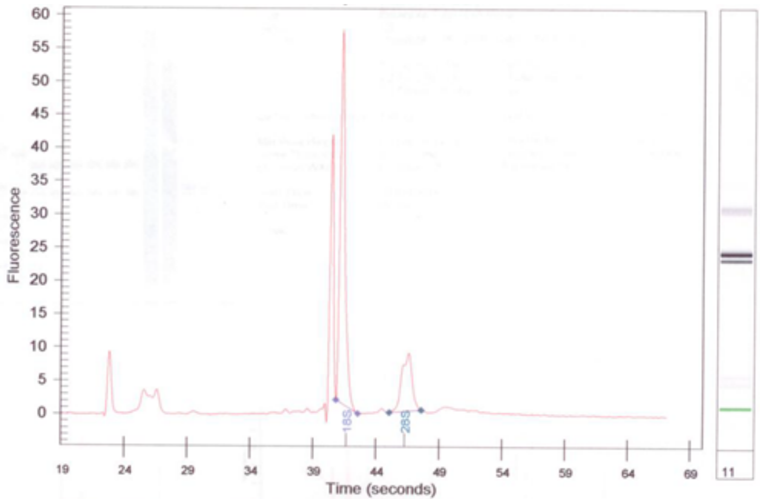


Figure 2-1 Good quality of *Drosophila* tubules Agilent profile.

Profile from Agilent Bioanalyzer indicating good quality for RNA extracted from *Drosophila* tubules. Note that the *Drosophila* RNA profile is different from human with two peaks at 18S and one peak at 28S.

2.6 **Genomic DNA extraction**

Two methods were used in genomic DNA extraction. The 30 fly genomic DNA protocol was used for generating a large amount of *Drosophila* genomic DNA. The DNeasy Blood and Tissue kit (Qiagen, Cat No. 69504) for high quality DNA extraction was used for genomic DNA array.

2.6.1 ***30 fly Genomic DNA extraction***

30 Canton S flies were put in a 1.5ml eppendorf tube containing 200µl buffer A (100mM Tris-Hcl pH 7.5, 100mM EDTA pH 8.0, 100mM Nacl, 0.5% SDS). The flies were grounded using a pestle. The grounded samples were incubated at 65^{0}C for 30 mins. Then 800µl buffer B (mix one part 5M potassium acetate with 2.5 parts 6M lithium chloride) was added, followed by incubation on ice for more than 10 mins. The sample was spun for 15mins at room temperature (RT), 13,000rpm. 1ml of the supernatant was transferred into a new tube to avoid crud. If crud was carried over, the transfer and spin step was repeated. 600µl isopropanol was added, mixed and spun for 15 mins at RT, 13,000rpm. The supernatant was removed, and the remaining pellet was washed with 200 µl 70% ethanol, and spun 5mins at 13,000rpm. The DNA pellet dried in 37^{0}C hybridization oven. The pellet was resuspended in 150 µl TE. The pellets that didn't dissolve were left

overnight in the fridge to dissolve and stored at -20^{0}C. The quality of genomic DNA was tested by running PCR using the same primers to span an intron using the genomic DNA and cDNA samples to compare the size.

2.6.2 *DNeasy Blood and Tissue kit*

DNeasy Blood and tissue kit (Qiagen, Cat No. 69504) was used to accelerate the high quality genomic DNA extraction for use with genomic DNA array.

10 flies (5 females and 5 males) around 8-10mg were put in 1.5 µl eppendorf tubes, and 180 µl Buffer ATL added and the tissue was homogenized by small pestle. 20µl proteinase K was added, mixed by vortexing, and incubated at 55^{0}C until the tissue was completely lysed, with vortexing occasionally during the incubation. Vortexed for 15s then 200 µl Buffer AL was added to the sample, mixed thoroughly by vortexing, and incubated at 70C for 10 min. 200 ul ethanol (96-100%) was added to the sample, and mixed thoroughly by votexing. The mixture was transferred by pipette to the DNeasy Mini spin column and placed in a 2ml collection tube, centrifuged at > 6000g (8000rpm). The flow-through and collection tube were discarded. Placed the column in a new collection tube, added 500µl Buffer AW1 (add 25ml ethanol), and centrifuged for 1 min at >6000g (8000rpm). The flow-through and collection tube was discarded. Placed the column in a new collection tube, added 500 ul Buffer AW2, and centrifuged for 3 min at 20, 000g (14,000rpm) to dry the column membrane. Discarded flow-through and collection tube. Placed the DNeasy Column in a clean 1.5ml eppendorf tube and pipette 200µl Buffer AE , incubated at room temperature for 1 min, and then centrifuged for 1 min at >6000g 98000rpm) to elute. The elution was repeated once in a new tube.

2.6.3 *DNA quality control*

DNA and RNA quantification was performed using a NanoDrop 1000 spectrophotometer (Thermo UK) in 1.5µl sample volume. The 260/280 ratio between 1.8 -2.0 indicates a good quality of DNA. A 260/230 ratio of greater than 1.8 indicates no buffer or ethanol contamination.

2.7 **Microarray**

2.7.1 ***Three-prime expression microarrays (Drosophila Genome 2.0 Array)***

Targets were prepared using the One-Cycle Eukaryotic Target labelling Assay protocol and kit (PN 900431) from Affymetrix were used in this experiment. 2ug RNA was used in line with the manufacture's protocol. Four replicate samples were used when preparing targets from tubules of *Drosophila* Simulans, *Drosophila* Pseudoodbscura and *Drosophila* melanogaster. The outline of the assay is shown in Figure 2-2. The RNA quality, quantities and integrity were checked by Nanodrop and Agilent bioanalyzer (chapter 2.5.2). 2ug total RNA, PolyA control and T7 oligo (dT) primer mixture were denatured at 70^0C for 10 mins to open the RNA second structure. First strand synthesis was performed using superscript II. Second strand synthesis was performed DNA ligase and DNA polymerase and T4 DNA polymerase was used to polish the end of dsDNA. The dsDNA was then cleaned up by the DNA clean up kit. The Biotin-labelled ribonucleotides and T7 RNA polymerase were used in '*In Vitro* Transcription' to make the biotin-labelled complementary RNA (cRNA). The quality of cRNA was then determined by Agilent 2100 Bioanalyzer.

Target Hybridization was performed by first fragmenting 10ug corrected cRNA. 200µl array target was made with control oligo B2 and 20x hybridization control. 130ul of the target was hybridized in *Drosophila* Genome 2.0 Array at 60rpm, 45^0C oven for 16-18 hours.

Fluidic Station Setup: Fluidic station 450 was set up by prime wash using the GeneChip Operating software (GCOS) in line with Affymetrix's instruction to operate the whole process. The samples and project information was entered and saved as an experiment file.

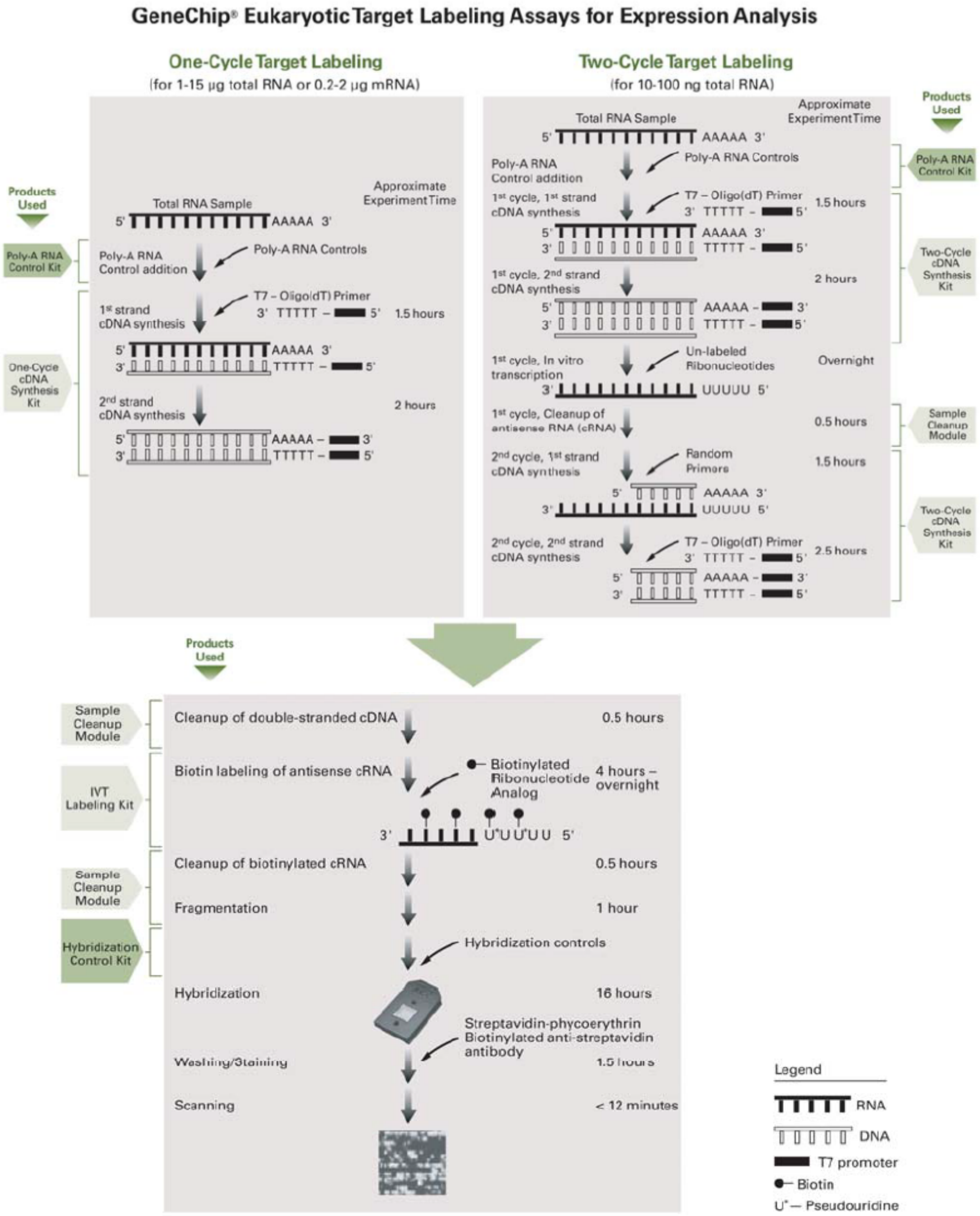

GeneChip® Eukaryotic Target Labeling Assays for Expression Analysis
One-Cycle Target Labeling
(for 1-15 µg total RNA or 0.2-2 µg mRNA)
Two-Cycle Target Labeling
(for 10-100 ng total RNA)
Products Used
Approximate Experiment Time
Total RNA Sample
5' AAAAA 3'
Poly-A RNA Control Kit
Poly-A RNA Control addition
Poly-A RNA Controls
One-Cycle cDNA Synthesis Kit
1st strand cDNA synthesis
T7 – Oligo(dT) Primer
3' TTTTT – 5'
1.5 hours
5' AAAAA 3'
3' TTTTT – 5'
2nd strand cDNA synthesis
2 hours
5' AAAAA – 3'
3' TTTTT – 5'
Products Used
Approximate Experiment Time
Total RNA Sample
5' AAAAA 3'
Poly-A RNA Control addition
Poly-A RNA Controls
Poly-A RNA Control Kit
1st cycle, 1st strand cDNA synthesis
T7 – Oligo(dT) Primer
3' TTTTT – 5'
1.5 hours
5' AAAAA 3'
3' TTTTT – 5'
Two-Cycle cDNA Synthesis Kit
1st cycle, 2nd strand cDNA synthesis
2 hours
5' AAAAA – 3'
3' TTTTT – 5'
1st cycle, In vitro transcription
Un-labeled Ribonucleotides
Overnight
3' UUUUU 5'
1st cycle, Cleanup of antisense RNA (cRNA)
0.5 hours
Sample Cleanup Module
2nd cycle, 1st strand cDNA synthesis
Random Primers
1.5 hours
5' AAAAA 3'
3' UUUUU 5'
Two-Cycle cDNA Synthesis Kit
2nd cycle, 2nd strand cDNA synthesis
T7 – Oligo(dT) Primer
3' TTTTT – 5'
2.5 hours
5' AAAAA – 3'
3' TTTTT – 5'
Products Used
Sample Cleanup Module
Cleanup of double-stranded cDNA
0.5 hours
IVT Labeling Kit
Biotin labeling of antisense cRNA
Biotinylated Ribonucleotide Analog
4 hours – overnight
3' U*UU*UU 5'
Sample Cleanup Module
Cleanup of biotinylated cRNA
0.5 hours
Fragmentation
1 hour
Hybridization Control Kit
Hybridization controls
Hybridization
16 hours
Streptavidin-phycoerythrin
Biotinylated anti-streptavidin antibody
Washing/Staining
1.5 hours
Scanning
< 12 minutes
Legend
RNA
DNA
T7 promoter
Biotin
U* — Pseudouridine

Figure 2-2 Overview of the GeneChip 3' IVT Express Labelling Assay

The process of 3'end expression microarrays includes one cycle (2µg RNA) and two cycles (100ng RNA) reverse transcription, *in vitro* transcription and labelling procedure. Picture is taken from www.affymetrix.com.

Probe Array Wash and Staining: SAPE, Antibody and SAPE stain were performed using fluidic station in line with Affymetrix's instruction using the Midi_euk2v3 protocol in this process.

Probe Array Scan: The array was scanned by Affymetrix Scanner 3000 7G. The cell files were created as in raw data format for further analysis by other software such as Partek.

2.7.2 *Drosophila tiling microarrays*

GeneChip Whole Transcript (WT) Double-Strand Target Assays protocol and kit (PN 900652) were used in this experiment. The outline of this procedure is shown in FIgure2-3. Four replicates of *Drosophila* whole flies, testes, heads and tubules were processed. The quality of total RNA of all the samples was checked on Nanodrop ND-1000 (Chapter 2.2.5.2) and Agilent 2100 Bioanalyzer (Chapter 2.2.5.2). The experiment was started with 7µg good quality total RNA. First strand cDNA synthesis used random primer and superscript II, second strand synthesis used DNA polymerase I, dUTP was incorporated in both strands for later recognition by the fragment enzyme. The double-stranded DNA was cleaned up by GeneChip Sample Clean up Module (PN 900371). The 7.5µg dsDNA was then fragmented by using enzyme UDG and APE to recognize the dUTP in the dsDNA. The fragmented dsDNA was labelled by TdT for the end labelling procedure (terminal labelling). The quality of the terminal labelling fragmented dsDNA can be checked by 'gel shift assay'. The hybridization target was made by the labelled dsDNA, control oligoB2, herring sperm DNA and BSA. The Fluidic Setup, Wash and Stain, Scan was performed as in Section 2.7.1. The Tiling chips in this experiment were GeneChip *Drosophila* Tiling 2.0R Array.

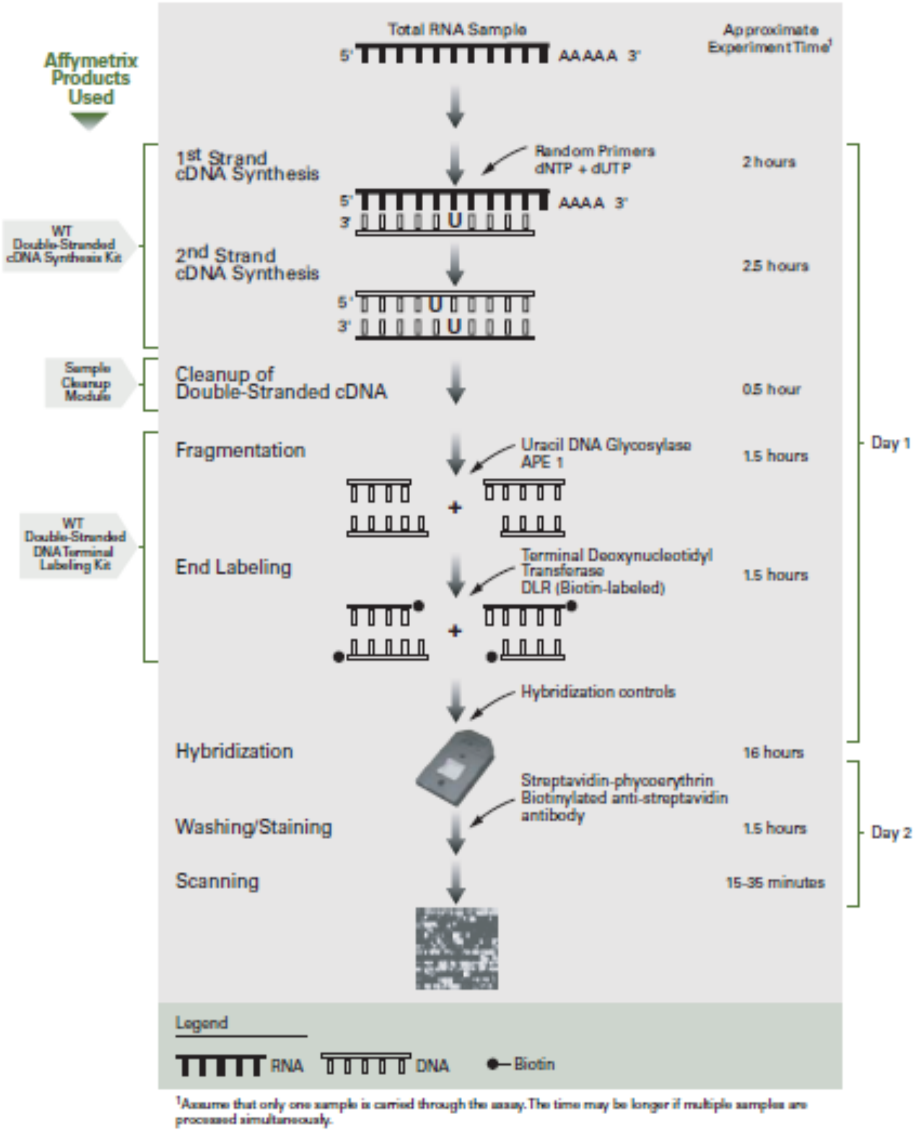

Affymetrix Products Used
Total RNA Sample
5' AAAAA 3'
Approximate Experiment Time[1]
1st Strand cDNA Synthesis
Random Primers
dNTP + dUTP
2 hours
5' AAAA 3'
3'
U
WT Double-Stranded cDNA Synthesis Kit
2nd Strand cDNA Synthesis
2.5 hours
5'
3'
Sample Cleanup Module
Cleanup of Double-Stranded cDNA
0.5 hour
Fragmentation
Uracil DNA Glycosylase
APE 1
1.5 hours
Day 1
+
WT Double-Stranded DNA Terminal Labeling Kit
End Labeling
Terminal Deoxynucleotidyl Transferase
DLR (Biotin-labeled)
1.5 hours
Hybridization controls
Hybridization
16 hours
Streptavidin-phycoerythrin
Biotinylated anti-streptavidin antibody
Washing/Staining
1.5 hours
Day 2
Scanning
15-35 minutes
Legend
RNA
DNA
Biotin
[1]Assume that only one sample is carried through the assay. The time may be longer if multiple samples are processed simultaneously.

Figure 2-3 GeneChip Whole Transcript Double-Stranded Target Assay Schematic

The process of tiling microarrays includes reverse transcription, fragmentation and end labelling producture. Picture is taken from www.affymetrix.com.

2.7.3 *Genomic DNA Array*

Invitrogen BioPrime® DNA Labelling System (Cat. No. 18094-011) was used to generate DNA target in this experiment. 500ng DNA was used; Random primers were annealed to the denatured DNA template and extended by Klenow fragment in the presence of biotin-14-dCTP to produce sensitive biotinylated-DNA probes. The entire labelled genomic DNA with Control oligo B2, 20x Hybridization control, BSA, Herring sperm to produce the hybridization cocktail. The hybridization, Fluidic Setup, Wash and Stain, Scan protocols were performed as in Section 2.7.1. The labelled DNA was hybridized to *Drosophila* Genome 2.0 Array.

2.7.4 *Microarray data analysis*

2.7.4.1 **Normalization method**

Three popular normalization methods applied to microarray are MAS5.0, RMA and GC-RMA. This can be referred to as low-level microarray analysis.

MAS 5.0 method. Li and Wong (2001) were the first to propose model-based expression measures. They observed a very strong probe effect in that *PM-MM* values, the need for non-linear normalization, and the advantages of using multi-array summaries for detection and removal of outliers (Li and Hung Wong, 2001; Li and Wong, 2001)(Li and Hung Wong, 2001; Li and Wong, 2001)(Li and Hung Wong, 2001; Li and Wong, 2001).

RMA (Robust Multiarray Average) normalization method. It is linear and performs the background correction, normalization across arrays, probe level intensity calculation and probe set summarization. The RMA method is notable for employing quantile normalization that forces the distributions of probe-level measurements to be equal across multiple arrays before median-polish probe-set summaries are calculated.

GC-RMA normalization method. A modification to RMA, GC-RMA performs the background correction by considering the GC contents. G/C in sequence leads to stronger hybridization because each G-C pair forms three hydrogen bonds whereas each A-T pair forms two. GC-RMA uses the mismatch data that RMA ignores to model the effects of GC-content on nonspecific binding.

The differential gene expression using statistical hypothesis testing methods including analysis of variance (ANOVA) can be referred to as high-level microarray analysis. The fundamental idea behind analysis of variance (ANOVA) is that, given an appropriate experimental design, variability in the quantity being measured (gene expression) can be partitioned into various identifiable sources. The assumed sources of variability will include the experimental factors, as well as random noise (Pavlidis, 2003).

2.8 mRNA Sequencing (mRNA-seq)

RNA sequencing experiments were carried out using the Illumina protocol and the mRNA-seq Sample Prep kit part (Part No1004814) in line with manufacturer's instructions. Samples were prepared using three replicates of tubules, three replicates of heads, three replicates of testes, three replicates of heads. Experiments were performed in two batches where the first batch was a pilot study including one sample of tubules, testes, whole flies and heads and the second batch included two replicates of tubules, testes, whole flies and heads.

2.8.1 *mRNA-seq samples library preparation*

Library preparation of mRNA-seq was carried out using 1-10ug total RNA (use 9ug for first batch, 2ug for second batch). The mRNA was purified and fragmented by using Sera-Mag oligo (dT) beads. The fragmented mRNA was precipitated by using 3M NaoAC, pH 5.2, Glycogen, 100% ETOH. Synthesis of first strand cDNA was performed using random primers and superscript II system. Second strand cDNA synthesis and purification of the dsDNA used the QIAquick PCR Purification Kit (Cat. No. 28104). Ends were repaired by using T4 DNA polymerase and klenow DNA polymerase. 'A' base to 3' ends was added. Ligation of adapters (including the sequences of primers, sequences of flow cell and sequences for PCR amplification) was performed and the ligation product purified by using Qiagen PCR purification kit. The cDNA templates were purified on a gel to select the 200bp (±25bp) range. PCR was used to enrich the purified cDNA templates. PCR was performed with two primers that anneal to the end of the adaptors. The library was validated on an Agilent Technologies 2100 Bioanalyzer using Agilent DNA-1000 chip (Figure 2-4). The size, purity and the concentration of the samples was decided. The final product should be at approximately 200bp.

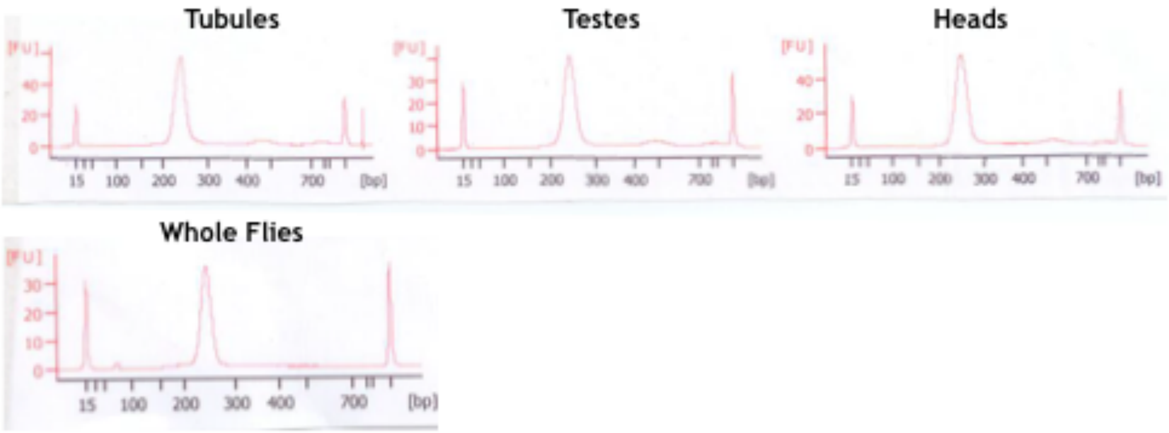


Figure 2-4 RNA-seq library run by Agilent Bioanalyzer 2100

The RNA-seq libraries indicated in Agilent Bioanalyzer 2100 around 220bp peak, and the concentration also can be determined by Agilent. Picture was run on Agilent DNA-1000 chip. The first and the third peak are marker peaks. The middle peak is the peak of RNA-seq library.

2.8.2 *Cluster generation on Illumina Cluster station*

The Single-read cluster generation kit v2 was used to generate the cluster on the flow cell. The flow cell contains adapters which are complementary to the adapters of the library samples. 1-4pM libraries samples hybridize to the lawn of primers on the flow cell; bound molecules are then extended by polymerases; double-stranded molecule is denatured; original template is washed away. Newly synthesized template is covalently attached to the flow cell surface. Single-strand flips over to form bridges. Hybridized primer is extended by polymerase. Bridges amplification cycle repeated till multiple bridges are formed. Double strand bridges are denatured; reverse strands are cleaved and washed away leaving a cluster with forward strand only. Free 3'-ends are blocked to prevent unwanted DNA priming. Sequencing primers are hybridized to adapters.

2.8.3 *Sequencing by synthesis on Illumina GAII*

After the cluster generation, the flow cell is then cleaned and put in the Illumina GAII Genome Analyser Reader. The Illumina sequencing kits are used in the sequencing process. 18 cycles and 36 cycle kits was for the first replicate of the four samples (a wholes flies, tubules, testes and heads). The total sequence length is 54bp. Two 36 cycles kits were used for the remaining samples which the sequence length is 72bp. The first base incorporation is performed on the

Illumina GAII Genome Analyzer, and then the quality was checked from the machine report. The report shows the clusters of the eight lanes of the flow cell. Validation was performed by checking Goodness of fit is greater than 0.9900 and the absolute value of the sensitivity is in the range 350-400. Sequencing commerces if the specification is met. To determine the sequence, four types of reversible terminator bases (RT-bases) are added and non-incorporated nucleotides are washed away. A camera takes images of the fluorescently labelled nucleotides, then the dye along with the terminal 3' blocker is chemically removed from the DNA, allowing the next cycle. The DNA chains are extended one nucleotide at a time and image acquisition can be performed at a delayed moment, allowing for very large arrays of DNA colonies to be captured by sequential images taken from a single camera.

2.8.4 *Integrated Data Analysis and Pipeline*

Integrated Data Analysis (IPA) and reporting was carried out by IPA instrument control software, which performs real-time reporting. IPA displays values for signal intensity, focus quality and cluster number. Images from each cycle of sequencing by synthesis are moved from the instrument control PC (IPA) into a run folder (pipeline) residing on a LINUX server. A series of programs in the pipeline perform image analysis, base calling, quality assessment and either sequence alignment and allele calling or tag alignment, binning and counting. Data analysis to the point of alignment (ChIP), allele identification (resequencing) or tag counts (gene expression and small RNA analysis) are provided with the system (see chapter 4 for details).

2.8.5 *Directional mRNA-Seq (Strand specific mRNA-seq)*

This experiment used the directional mRNA-seq Pre-release Library. Prep. Protocol v 1.0. The kit is from Illumina mRNA-seq library pre kit (RS-100-0801). library preparation started with 2ug of good quality total RNA (260/280 1.9-2.1, 260/230 1.8-2.0, Agilent Bioanalyzer RIN is greater than 8). Firstly, polyA selection of mRNA from total RNA was performed by using Sera-mag oligo (dT) beads, then mRNA were fragmented by fragmentation buffer and purified by Qiagen RNeasy MinElute clean up kit (Cat. No. 28004). Secondly, end repair was performed of RNA with phosphatase & PNK treatment and purified by Qiagen

RNeasy MinElute clean up kit. Thirdly, the 3' and 5' adaptors was ligated (the diluted v1.5 sRNA 3' Adaptor and the NEB supplied 10xT4 RNL2 truncated buffer), so that the adaptors have the desired sequences (including the sequences primers, sequences of flow cell and sequences for PCR amplification). Fourthly, RT-PCR amplification was performed, with the reverse transcription reaction using superscript II and SRA RT primer. The amplification PCR used GX1, GX2 primers that bind the sequences on the adapters. The purification of library uses Agencourt AMPure beads (Item No.A63880). Characterization of the library was determined by Agilent 2100 bioanalyzer. The peak size and range is indicative of the purity and quantity of the library. The normally size is between 200-250bp.The library was now ready for amplifying on the cluster station (section 2.8.2) and sequencing on GAII (section 2.6.3).

2.8.6 ***RNA-seq and Directional RNA-seq analysis methods***

A TopHat and Cufflinks analysis pipeline was the main analysis tool applied in this thesis. This is the only pipeline for RNA-seq analysis so far.

RNA-seq results from Illumina technology (GAIIx) were trimmed by in-house script then aligned by Bowtie; the unaligned reads were split and realigned by TopHat. All the alignment files were merged together by Cuffmerge, and the merged files were compared with the reference annotation by Cuffcompare to find the novel unannotated features. The tubule-, testes- and head-specific novel expression genes were found by Cuffdiff combined with the Cuffcompare results. The results can be visualized using CummeRbund.

2.9 **Complementary DNA (cDNA) synthesis**

cDNAs for PCR and qPCR were synthesised using 500 - 1000 ng of total RNA. Recombinant reverse transcriptase (SuperScript® II; Invitrogen UK) was used to reverse transcribe the RNA in a total of 20 µl of reaction volume. Firstly, Oligo (dT)12-18mers (500 µg/ml), 500 – 1000 ng total RNA, 1 µl dNTP (10 mM each of dCTP, dGTP, dATP and dTTP) and nuclease free water (Ambion, Cat.No.9932) to make up to 12 µl total volume were assembled in a PCR tube. This mixture was heated to 65°C for 5 min and chilled for 2 min on ice (PCR machine).

The contents were collected by brief centrifugation and mixed with 4µl of 5x first-strand buffer, 2 µl of 0.1 M DTT and 1 µl of RNAaseOUT® (40 units/µl; Invitrogen). The reaction mixture was then incubated for 2 min at 42°C. After the incubation, 1 µl (200 units) of SuperScript® II RT was added and mixed by pipetting gently up and down. Then mixture was incubated at 42°C for further 50 min in a Peltier thermal cycler (MJ Research). Finally, the reaction was terminated by heating at 70°C for 15 min and centrifuged briefly to collect the cDNA contents at the bottom. The cDNA was stored on ice for subsequently use or stored in -20°C. The cDNA was used for normal PCR, qPCR or *Pfu*-based PCR as described in the following sections.

2.10 **Oligonucleotide (primer) synthesis**

2.10.1 ***Standard PCR primer design***

Oligonucleotide primers were designed using MacVector 11.1.1 (MacVector, Inc. UK) or other web resources (Primer3, NCBI; OligoPerfect ™ designer, Invitrogen UK; SnapDragon - dsRNA Design). The sequences were sent to the MWG Biotech custom primer service for synthesis on a 0.01 µmol scale. Primer stock concentrations at 100 µM were obtained for each primer by resuspending the lyophilised powder in ddH_2O and a working concentration of 6.6µM was prepared from the stocks. Primers were stored at -20 ºC until further use.

2.10.2 ***Taqman qPCR primer design***

qPCR primers and probes were designed using Integrated DNA Technologies (IDT) PrimerQuest primer and probe design tool. Primers and probes sequences were sent to IDT customer primer service for synthesis. The qPCR Assay tubes (primers and probes) were centrifuged at 750g for 10 sec, then resuspended to 20X stock by adding 500 ul IDTE buffer (10mM Tris, 0.1mM EDTA, pH8.0). The final 1X concentration of 500 nM primers, 250 nM probe and 5 ng cDNA will be used in the assays. Taqman primers are designed on two near exons and probes are across the intron-exon boundary. TaqMan® Gene Expression Assays consist of a pair of unlabelled PCR primers and a TaqMan® probe with a fluorescent reporter or fluorophore such as 6-carboxyfluorescein FAM™ or VIC® dye label on the 5' end, and minor groove binder (MGB) nonfluorescent quencher (NFQ) on the 3' end.

2.11 **Polymerase chain reaction (PCR)**

2.11.1 ***Standard PCR***

Standard PCR protocols were used for most amplifications of DNA. Amount of template DNA varied, however, 0.5 µg of genomic DNA or 0.1 µg of plasmid DNA were typically used per reaction. Standard PCR (Table 2-3) was performed using pre-aliquoted ready-to-use master mix in a Peltier Thermal Cycler (MJ Research).

Table 2-3 Typical cyclic conditions for PCR.

Step	Number of cycles	Temperature	Duration	Principle
Initial denaturation	1	95°C	5 - 10 min	To denature secondary structures
Denaturation	30	95°C	30s	To denature the end products of each PCR cycle
Annealing		55 - 62°C	30s	Temperature is set depending on the melting temperature of the primers used; typically ~5°C lower than Tm
Extension		72°C	30s - 5 min	Extension time is set at the rate of 20 base pairs/sec
Final extension	1	72°C	10 min	For the final extension of incomplete ssDNA

This mix uses the Taq-Polymerase (modified) which has 5’ to 3’ polymerization and exonuclease activity but lacks 3’ to 5’ exonuclease (proofreading) activity. The master mix includes Thermoprime plus DNA polymerase (1.24 U) (Thermo UK), Tris-HCl (75 Mm; pH 8.8 at 25°C), $(NH_4)_2$ SO_4 (20 Mm), $MgCl_2$ (1.5 Mm), Tween 20 (0.01%), dNTPs (0.2 mM each). PCR reactions were normally performed in a total of 25 µl (1µl cDNA, 1µl forward primer, 1µl reverse primer and 23µl master mix) volume. The cycling parameters used are presented in the table below. PCR products were separated by agarose gel electrophoresis described in section 2.13.

2.11.2 *Pfu-based Herculase II Fusion polymerase PCR*

Pfu-based Herculase II fusion polymerase (Agilent UK) was used for amplifying longer products. It has a high affinity double-stranded DNA binding domain that enhances the processivity and increases the yield. The PCR products were used for pENTR clone, in situ hybridization. The protocol used is presented in the Table 2-4 below.

Table 2-4 *Pfu*-based Herculase II fusion polymerase PCR reaction mix and protocol.

Parameter	Targets <1 kb	Targets 1 - 10 kb	cDNA Targets
Input template DNA	100 - 300 ng genomic DNA or 1 - 30 ng vector DNA	100 - 400 ng genomic DNA or 1 - 30 ng vector DNA	1 - 2 µl cDNA from RT-PCR reaction
Herculase II polymerase	0.5 µl	1 µl	1 µl
DMSO	0 - 8% final concentration	0 - 8% final concentration	0 - 8% final concentration
Primers (each)	0.25 µm	0.25 µm	0.25 µm
dNTPs	250 µm each dNTP	250 µm each dNTP	400 µm each dNTP
Extension time	30s	30s per kb	60s per kb
Denaturing temperature	95°C	95°C	95°C
Extension temperature	72°C	72°C	68°C

2.12 **Quantitative reverse-transcriptase PCR (qPCR)**

qPCR was performed using TaqmanGene Expression Master Mix (Part No. 4374657) and the primers and probes mixture from IDT (see section 2.10 qPCR primer designs).

qPCR was performed using cDNA as the starting material. The cDNA was diluted 5ng/ul assuming 1ng RNA reverse transcript to 1 ng cDNA. The mixture for qPCR includes 10 µl Taqman master mix, 1 µl 20x Taqman probe, 8ul water and 1 µl cDNA (5 ng/µl) making the total volume 20 µl. The negative controls (without Superscript II) and/or a blank (without cDNA) were maintained to monitor and subtract the genomic DNA contamination and background fluorescence respectively. The standard curve was generated using target gene primers and the reference control primers by adding a serial dilution of fly cDNA. PCR conditions commonly used included denaturation of the primers at 50°C for 2mins, 95 ºC 10mins, followed by 95 ºC for 15mins and 60 ºC for 1mins for total number of PCR cycles used of 40. The amplification curve shows the amplified PCR product, the Taqman probe has fluorophore on 5'-end and a quencher on 3'-end. When the forward and reverse primers are extended by Taqman DNA polymerase and degraded of the probe releasing the fluorophore, then fluorescence is detected.

The expression data were further analysed. The *alpha-Tubulin84B* used as a reference controls (being house-keeping genes) to normalize the data. The fold change data was obtained using relative standard analysis by calculating the ratio of the two compared samples' CT values using the $2^{-\Delta\Delta CT}$ method (van Iterson et al., 2009). The standard error means (SEM) and *P*-values for statistical significance were calculated using GraphPad Prism statistics software (GraphPad version 5 Software, USA).

2.13 **Agarose gel electrophoresis**

PCR products or DNA were run on 1% agarose gel to assess quality and specificity. Gel was casted using 0.5x TBE [90 mM Tris, 90 mM boric acid (pH 8.3), 2 mM EDTA], containing 0.1 µg/ml EtBr as described in (Joseph Sambrook, 2001). 6x loading dye [0.25% (w/v) bromophenol blue, 0.25% (w/v) xylene cyanol, 30% (v/v) glycerol in water] was added to samples, using 0.5% TBE as the electrophoresis buffer and a 1 kb ladder (Invitrogen) to a final concentration of 1x 5 µl (500 ng) of ladder and 10-20 µl samples were loaded into the wells.

Typically these were run at 100 V; the dye front was followed for electrophoresis termination and the DNA was visualised using high performance ultraviolet transilluminator (UVP UK) and compared against the ladder band size. Where needed, PCR products were extracted from gel as described in section 2.14, and quantified using NanoDrop as described in section 2.6.3.

2.14 **PCR/Gel purification**

PCR products and the products excised from the gels were purified using Qiagen PCR/Gel purification kit (Part No. 28704) according to manufacturer's guidelines. DNA was eluted in 20-30 µl of nuclease-free water. Purified PCR products were quantified on NanoDrop and are suitable for molecular cloning.

2.15 **Molecular cloning**

2.15.1 ***Plasmid vectors***

Table 2-5 Plasmids used in this project

pCR®2.1-TOPO®	For cloning poly-adenylated PCR products according to the TOPO TA cloning kit protocol (Invitrogen).
pTW	Gateway cloning destination vector for recombining entry clones (in pENTR) to generate final clones for germline transformation of cDNA of interest under the control of the upstream UAS in the UAS/GAL4 binary induction of transgenes *in vivo*.
pTWV	Gateway cloning destination vector for recombining entry clones (in pENTR) to generate final clones for germline transformation of cDNA of interest under the control of the upstream UAS. It also incorporates a C-terminal yellow fluorescent protein (YFP) sequence for fluorescent tagging.
pRISE	Gateway cloning destination vector for recombining entry clones (in pENTR) to generate final clones for germline transformation of dsRNA of interest under the control of the upstream UAS (Kondo et al., 2006).

2.15.2 ***Normal cloning procedure***

PCR products were directly cloned using the Invitrogen TOPO® cloning kit into appropriate TOPO® vectors according to the manufacturer's instructions and transformed into E.coli TOP10 cells or DH5α (see 2.15.4). 100 µl of the transformed cells was then spread onto L-agar plates containing 100 µg/ml ampicillin or the antibiotic appropriate to the resistant marker of the plasmid, and incubated overnight at 37°C. TOPO TA cloning kit required the PCR products to contain 'A' overhangs.

The transformants were removed as single colonies and grown overnight (with shaking) at 37°C in 5 ml or 100 ml L-broth (Appendix IV) using appropriate antibiotic for selecting the clones.

2.15.3 *Gateway® cloning*

The Gateway® cloning system (Invitrogen), which uses a homologous recombination technique, was used to clone the cDNA or dsRNA amplicons for germline transformation of *Drosophila* embryos for GAL4/UAS system induction of transgene expression *in vivo* in the flies. The system uses entry (pENTR) and destination vectors (*P*-element containing germline transformation vectors).

2.15.3.1 **Primer design and PCR amplification**

For Gateway® entry cloning, a forward primer was designed to contain a CACC sequence on 5' end for directional cloning into the entry vector: pENTR (Part No. 45-0218) for RNAi vector pRISE. However, the sequence for overexpression destination vector PTW, the primers design sequence for entry cloning was used the longest ORF sequence with taa in the end but for PTWV which had YFP in the end, the primers design sequence was used without taa. PCR amplifications were performed using Herculase® fusion polymerase according to the protocol in section 2.11.2 and the PCR product was purified by QIAquick PCR purification kit (Cat. No. 28104).

2.15.3.2 **Entry clones**

Entry clones were made using the pENTR vector according to the manufacturer's instructions (Invitrogen). pENTR/D-TOPO® vectors take advantage of fast, efficient Directional TOPO® cloning that delivers the insert in the correct orientation for expression. These vectors contain the necessary attL sequences for recombination into any Destination vector.

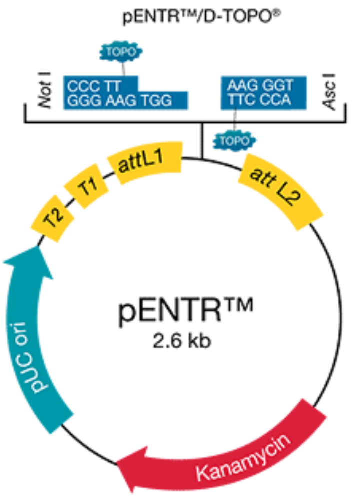


Figure 2-5 pENTR™ vector for Directional TOPO® cloning

This vector is used in this project and direct access to the multitude of Gateway® expression vectors. The attL sequence is important to recombination into destination vector.

2.15.3.3 Destination vectors

Destination vectors used include pRISE for RNAi, pTW for normal overexpressor and pTWV for tagged overexpressor constructs.

2.15.3.4 Gateway® recombination using LR Clonase

Gateway® recombination of entry and destination clones was performed using LR clonase enzyme mix according to the manufacturer's protocol (Invitrogen). Essentially, the enzyme catalyses the *in vitro* homologous recombination between an entry clone (pENTR-attL-GENE OF INTEREST-attL) and a destination vector (containing attR sites) to generate an expression clone of interest.

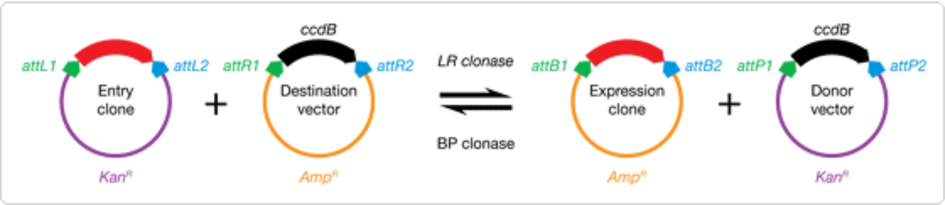


Figure 2-6 Generate an expression clone of interest.

LR Clonase enzymes that catalyze the in vitro recombination between an Entry clone (containing a gene of interest flanked by attL sites) and a Destination vector (containing attR sites) to generate expression clone.

2.15.4 *Transformation of E. Coli*

Competent *E. Coli* cells were transformed with the construct of interest according to the manufacturer's protocol. Briefly, cells were thawed on ice and the plasmid vector, PCR products, salt solution were mixed and incubated for 5 min on ice. The plasmid mixture was then transferred into the cells. The positive clones were identified using the antibiotic resistance markers of the clones generated.

Table 2-6 Competent bacteria strains used in this project

Strain	Genotype	Use
TOP10 competent cells (Invitrogen)	(F- mcrA, D(mrr-hsdRMS-mcrBC), f80lacZ DM15, DlacX74, recA1, deoR, araD139, D(ara-leu)7697,galU, galK, rpsL, (StrR), endA1,nupG).	For plasmid transformation and propagation of TOPO-related clones
DH5α subcloning efficiency competent cells (Invitrogen)	(F- f80dlacZ DM15, D (lacZYA-argF), U169, deoR, recA1, endA1, hsdR17 (rK-, mK+), phoA, supE44, l-, thi-1, gyrA96, relA1).	For normal plasmid transformation and propagation

2.15.5 *Purification of plasmid DNA*

Purification of plasmid DNA was performed using Qiagen mini (Cat. No.12125) or maxi kits (Cat. No.12165) (Qiagen UK). The overnight grown cultures were spun down to pellet the cells. The cells were lysed in the lysis buffer and DNA was either column eluted in 30 µl of water (for minipreps) or resuspended in 500 µl of water (for maxipreps).

2.15.6 *Validation of cloning products*

The cloning products obtained using different cloning procedures were validated for sequence, direction and length using PCR, restriction enzyme digestion and/or sequencing.

2.15.6.1 **PCR**

For PCR validation, the clones were amplified using the combination of primers with one from the transgene and the other from the vector. This allows confirmation at whether the transgene is inserted in the right direction and has the full length transgene. However, this approach was only employed for transgenes less than 2000 bp as the increase in length causes the cycling conditions to vary greatly.

2.15.6.2 **Restriction enzyme digestion**

Restriction enzyme digestion was employed to confirm whether the transgenes were inserted in the right direction and if they were right size.

2.15.6.3 **Sequencing**

Before they were sent off to be microinjected, DNA sequencing was performed on the constructs to check for any possible errors in the proof reading of the polymerase.

2.15.7 *Normal cDNA constructs*

pTW destination vector was used to recombine the pENTR entry clones for the normal overexpression constructs listed in Table 2-5.

2.15.8 *YFP fusion cDNA constructs*

pTWV destination vector with a C-terminal YFP tag was used to recombine the pENTR entry clones for the tagged constructs listed in Table 2-5.

2.15.9 *Double-stranded RNA constructs*

pRISE vector was used for making double-stranded RNA constructs for transgenic RNAi flies for GAL4/UAS system induction of RNAi *in vivo*. Gateway recombination system was used for RNAi constructs where pRISE (Kondo et al., 2006) is used as donor and pENTR D TOPO® as an entry vector. Three RNAi constructs were made for the novel gene 3L (23777335-23780626), 3L-1b, 3L-2b, 3L-4a. These were sent for *Drosophila* embryo germ line transformation to BestGene Company (USA).

2.15.10 *Dual promoter constructs for in situ hybridization*

RNA probe constructs for *in situ* hybridization were made using the PCR II TOPO vector. This vector has nucleotide 'T' overhangs and dual promoters either side of the multiple cloning sites. Overhangs of nucleotide 'A' were added to the PCR products obtained from *pfu* PCR that were to be used for *in situ* hybridization.

2.16 *Drosophila* S2 cell culture

2.16.1 *Passaging*

Drosophila S2 cells (Invitrogen) were maintained in complete Schneider's medium or CSM [Schneider's medium supplemented with 10 % fetal calf serum (FCS)] at 28°C. 15 ml of cells were kept in T75 flasks. Cells were passaged when their density reached 10^7 cells/ml. The weakly adherent S2 cells were resuspended gently by pipetting and then diluted by adding 6 ml of cells into 9 ml of CSM in a fresh flask.

2.16.2 *Transient transfection (Insect GeneJuice Protocol)*

Transient transfection was carried out in tissue culture six-well plates. The Insect GeneJuice Transfection Reagent (Part No. 71259) and protocol were used for the transfection (Novagen). Exponentially growing cells were spun and resuspended in Schneider's only. The cells were counted under Microscope. 1 million cell volumes were taken and made up to 500 µl. The cells were allowed in the Shneider's plate for 1 hour in the incubator. Plasmid DNA was prepared using a maxi-prep kit (Qiagen) and eluted in TE buffer. For each transfection, 2-3 µg of Plasmid DNA was added with 80 µl Schneider.12-15 µl of insect gene juice was added in 80ul Scheider medium. The diluted DNA was slowly added *dropwise* to the diluted Insect GeneJuice Transfection Reagent, then mix immediately by gently vortexing and incubated at room temperature for 15 mins. 640ul Schneider's medium was added to Insect juice/DNA transfection mixture. The cells were incubated for 4 hours at 28^0C. After 4 hours the transfection medium was removed and replaced with 2.5ml CSM.

If a plasmid encoding a metal inducible promoter was used, 20 µl $CuSO_4$ was added to the cells, mixed by shaking to induce expression and expression was allowed to proceed for 48-72 hours at 28^{0}C. Cells were then harvested by centrifugation at 1,500 g for 1 min at RT, washed once in PBS, pelleted and either frozen at -70ºC before use or used immediately.

2.17 **Protein analysis**

2.17.1 ***Extraction***

Table 2-7 Protein lysis buffer (RIPA) components

Component	Volume	Manufacturer	Catalog no.
100mM Tris-Cl (pH 7.4), 300mM NaCl	5 ml		
10% Triton®X100	1 ml	Sigma	T8787
10% Na deoxycholate	1 ml	Sigma	D6750
200mM PMSF (in isopropranol)	50 ul	Sigma	P7626
10% SDS	100 ul	Sigma Aldrich	L45090
Pierce® protease and phosphatase inhibitor	1 tablet	Thermo	NB 167568
0.01 M EDTA (pH 7.4)	100 ul	Sigma Aldrich	27285
H2O	2.75 ml		

Total protein was extracted from 15 whole flies. Flies were homogenized in 100 ul RIPA lysis buffer (see Table 2-7) using a hand-held pestle and then an ultrasonic cell disrupter, until the sample appeared homogeneous. The homogeneous protein lysate was kept on ice for 10 minutes and then clarified by centrifugation at 13000rpm for 5 minutes at room temperature. The supernatant was transferred into a fresh eppendorf tube and pellet was discarded.

2.17.2 ***Protein quantification***

The Bradford assay kit (BIO-RAD, 500-0006) was used for total protein quantification. Assay was carried out in a 96-well microtiter plate. Six BSA standards of 0-5 µg in water and 5 µl of protein supernatant were set up in triplicate in a final volume of 50 µl respectively. 200 µl of diluted Bradford dye reagent with ratio of 1:5 in distilled water was added to each well and mixed by pipetting. The plate was then incubated at room temperature for 5 minutes. The absorbance at 590 nm was read using a plate reader, and each standard absorbance was plotted against the known concentration to interpolate the unknown protein sample absorbance to calculate their quantities.

2.17.3 *SDS-PAGE separation*

Protein separation was performed using Novex NuPAGE™ electrophoresis system. Protein samples (20 ug) were prepared by adding 4X SDS-PAGE loading buffer to final volume of 28 ul. Samples were briefly vortexed and then spun down. Samples and protein marker (SeeBlue® Plus2 prestained standard, LC5925, and Life technology) were heated at 95^{o}C for 5 minutes. Protein marker and samples were then loaded on 10 wells 4-12 % Bis-Tris NuPAGE® Gel (Invitrogen, NP0321). The gels were then run in 1x NuPAGE™ MOPS SDS running buffer (diluted from NuPAGE™ MOPS SDS buffer (20x), NP0001, Invitrogen) at 200 V constant for 55 minutes (Appendix V).

2.17.4 *Western blotting*

Proteins separated on NuPAGE® Gel were then transferred onto Hyband transfer membrane (catalog no. RPN203B, Amersham) using Xcell II blot module at 30 V constant for one hour. After transfer, Hyband membrane was blocked in blocking solution (1 g non-fat dry milk in 20 ml 0.1% PBST) at room temperature for one hour. The blocked membrane was washed three times in 0.1% PBST buffer with 10 minutes for each time. Primary antibody (anti-GFP, mouse monoclonal, ZYMED) was diluted in blocking solution with ratio of 1:1000. Membrane was then incubated in primary antibody at 4^{o}C overnight. Membrane were washed in 0.1% PBST three times and then incubated in secondary antibody (goat anti mouse IgG-HRP, SC-2031, Santa Cruz Biotechnology) diluted in blocking solution with ratio of 1:5000 at room temperature for one hour. The blot was washed well for at least three times before detection. Blot was detected by incubation in detection solution (Immobilon^{TM} Western, Catalogue no. WBKLS0100, Millipore) for one minute and then developed in XOMAT film processor with varying exposure time (15 seconds to 2 minutes).

Table 2-8 Antibodies used for western blotting and immunocytochemistry.

Antibody and Source	Dilution and Use
Anti-GFP (mouse monoclonal, ZYMED)	1:1000 (Western and ICC)
Fluorescent Goat anti-mouse- IgG-FITC (goat polyclonal, Molecular Probes)	1:500 (ICC)
goat anti mouse IgG-HRP, SC-2031, Santa Cruz Biotechnology	1:5000 (Western)

2.18 **Immunocytochemistry (ICC)**

Immunocytochemical staining of cells and tissues as described in the following sections was performed for *in vitro* and *in vivo* localisation studies.

2.18.1 ***ICC of S2 cells***

S2 cells were resuspended and collected into 15 ml falcon tubes from the tissue culture flasks. These were spun at 3000 g in a free rotating bench top centrifuge, and supernatant was removed and cells were washed with PBS two times. About 100 µl of cells at a density of 6x10^6 cells/ml were plated and left for 15 min to allow the cells to settle and adhere. Excess solution was removed and the samples washed three times with PBS (Appendix II). Samples were then fixed by the addition of 4 % (w/v) paraformaldehyde in PBS for 15 min at RT. Samples were then washed 3 times with PBS, and blocked in PBS, 0.2 % (w/v) BSA, 0.1 % Triton X-100 for 10 min at RT.

Samples were then incubated overnight at RT in a humidified box with primary antibody diluted to the desired concentration in PBS/BSA/Triton X-100. Samples were then washed 3 times with PBS and incubated for 1 h at RT with the appropriate secondary antibody, diluted to the desired concentration in PBS/BSA/Triton X-100. Samples were then washed 3 times in PBS and, if required, DAPI stained as described in section 2.18.2. The coverslips to which samples were attached were then mounted on slides using VectaShield mounting medium (Vector Laboratories UK) and sealed with glycerol-gelatin. Samples were imaged by a confocal microscope system, as described in section 2.20.2.

2.18.2 *ICC of intact Drosophila tissues*

Intact tissues were dissected carefully in Schneider's medium (Appendix III) and transferred into a 1.5 ml tube containing PBS (pH 7.4). Then the tissues were washed with PBS 2 more times and the PBS was carefully removed. Tissues were then fixed in 4 % (w/v) paraformaldehyde in PBS at RT for 10-30 min. The tissues were washed three times in PBS and permeabilised using PBS, 0.3 % (v/v) Triton X-100 (PBT) for 30 min. This was followed by incubation with PBT with 10 % (v/v) goat serum (Sigma) (PBT-GS) for 3 h at RT.

Primary antibody, diluted to the desired concentration in PBT-GS, was then applied and the tubes incubated in a humidified box overnight at 4ºC.

The following day the tubules were washed in PBT 5 x 30 min and incubated in PBT-GS (Sigma) for 3-4 h. Secondary antibody, diluted to the desired concentration in PBT-GS, was then applied and the tubes were incubated in a dark humidified box overnight at 4ºC.

The tissues were then washed with PBT 3 x 1 h and in PBS 3 x 5 min. Then the nuclei were stained using 500 ng/ml DAPI for 2 min in PBS, diluted from a 10 mg/ml (in H_2O) stock solution. Tissues were washed three times with PBS before mounting. They were mounted in Vectashield mounting medium on confocal microscopy slides (BDH UK) or plates (Matek corporation USA). For slides, a coverslip was used and sealed with glycerol/gelatin (Sigma UK. The samples were viewed using a fluorescent microscope and a confocal microscopy system (see section 2.20.1, 2.20.2).

2.19 **mRNA *in situ* hybridization**

The *in situ* protocol was adapted from those described by (Allan et al., 2005) and the Berkeley *Drosophila* Genome Project (BDGP) 96-well *in situ* protocol (http://www.fruitfly.org/about/methods/RNAinsitu.html). The primers used for *pfu*-PCR in the above-described method were used to generate in situ probes. Two pairs of primers were used for generating two probes for the same gene.

2.19.1 *Cloning of template DNA*

The sequences of all PCR products were analysed using National Centre for Biotechnology Information (NCBI) and BDGP databases with Basic Local Alignment Search Tool (BLAST) to check the cross-hybridization potential of the sequences were found no significant matches with other *Drosophila melanogaster* sequences in the database.

PCR products were added 'A' overhangs, then cloned into the dual promoter PCR II TOPO vector (Invitrogen UK), and the orientation of the insert was checked using colony PCR with the combination of either M13 forward or reverse with a forward gene specific primer.

2.19.2 *Labelling and amplifying of RNA with digoxigenin*

Two different restriction enzymes were used to generate fragments with two different promoters (Sp6, T7) and the size checked by the gel (chapter 2. 2.13). Two types of DIG-labelled RNA *in situ* probes (sense and anti-sense) were generated by *in vitro* transcription by using DIG RNA labelling kit (DIG RNA labelling Kit SP6/T7 PN 11175025910 Roche). The sense probes were used as negative controls. The sense and anti-sense RNA were cleaned up by RNeasy column (Qiagen UK).

2.19.3 *In situ hybridization*

Adult tubules were dissected in Schneider's medium (Invitrogen UK) and placed into 1.5ml tubes with 100 μl of Schneider's medium. Samples also included no probe control and no antibody control. Schneider's medium was removed by the 20ul pipette. Postfix solution [10 mM potassium phosphate buffer (pH 7.0) containing 140 mM NaCl, 0.1% (v/v) Tween-20, and 5% (v/v) formaldehyde] was added for 20 min, followed by three washes with PBT [10 mM potassium phosphate buffer (pH 7.0) containing 140 mM NaCl and 0.1% (v/v) Tween-20]. The tissues were incubated with proteinase K in PBT (4 μg/ml) for 3 min at RT. The reaction was stopped with two washes of PBT containing 2 mg/ml glycine. The samples were washed twice with PBT before incubation with postfix for a further 20 min at RT.

The tissues were washed with five changes of PBT, followed by one wash with 50% hybridization buffer [5x SSC containing 50% (v/v) formamide, 10 mM KH_2PO_4, 140 mM NaCl, 1 mg/ml glycogen, 0.2 mg/ml sheared salmon sperm DNA, and 0.1% Tween-20 (pH 7.0)] plus 50% (v/v) PBT. The samples were washed once with hybridization buffer before a 1 h pre-incubation with hybridization buffer at 55°C and subsequently incubated for 43 h at 55°C with 100 μl of hybridization buffer containing 200-300 ng of either the sense or anti-sense riboprobe, taking care to seal the tubes with parafilm to prevent evaporation.

After hybridization, the samples were washed four times with hybridization buffer at 55°C, followed by a final wash overnight with hybridization buffer at 55°C with rotating. Samples were washed once with 50% (v/v) hybridization buffer and 50% (v/v) PBT, followed by four washes with PBT, and then incubated overnight at RT with 100 μl of pre-absorbed alkaline phosphatase-conjugated anti-digoxigenin Fab fragment (Roche UK) diluted 1:2,000 with PBT with shaking. The unbound antibody was removed with extensive washing in PBT, at least 10 times for 5-10 min. The samples were incubated with DIG detection buffer (100 mM Tris-HCl, pH 9.5, 100 mM NaCl, 50 mM $MgCl_2$) for 5 min and then repeated again.

The colour reaction was initiated by the addition of DIG detection buffer 5-bromo-4-chloro-3-indolyl phosphate (BCIP) and nitro blue tetrazolium (NBT) (BCIP/NBT Liquid Substrate System B1911-100ml, SIGMA) and left for 10 min to 2 h at RT. Development was stopped with extensive washing with PBT containing 50 mM EDTA, and the tissues were removed from the wells and mounted on slides with 70% glycerol and viewed with the Axiocam imaging system (Carl Zeiss UK).

2.20 **Imaging**

2.20.1 ***AxioVision fluorescent microscope***

Fluorescent imaging was carried out using AxioVision fluorescent microscope (Carl Zeiss imaging AxioVision 40 V4.6.3.0) for imaging the tissue or selecting the slide for further confocal microscopy. The defaut DAPI and GFP channel defined for AxioVision multi-channel microscope settings were chosen for visualisation of samples as appropriate. The microscope was adjusted to the required settings for DAPI or GFP, then the images were taken and the merged images were created. Control images were taken by using the same settings.

2.20.2 *Confocal Microscope*

Fluorescent imaging was carried out using the confocal microscope system LSM510 Meta from Zeiss Technologies UK. A HeNe1 543nm laser and a 561-625 band pass filter were used for imaging the Alexafluor® 568 secondary antibody. An Argon 488 laser and a 505-530 band pass filter were used for imaging the FITC antibody or fluorescent proteins. For visualization of DAPI, a pseudo-DAPI technique was used. The DAPI was excited using the standard UV source (mercury lamp) and the image captured using the confocal photomultipliers. The DAPI image was then merged with the other channels retrospectively using the proprietary LSM Meta software. A 40x objective was used in most cases.

2.21 Fluid secretion assay

The miniaturised version of classical Ramsay assay for tubule fluid secretion was used for measuring rates of secretion (Dow et al., 1994) as illustrated in Figure 2-7. The pairs of tubules were dissected along with the ureter and transferred to a 9 µl drop of *Drosophila* saline: Schneider's (50:50) under 25X Microscope. One end was wrapped around a metal pin under white, heavy mineral oil (Sigma UK) whilst the other tubule was immersed in 9 µl drop containing trace amounts of the red dye, Amaranth, for easy viewing of the emerging bubbles. Care was taken to ensure that the ureter remained in the oil but out of the 9 µl drop. Drops emerging from the ureter were removed with a fine glass rod; the diameter of each drop of the secreted fluid was measured at 10 minutes intervals under a microscope graticule (50x). From this diameter the volume of secreted fluid in nl/min was calculated using equation: volume= $(4/3)\ \pi r^3$ (Figure 2-7).

Drugs including antagonists and agonists were added to the Schneider: saline as a 10x stock in 1 µl when required. The drugs used in these experiments were the neuropeptides capability-1 (capa-1) and leucokinin (also called drosokinin) (both custom synthesised by Research Genetics Inc.). All compounds were initially dissolved in an appropriate solvent then diluted to the 10x stock in 1:1 (v/v) Schneider: saline.

Results were analysed using a Microsoft Excel sheet, where the secretion in nl/min was calculated from the volume of fluid secreted in 10 min. All data was reported as mean ± SEM and viewed using GraphPad Prism software v5 (GraphPad Software Inc) USA.

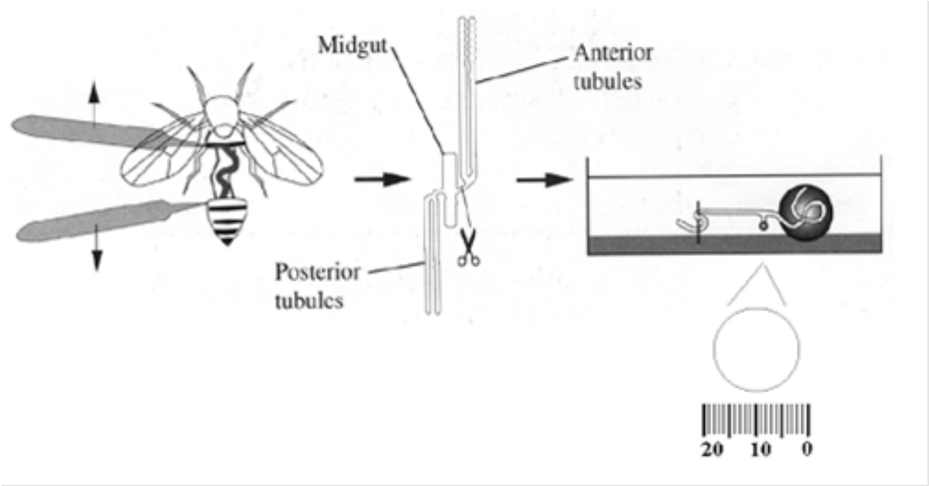


Figure 2-7 Fluid secretion assay schema (Dow et al., 1994).

Intact tubules are dissected along with their common ureter from the flies in Schneider's medium using fine forceps (left panel). Tubule ureter is cut just before its joining with gut (middle panel). One tubule is wrapped around the needle and other tubule is in the Schneider: saline mix; all are immersed in the mineral oil (right panel; above). Finally tubule secreted droplets emanating from the ureter are measured using the microscope graticule and converted in to nl/min (right panel; below)

3. Three-way analysis of *Drosophila* expression microarray, *Drosophila* tiling microarray and RNA-seq

Summary

In this chapter, three technologies, *Drosophila* RNA-seq, *Drosophila* tiling microarray and *Drosophila* expression microarrays were compared in order to find out the correlation between them, and the advantages and drawbacks of each other. The results suggested that the three technologies were correlated well in both absolute gene expression and relative gene expression level, both from the technical and biological view points. The correlation between RNA-seq and expression microarray was better than that with tiling microarray when detecting gene expression. However, tiling microarray was able to discern novel genes and verify RNA-seq results. RNA-seq has a number of merits over both tiling microarrays and expression microarrays, such as large dynamic range, ability to detect low expression genes, and the ability to identify novel genes and novel alternative splicing isoforms. For these reasons, RNA-seq was chosen in this thesis as a tool to discover novel genes in *Drosophila* tubules.

3.1 Introduction

There are many gene expression profiling and gene discovery methods in the history. Northern blots, reverse-transcription PCR (RT-PCR) and the automated Sanger-sequence-based technologies including Expression Sequencing Tag (EST) (Adams et al., 1991) and Serial Analysis of Gene Expression (SAGE and SuperSAGE) (Velculescu et al., 1995) etc. In the past decade, there has been tremendous progress in the development of methods to measure gene expression and search for novel genes at the whole transcriptome level. Among these methods, RNA-seq and DNA microarrays stand out as the two most widely used methods for genome-wide gene expression quantification and novel gene and splicing discovery (Gros, 2003; Wang et al., 2009). RNA-seq and DNA expression microarrays are two popular methods to measure gene expression, whilst RNA-seq and DNA tiling microarrays are two popular methods for gene discovery at the genome-wide scale. How are these technologies correlated to each other?

There were a number of publications of two-way comparison, either between RNA-seq and expression microarrays or between RNA-seq and tiling microarrays. There is no publication describing a three-way comparison of RNA-seq, tiling microarrays and expression microarrays so far. Such a comparison will provide a better understanding of how these technologies work.

3.2 **Background**

3.2.1 ***Description of technologies***

3.2.1.1 ***Drosophila* RNA-seq**

RNA-seq is a recently developed approach based on the high-throughput deep sequencing technologies also called ‘next-generation sequencing’. Transcriptome analysis aims to measure all transcription within the genome including coding RNA (mainly mRNA) and noncoding RNA information. However the majority of RNA molecules are tRNAs and rRNAs, which are not transcribed but serve as the primary site of biological protein synthesis (translation). mRNA accounts for only 1–5% of whole RNA population, and can be distinguished from tRNA and rRNA by the presence of a poly(A) tail. In order to obtain accurate mRNA information, RNA-seq either has to select positively for the poly(A) information, or reduce the ribosome information. So RNA-seq methods can be separated into the poly(A) selection method and the ribosome reduction method. These two categories of method address different aims. The first one aims to measure the transcriptome of coding genes and to discover novel coding and poly(A) based noncoding genes by using oligo (dT) captured methods in order to pull poly(A) RNA out of the whole RNA population. The second one aims to measure all transcripts within the genome as well as discovering all novel transcripts at the genome level by reducing the ribosome RNA information and obtaining all the transcription information including coding RNA and noncoding RNA. The RNA-seq work flow is detailed in section 1.1.7.2 and illustrated in Figure 3-1.

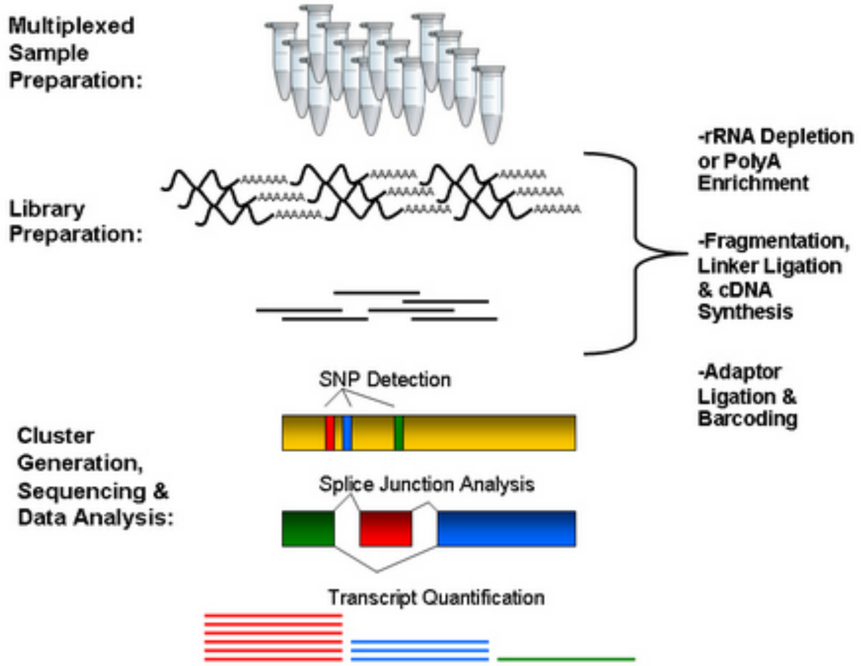

Multiplexed Sample Preparation:
Library Preparation:
Cluster Generation, Sequencing & Data Analysis:
-rRNA Depletion or PolyA Enrichment
-Fragmentation, Linker Ligation & cDNA Synthesis
-Adaptor Ligation & Barcoding
SNP Detection
Splice Junction Analysis
Transcript Quantification

Figure 3-1 RNA-seq work flow.

Picture is taken from seq Wright Genomic Services website http://www.seqwright.com. A representation showing the essential protocol by which the two types RNA-seq work. Total RNA started with Poly(A) selection by using oligo(dT) captured method or rRNA depletion with ribosome reduction method.

There are several different sequencing platforms on the market that support RNA-seq such as the Illumina Genome analyser, Roche 454 Genome Sequencer, Applied Biosystem SOLiD, and Life Technologies IonTorrent. These technologies have their own relative advantages and disadvantages (Zhou et al., 2010).

3.2.1.2 ***Drosophila*** **tiling microarrays**

Tiling microarrays are a subtype of microarray chip. Like traditional microarrays, they function by hybridizing labelled DNA to probes fixed onto a solid surface. Tiling microarrays differ from traditional microarrays in the nature of the probes. Instead of probing for sequences of known or predicted genes that may be dispersed throughout the genome, tiling microarrays probe intensively for sequences that are known to exist in a contiguous region of the genome. There are two types of tiling microarrays design, one using partially overlapping probes and the other one non-overlapping probes (Figure 3-2). This is useful for characterizing regions of the genome that are sequenced but with local functions that are largely unknown.

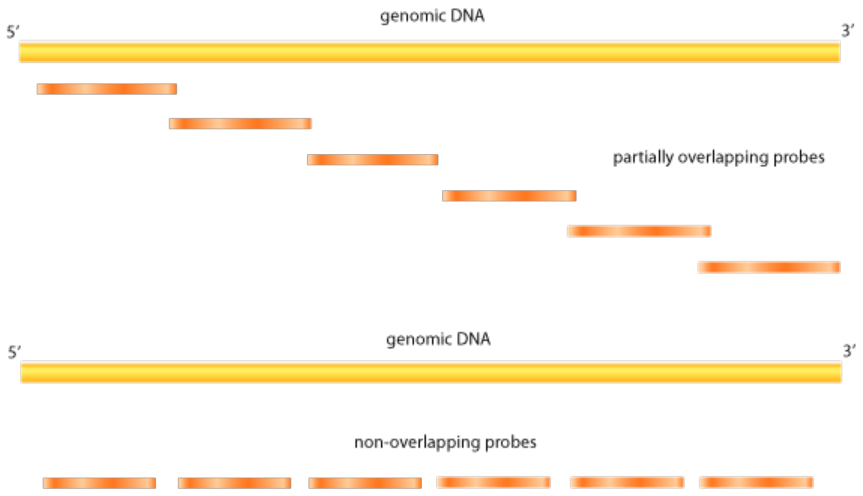


Figure 3-2 Tiling microarrays probe design.

There are two types of tiling microarrays design, one is the partially overlap probes, the other one is the non-overlapping probes. Picture taken from http://www.absoluteastronomy.com/topics/Tiling_array.

The GeneChip *Drosophila* Tiling 2.0R Array (Tiling 2) was designed using the *Drosophila* sequence release from the Berkeley *Drosophila* Genome Project (BDGP), March 2006 (details also in 1.1.6.1). The array includes both euchromatic and heterochromatic sequences. Average probe spacing on the array is 39 base pairs. The *Drosophila* Tiling 2.0 Array is designed to detect sequences in the reverse (-) orientation. It contains a 25bp mismatch (MM) probes for each perfect match (PM) probes, with a single base mismatch is located in the middle at 13th base.

There has been a number of publications report the use of *Drosophila* Tiling 2.0R Arrays for gene expression detection (Cherbas et al., 2011), novel transcript region detection (Graveley et al., 2011; Manak et al., 2006), and transcription binding site detection (Abruzzi et al., 2011; Smith et al., 2009). These publications confirmed that *Drosophila* tiling microarrays are still a useful tool for *Drosophila* research.

3.2.1.3 ***Drosophila* expression microarrays**

Microarrays are based on the principle of mRNA hybridization to a complementary probe as used in Southern blotting, which enables the detection of a specific transcript. The single colour type of microarray is represented by

Affymetrix GeneChip (Figure 3-3). The dual colour type of microarray is represented by Agilent technologies.

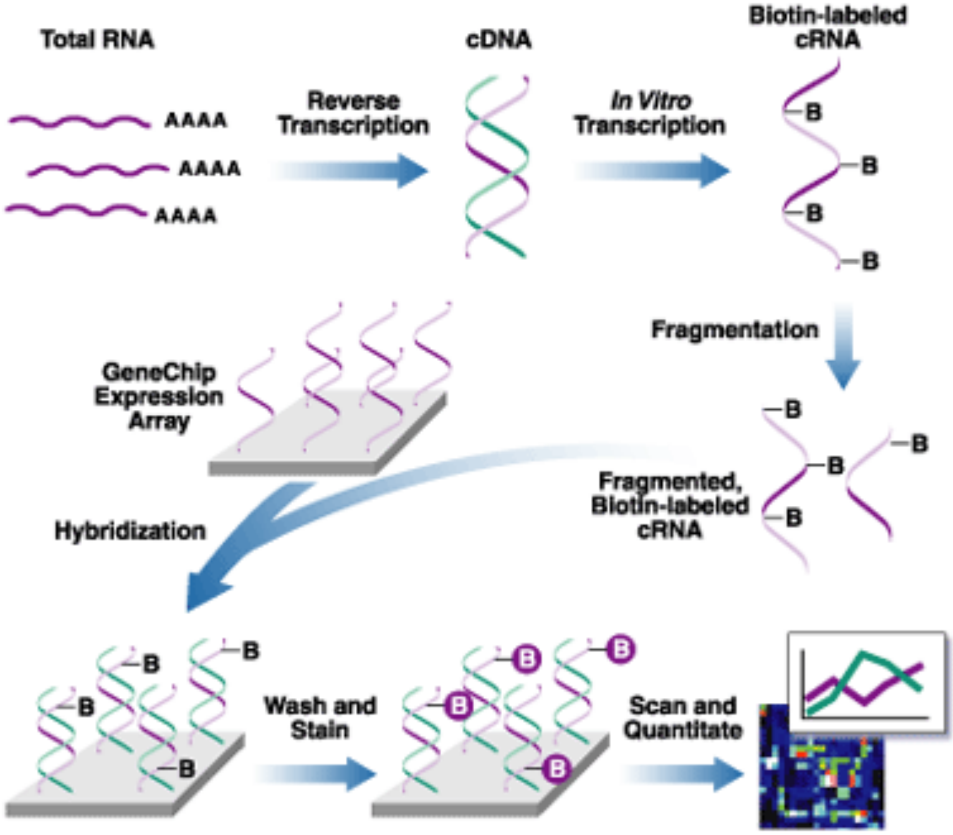


Figure 3-3 The essential protocol showing how expression of an organism can be analysed using DNA microarrays.

Picture was taken from http: //www. affymetrix.com

*Drosophil*a Genome 2.0 Array (details also in 1.1.6.1) is a three-prime expression microarray which has been used for *Drosophila* Genome profiling for nearly a decade, delivering results to the satisfaction of the *Drosophila* community. There are a number of publications available using these Genechips (Wang et al., 2004). The FlyAtlas website was set up using all the information generated using *Drosophila* Genome 2.0 Arrays to measure the gene expression of all individual tissues of *Drosophila* in comparison to a matched whole-fly sample (Chintapalli et al., 2007; Estrada and Michelson, 2008; Willis et al., 2010; Zhu et al., 2013a). The FlyAtlas presents an excellent opportunity to study gene expression in multiple tissues and provides a complementary resource to published developmental data sets (Arbeitman et al., 2002).

3.2.2 *Previous technology comparisons*

3.2.2.1 **Correlation between RNA-seq and expression microarrays**

To date, several studies have been conducted to compare the performance of RNA-seq and microarrays in quantifying the expression levels of genes from different aspects mainly focusing on reproducibility, accuracy, technical and biological variabilities (Kogenaru et al., 2012; Malone and Oliver, 2011; Marioni et al., 2008),. The two technologies have also been compared at the proteomics level (Fu et al., 2009).

Comparisons between the two techniques have been reported in *Candida parapsilolis* (Guida et al., 2011), *Drosophila melanogaster* (Malone and Oliver, 2011), *Saccharomyces cerevisiae* (Nookaew et al., 2012), on the fission yeast *Schizosaccharomyces pombe* (Wilhelm et al., 2010; Wilhelm et al., 2008), *Lactobacillus plantarum* (Leimena et al., 2012), pathogenic bacteria *HrpX regulome* (Kogenaru et al., 2012), in mouse tissues (Liu et al., 2007; t Hoen et al., 2008), in *Canis familiaris* (Mooney et al., 2013), in several human cells and cell lines (Bradford et al., 2010; Fu et al., 2009), and in human and non-human primate tissue (Liu et al., 2011a; Marioni et al., 2008; Raghavachari et al., 2012).

In all cases the expression levels have showed strong agreement between the two technologies, with correlation ranging from 0.6-0.8 for biological replicates, and from 0.9-0.96 for technical replicates. Correlation of the relative measurement of differential gene expression between the two technologies also agreed, with correlation range from 0.7-0.9 (Malone and Oliver, 2011; Marioni et al., 2008). RNA-seq had better reproducibility, accurrance and dynamic range than microarrays due to the microarray hybridization background, which affects the measurement (Mooney et al., 2013; t Hoen et al., 2008). However, both technologies have an increased error rate for lowly expressed genes (Liu et al., 2011a). RNA-seq needs to increase the read coverage and microarrays need more density to cope with the lowly expressed genes.

Two studies also addressed RNA-seq and microarrays in terms of having complementary strengths and limitations with each other. Detecting genes with low expression will remain a problem for both technologies (Kogenaru et al., 2012; Malone and Oliver, 2011). RNA-seq has been shown to be better for novel gene discovery and novel isoform identification than microarrays. The correlation between *Drosophila* expression microarrays (*Drosophila* Genome 2.0 Array) and RNA-seq (Illumina GAIIx) will be addressed in this chapter.

3.2.2.2 **Correlation between RNA-seq and tiling microarrays**

In terms of the correlation between RNA-seq and tiling microarrays, most studies investigated the gene expression levels as well as the differential levels of gene expression detection.

There are some different opinions regarding the correlation between RNA-seq and tiling microarrays. One study found the agreement for gene expression detection between RNA-seq and tiling microarrays to be very good with correlation given as a Spearman's correlation coefficient of approximately 0.9. The same study also found the differential expression of genes between the two technologies to show a correlation given as a Spearman's coefficient around 0.7 (Agarwal et al., 2010). Another study found that 80% of the bases detected as expressed by RNA-seq overlapped with those found using DNA-tiling microarrays (Nagalakshmi et al., 2008). On the other hand, one studies found the comparison between tiling microarrays and the tag based MPSS (Massively Parallel Signature Sequencing) technology revealed a very good overlap only at the protein-coding gene loci (Sasidharan et al., 2009). There was poor agreement between tiling microarrays and RNA-seq for detecting certain genes expression for example; *Wnt* gene in *Drosophila* cell lines could be detected in RNA-seq but undetectable by tiling microarrays. However, other evidence confirmed RNA-seq result was the more accurate one for the *Wnt* genes (Cherbas et al., 2011; Sasidharan et al., 2009).

Studies using tiling microarrays and RNA-seq report agreement at nearly 90% for the discovery of novel transcribed regions (Cherbas et al., 2011; Graveley et al., 2011) and detection of the 3' and 5' extension of transcriptional activities of genes (Graveley et al., 2011). However, another report suggests that tiling microarrays overestimated the transcriptional signals due to the high false-positive rates, especially overestimating the novel genes within the genome (van

Bakel et al., 2010). Taken together, these studies suggest that tiling microarrays can still be treated as a useful tool to verify RNA-seq findings at global levels.

3.3 Experimental design for three-way comparison

Evaluation of the three technologies was performed using a set of three-way comparisons, namely a comparison of absolute expression using technical replicate samples, a comparison of absolute expression using biological replicates samples, and a comparison of relative expression using biological replicate samples.

Technical replicate comparison was used in order to compare the performance of the three technologies in measuring the expression level of known and novel genes. Biological replicate comparison was used to investigate how the difference between the three technologies affects the biological interpretation.

3.3.1 *Absolute gene expression comparison using technical replicate samples*

The aim of the comparison of absolute expression using technical replicate samples was to compare the performance of the three technologies in measuring the expression level of known and novel genes. The design for this comparison was to analyze three technical replicate samples of *Drosophila* whole fly (Canton S) using *Drosophila* Genome 2.0 Array, three technical replicate samples of whole fly using *Drosophila* Tiling Array 2.0R (same RNA of Canton S as *Drosophila* Genome 2.0), and one technical sample of whole fly using RNA-seq (same RNA of Canton S as *Drosophila* Genome 2.0). The absolute gene expression results obtained were then compared by using scatter plot analysis.

As shown in Figure 3-4, three scatter plots were generated for technical replicate samples comparison; whole fly RNA-seq *versus* whole fly *Drosophila* Genome 2.0 Array (Dros 2), whole fly RNA-seq *versus* whole fly *Drosophila* Tiling Array 2.0R (Tiling 2) and whole fly *Drosophila* Tiling 2.0R Array *versus* whole fly *Drosophila* Genome 2.0 Array.

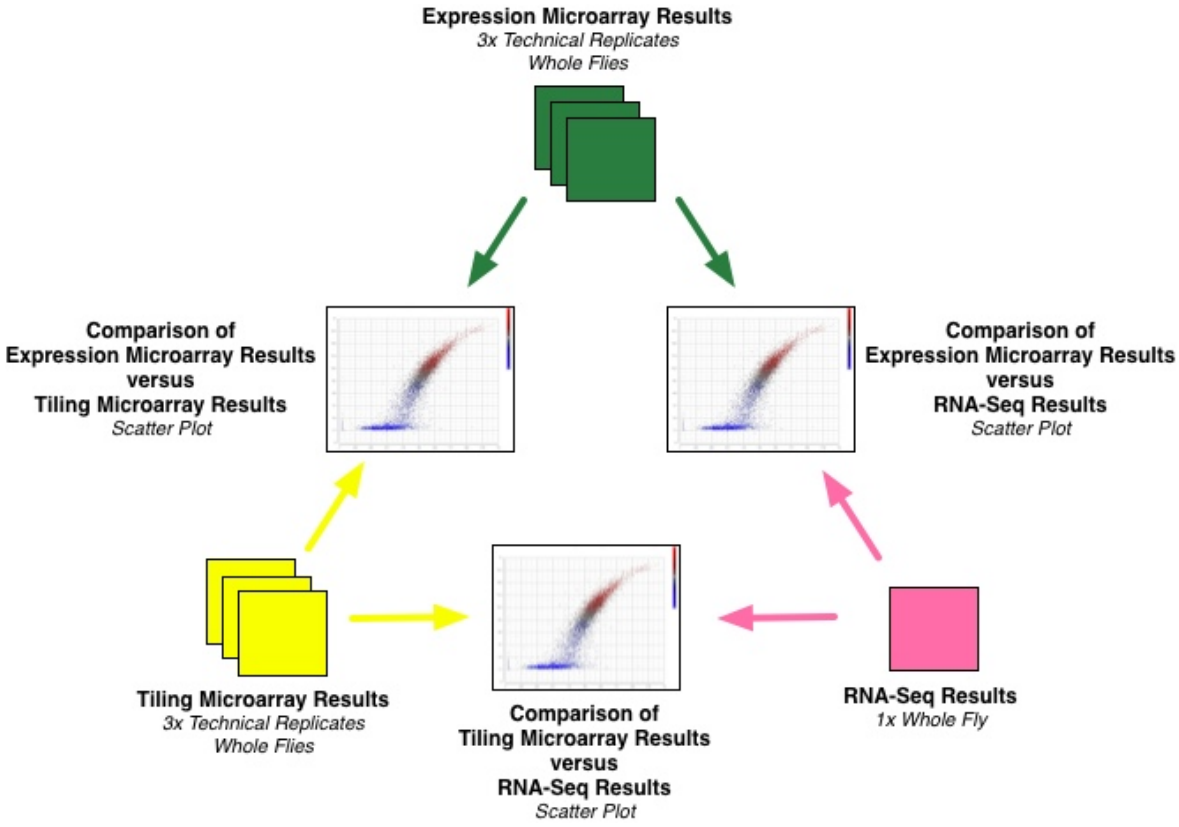


Figure 3-4 The flow diagram of absolute gene expression analysis using technical replicate samples.

The same RNA from whole flies samples were used in these comparisons by applying to *Drosophila* expression microarrays, *Drosophila* tiling microarrays and RNA-seq platform. Three scatter plots were generated as results.

The results for comparison of absolute expression using technical replicates are shown in 3.4.1.1 to 3.4.1.3 respectively.

3.3.2 *Absolute gene expression comparison using biological replicate samples*

The aim of the comparison of absolute expression using biological replicate samples was to investigate how the difference between the three technologies affects the biological interpretation. The design for this comparison was to analyze three biological replicate samples of *Drosophila* whole fly (Canton S) using *Drosophila* Genome 2.0 Array, three biological replicate samples of whole fly using *Drosophila* Tiling Array 2.0R (same RNA of Canton S as *Drosophila* Genome 2.0), and two biological samples of whole fly using RNA-seq (same RNA of Canton S as *Drosophila* Genome 2.0). The absolute gene expression results obtained were then compared by using scatter plot analysis.

As shown in Figure 3-5, three scatter plots were generated for biological replicate samples comparison; whole fly RNA-seq *versus* whole fly *Drosophila* Genome 2.0 Array, whole fly RNA-seq *versus* whole fly *Drosophila* Tiling 2.0R Array and *Drosophila* Tiling Array 2.0R *versus* whole fly *Drosophila* Genome 2.0 Array.

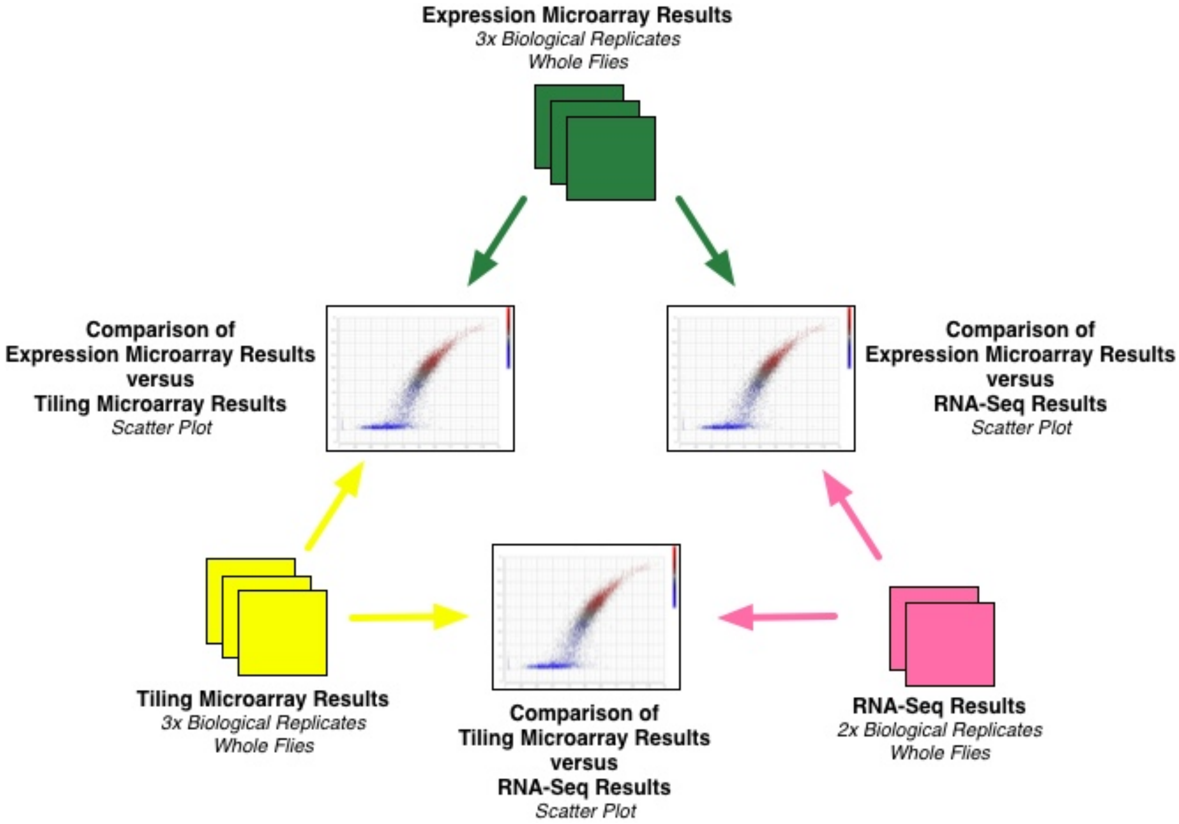


Figure 3-5 The flow diagram of absolute gene expression analysis using biological replicate samples.

The RNA from whole flies samples which all replicate samples being raised in the same conditions were used in these comparisons by applying to *Drosophila* expression microarrays, *Drosophila* tiling microarrays and RNA-seq platform. Three scatter plots were generated as results.

The results for comparison of absolute expression using biological replicates are shown in 3.4.1.4 to 3.4.1.7 respectively.

3.3.3 *Relative gene expression comparisons using biological replicates*

The aim of the comparison of relative expression using biological replicate samples was to investigate how the difference between the three technologies affects the biological interpretation. The design of the biological replicate samples comparison was based on analysing the relative expression in tubules relative to whole flies. Three biological replicate samples of whole fly (Canton S) and tubules were analysed using *Drosophila* Genome 2.0 Array, three biological replicate samples of whole fly and tubules were analysed using *Drosophila* Tiling 2.0R Array (Canton S), and three biological replicate samples of whole fly and tubules were analysed using RNA-seq (Canton S). The relative gene expression results obtained were then compared by using scatter plot analysis.

As shown in Figure 3-6. Three scatter plots were generated for biological replicate samples comparison on tubule *versus* whole fly fold change. RNA-seq tubule/whole fly ratio *versus Drosophila* Genome 2.0 Array tubule/ whole fly ratio, RNA-seq tubule/whole fly ratio *versus Drosophila* tiling arrays 2.0R tubule/whole fly ratio and *Drosophila* Genome 2.0 Array tubule/whole fly ratio *versus Drosophila* tiling array 2.0R tubule/whole fly ratio.

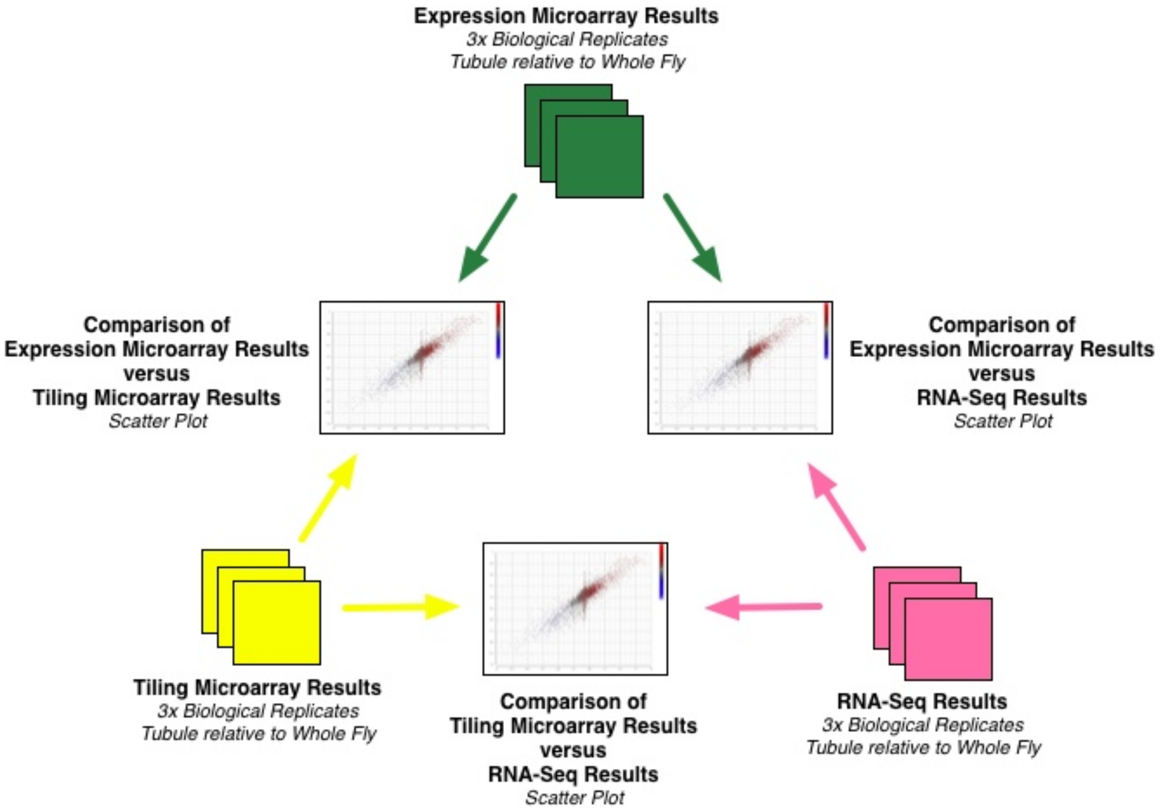

Expression Microarray Results
3x Biological Replicates
Tubule relative to Whole Fly
Comparison of
Expression Microarray Results
versus
Tiling Microarray Results
Scatter Plot
Comparison of
Expression Microarray Results
versus
RNA-Seq Results
Scatter Plot
Tiling Microarray Results
3x Biological Replicates
Tubule relative to Whole Fly
Comparison of
Tiling Microarray Results
versus
RNA-Seq Results
Scatter Plot
RNA-Seq Results
3x Biological Replicates
Tubule relative to Whole Fly

Figure 3-6 The flow diagram of relative gene expression analysis using biological replicate samples.

The RNA from whole fly and tubule samples which all replicate samples being raised in the same conditions were used in these comparisons by applying to *Drosophila* expression microarrays, *Drosophila* tiling microarrays and RNA-seq platform. The fold changes were calculated using tubule/whole fly. Three scatter plots were generated by using ratio of tubule/whole fly as results.

The results for comparison of relative expression using biological replicates are shown in 3.4.2 respectively.

3.4 Results

3.4.1 *Absolute gene expression level comparison*

3.4.1.1 Technical replicate samples of RNA-seq versus *Drosophila* Genome 2.0 Array (three-prime expression array)

RNA was extracted from Canton S whole flies, and the same RNA was used for the technical replicate sample comparison of RNA-seq, *Drosophila* Genome 2.0 Array, and *Drosophila* Tiling 2.0R Array.

RNA-seq was performed on Canton S whole fly using single-end Illumina GAIIx, 76bp library preparation kit according to Illumina RNA-seq sample preparation manual (section 2.8.1). The result of 13.7 million raw reads was cleaned using an in-house script, and then the cleaned reads imported into Bowtie (Langmead et al., 2009; Trapnell et al., 2009); Bowtie aligned the reads to the *Drosophila* Genome (dm3.refFlat) using unique reads only, and then generated aligned bam files. The bam files were imported into Partek Genomic Suite and normalized in Partek. The signal of RNA-seq data is defined as a count of the number of reads overlapping at each base pair, and the reads per kilobase of exon per million mapped reads (RPKM) for each gene (Mortazavi et al., 2008). The in-house script took normalized Partek data and generated $\log_2$ (RPKM+0.01) value for genes characterised by one transcript only. The $\log_2$ value can't be generated if RPKM is 0, so 0.01 was added to each RPKM in case the gene expression value is 0. Three technical replicate samples of whole fly using the same RNA as RNA-seq analysed using *Drosophila* Genome 2.0 Array started from 2.0ug RNA according to Affymetrix gene expression manual (section 2.7.1). The .CEL files of the samples were created after scanning, imported into Partek Genomic Suite (version 6.5), and normalized in Partek using GCRMA, which is the popular normalization method performs the background correction by considering the GC contents and increases the signal to noise ratio (annotation file Drosophila_2 from Affymetrix). A microarray signal is defined as an intensity value for each probe, and each probe's value as GCRMA normalized PM value. The signal intensity of 11-20 probes was combined as the signal intensity of a gene. An in-house developed script took probe-set level data for Affymetrix replicate samples of whole fly and produced $\log_2$ (signal intensity) median values for uniquely identified genes with only one unique probe.

An in-house Perl script was used to joined the RNA-seq and *Drosophila* Genome 2.0 Array two data sets together. A total of 9953 genes was included in this data set from the two platforms. Two values were produced for each gene, namely the median affy $\log_2$ (signal intensity) and $\log_2$ (RPKM). Figure 3-7 shows the scatter plot drawn by Partek.

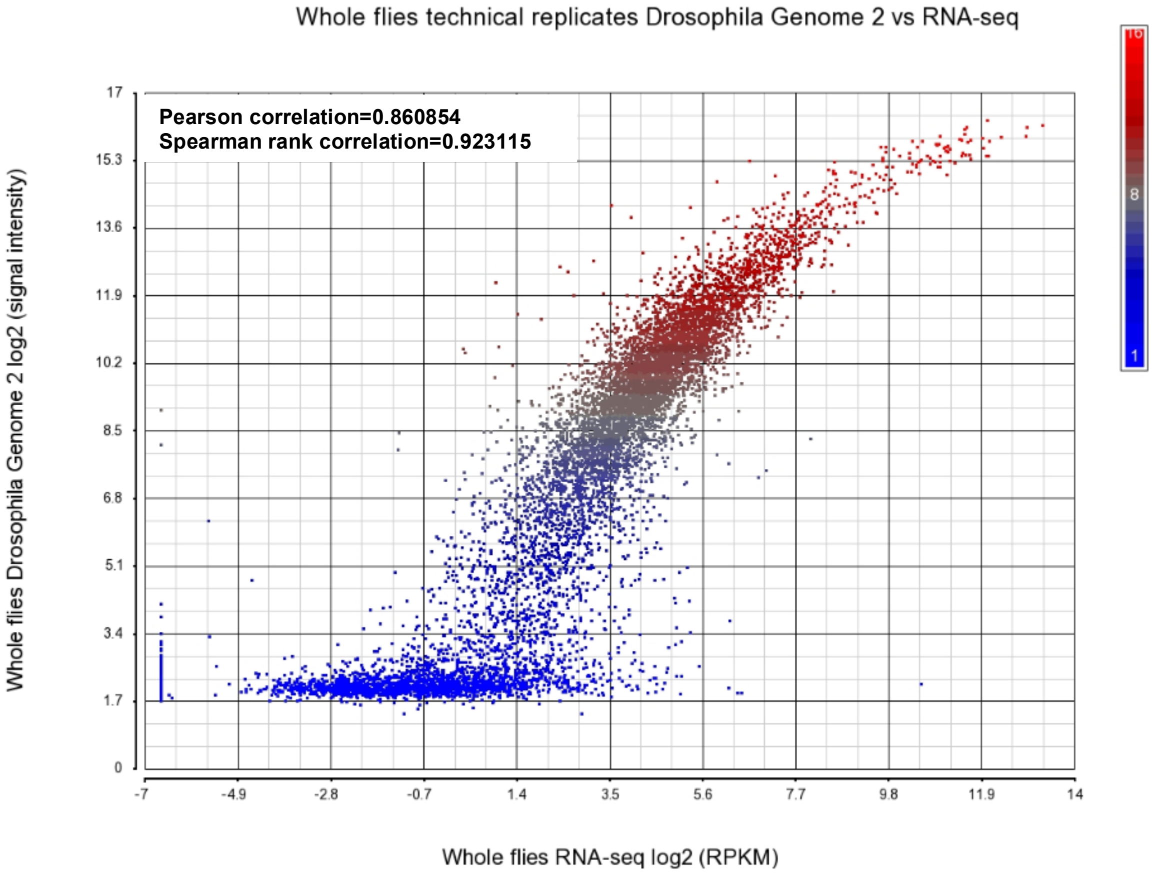


Figure 3-7 Scatter plot of whole flies technical replicate samples of RNA-seq compared with *Drosophila* expression microarrays.

The expression microarrays $\log_2$ (signal intensity) compared with 76bp Illumina $\log_2$ (RPKM). The scatter plot showed strong but nonlinear correlation between the two measurements with Pearson correlation coefficient 0.860854 and Spearman rank correlation at 0.923115. Genes with only one transcript, with unique mapped reads, and the unequivocal annotation were chosen for both platforms. The colour of the spots indicates the expression value of both platforms; the value range from low (<8) to high (>8), the colour changes from blue to red. Data values at the far bottom left quadrant for arrays signify genes highly expressed in *Drosophila* Genome 2.0 Array but undetectable by RNA-seq. The low end of RNA-seq corresponding to RNA-seq $\log_2$ (RPKM) in the range from 2-10 and Dros 2 $\log_2$ (signal intensity) <2.3 signifies the genes that were absent in *Drosophila* Genome 2 Array but some genes were detected by RNA-seq. These spots indicate the main different expressed genes between the two platforms.

The scatter plot in Figure 3-7 shows the technical replicate samples comparison between Dros 2 and RNA-seq. The correlation is better than a previous report which had a Spearman rank correlation 0.73 (Marioni et al., 2008); our data had Pearson correlation coefficient 0.86054 and Spearman rank correlation 0.923115.

The main reasons are that the samples were technical replicate samples, the genes were single transcript genes, RNA-seq used the unique reads, and both platforms chose the unequivocal annotation genes. These steps increased the correlation between the two platforms. However, there were still a group of genes at the low end on arrays that were highly expressed in Dros 2 but were undetectable by RNA-seq (left corner of Figure 3-7). Another group of genes located on the low end of the RNA-seq (bottom spots of RNA-seq of Figure 3-7) showed lowly expressed genes (the absent calls) in Dros 2 but most of them could be detected by RNA-seq. Some of the values were quite high, for example some genes Log_2 (RPKM) could reach to 5 to10. These genes would be further investigated in the RNA-seq and *Drosophila* expression microarrays comparison of the biological replicate samples.

3.4.1.2 **Technical replicate samples of RNA-seq versus *Drosophila* tiling microarrays**

The scatter plot shows that tiling microarrays and RNA-seq are reasonably correlated with a Pearson correlation coefficient of 0.66397 and Spearman rank correlation 0.756478 (Figure 3-8).

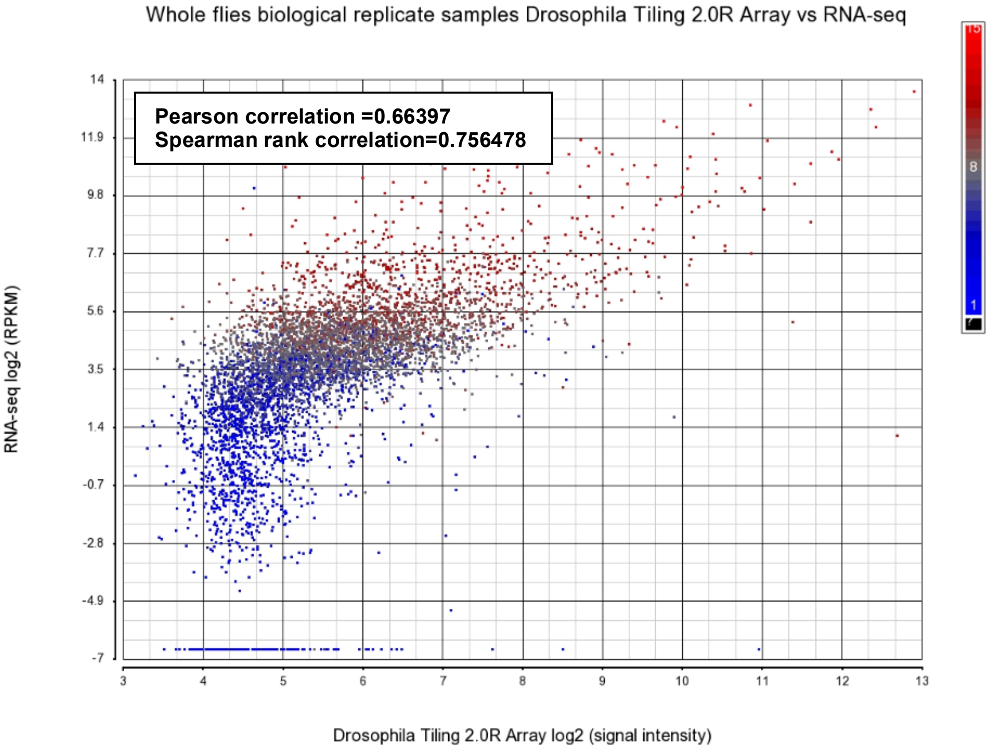


Figure 3-8 Scatter plot of whole flies technical replicate samples of RNA-seq compared with *Drosophila* tiling microarrays.

The scatter plot of technical replicate samples 76bp single-end RNA-seq $\log_2$ (RPKM) compared with *Drosophila* Tiling 2.0 Array. The scatter plot shows good nonlinear correlation between the two measurements with Pearson correlation coefficient 0.66397 and Spearman rank correlation 0.756478. Genes with only one transcript, and the unequivocal annotation were chosen for both platforms. The colour of the spots denote the expression values from both platforms, the value ranging from low (<8) to high (>8), the colour changed from blue to red respectively. The bottom left quadrant of tiling microarrays and RNA-seq were genes that were expressed in tiling microarrays and were undetectable by RNA-seq showed the high false-positive rates of tiling microarrays compared to RNA-seq. Genes cannot detect by both platforms (far bottom left quadrant) indicated most noise are from the lowly expressed genes from two platforms. The compressed signals at the top of tiling microarrays indicated the saturation signals and the low dynamic range of tiling microarrays.

Highly expressed genes are closely correlated than the genes at low expression levels. The genes at bottom left quadrant of RNA-seq and tiling microarrays are those genes which highly expressed in tiling microarrays but were undetectable by RNA-seq. These genes likely represent cross-hybridization of the microarrays which causes the high-false positive signals in tiling microarrays (Agarwal et al., 2010). Signals at the top of the range were compressed in the tiling microarrays indicating saturation of the signals caused by the scanner.

3.4.1.3 Technical replicate samples of *Drosophila* expression microarrays *versus Drosophila* tiling microarrays

Three technical replicate samples of whole fly were started from 7μg RNA according to the GeneChip Whole Transcript (WT) Double-Stranded Target Assay Manual from Affymetrix, as detailed in section 2.7.2. CEL files of the samples were imported into Partek Genomic Suite, and normalized by GCRMA in Partek (version 6.6) according to the Partek manual.

For Partek analysis, a tiling microarray signal is first defined as an intensity value for each probe. The PM minus MM values are computed for all replicates, and the replicates' signals of genes are cut into segments where the segments contain a minimum of nine probes in a region; The P-value threshold for testing the difference in intensity level between the test segments and neighbouring segments was 0.01; signal is higher than noise at least by 10%. The signal-to-noise threshold was set at 0.1 as default). Secondly, the segments were assigned to genes based on the annotation of *Drosophila* tiling microarrays allowing for 100bp extension of both ends of genes by a developed in-house script. Thirdly, those genes with only one transcript and an unequivocal annotation were chosen. The median value was chosen for the segments for one gene as an expression value. The tiling microarrays results were joined together with the Dros 2 and RNA-seq results in Partek giving a total of 9616 genes. The scatter plot showing the technical replicate samples comparison of *Drosophila* Genome 2.0 Array *versus Drosophila* Tiling 2.0R Array is shown in Figure 3-9.

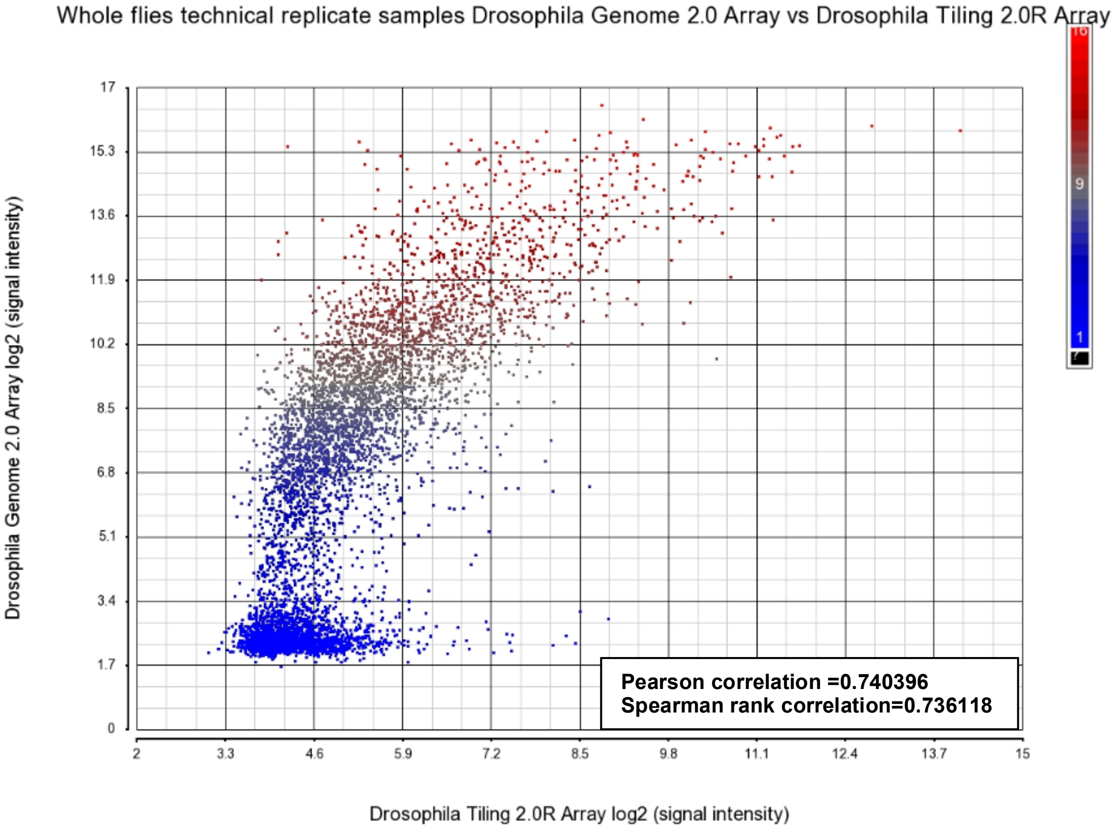


Figure 3-9 Scatter plot of whole flies technical replicate samples of *Drosophila* expression microarrays compared with *Drosophila* tiling microarrays.

Drosophila Genome 2.0 Array $\log_2$ (signal intensity) compared with *Drosophila* tiling 2.0R Array $\log_2$ (signal intensity). The scatter plot shows good nonlinear correlation between the two measurements with Pearson correlation coefficient 0.740396 and Spearman rank correlation 0.736118. Genes with only one transcript, and unequivocal annotation were chosen for both platforms. The colour of the spots vary according to the expression values from both platforms, the range from low (<8) to high (>8), indicated by the colour changing from blue to red. The bottom quadrant of tiling microarrays are genes expressed in tiling microarrays that were undetectable by Dros 2 array suggested the high-positive rates of tiling microarrays, and genes can't detect by both platforms (far bottom left quadrant) suggested that both platforms have low ability to detect the lowly expressed genes.

The scatter plot showing technical replicate samples for whole flies tiling microarrays and Dros 2 were reasonably correlated with a Pearson correlation of 0.740396 and Spearman rank correlation of 0.736118. However they were not correlated as well as the correlation of Dros 2 and RNA-seq, especially when measuring the low expression genes. However, Dros 2 and tiling microarrays data generally agreed at high expression levels. For the low expressed genes, both platforms suffered from hybridization background noise, making it difficult to detect the low expressed genes.

3.4.1.4 Biological replicate samples of RNA-seq *versus Drosophila* expression microarrays

Whole fly biological replicate samples in RNA-seq were compared with whole fly *Drosophila* Genome 2.0 Array, and with whole fly *Drosophila* Tiling 2.0R Array by using different batches of *Drosophila* Canton S flies with the same conditions.

Four whole fly biological samples were used for *Drosophila* Genome 2.0 Array and two whole fly biological samples for single-end 54bp Illumina GAIIx RNA-seq. Using the same methods which had been used with the whole fly technical replicate samples comparison, *Drosophila* Genome 2.0 Array was compared with RNA-seq to generate the joined *Drosophila* Genome 2.0 Array, RNA-seq and tiling microarrays expression data. This data was recorded in a Microsoft Excel sheet with gene ID indexes. The scatter plots were produced by Partek Genomic Suite. The scatter plot of biological replicate samples of *Drosophila* expression microarrays compared with RNA-seq was shown in Figure 3-10.

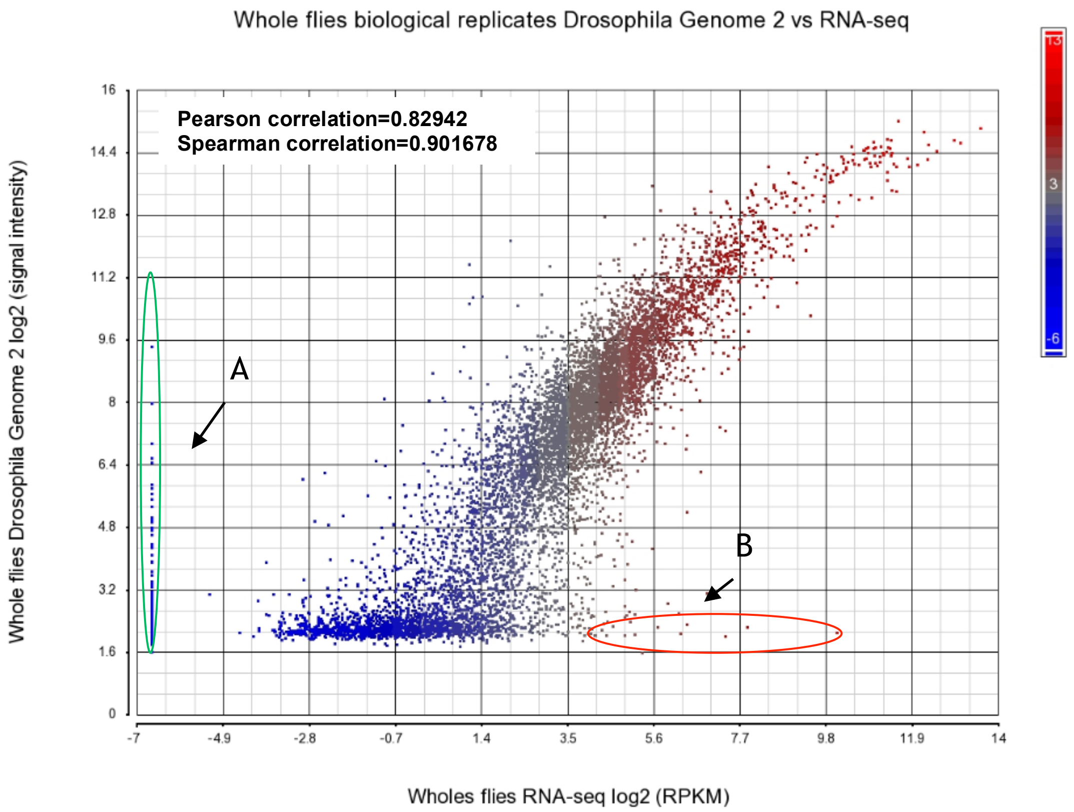


Figure 3-10 Scatter plot of whole fly biological replicate samples of RNA-seq compared with *Drosophila* expression microarrays.

The scatter plot of whole fly *Drosophila* Genome 2.0 Array $\log_2$ (signal intensity) compared with 76bp single-end Illumina GAIIx $\log_2$ (RPKM) showed strong but nonlinear correlation between the two measurements with a Pearson correlation coefficient 0.82924 and Spearman rank correlation 0.901678. Genes with only one transcript, with unique mapped reads, and the unequivocal annotation were chosen for both platforms. The dynamic range (ratio of largest observable value to apparent background value) of the RNA-seq data is clearly larger than that of microarray data. The microarray data appears to be slight compression at the top. The colour of the spots denote the expression value of both platforms, the value ranging from low (<8) to high (>8), the colour changed from blue to red respectively. The far left signals (arrow A) of RNA-seq were genes that were highly expressed in Dros 2 and were undetectable by RNA-seq. The low end of RNA-seq (arrow B) corresponds to RNA-seq $\log_2$ (RPKM) 4-10 and Dros2 $\log_2$ (signal intensity) <2 of the picture were the genes that were absent in Dros 2 but were detected by RNA-seq. These spots indicate the main difference in expressed genes between the two platforms.

Data from the biological replicate samples of Dros 2 compared with Illumina signal-end, 54-bp RNA-seq showed a strong but nonlinear correlation between the two platforms with a Pearson correlation coefficient 0.82942 and Spearman rank correlation 0.901678. This correlation was, however, not as good as the technical replicates. The dynamic range (ratio of largest observable value to apparent background value) of the RNA-seq data is clearly larger than that of the microarray data. The microarray data showed slightly compressed at the top, which meant that the microarray data was saturated when measuring the high signals. The main noise came from the highly expressed genes on Dros 2 but were undetectable by RNA-seq (arrow A of Figure 3-10) and from the absent genes on Dros2 but could be detected by RNA-seq (arrow B of Figure 3-10)

3.4.1.5 Investigation of transcripts unique to particular platforms

Further investigation of the transcripts unique to particular platforms was taken into two steps.

Table 3-1 Example of top ten genes which highly expressed on *Drosophila* expression microarrays, but with very low expression on RNA-seq

Gene ID	*Drosophila* Genome 2		RNA-seq	
	$\log_2$(Signal Intensity)	Present Call	$Log_{2(}$ RPKM)	Reads Present
CG31909	9.43915	P	-6.64386	A
Sdic3	7.97227	P	-6.64386	A
Lcp65Ag2	6.93959	P	-6.64386	A
Lcp65Ag3	6.57006	P	-6.64386	A
CG13068	6.46132	P	-6.64386	A
CG13705	5.88526	P	-6.64386	A
TwdlN	5.79068	P	-6.64386	A
TwdlM	5.66576	P	-6.64386	A
CG10598	5.52161	P	-6.64386	A
CG17290	5.32895	P	-6.64386	A

This table listed the top ten genes that were highly expressed on Affymetrix *Drosophila* Genome 2.0 Array but undetectable by RNA-seq. Calls labelled “P” indicated ‘present’ on Affymetrix *Drosophila* Genome 2.0 Array, mainly indicating that the array signal intensity >5 .Reads labelled “A” indicated undetectable reads by RNA-seq, which were indicated by the RPKM <4 (in this study).

Firstly, the RNA-seq data was sorted and 863 genes with RNA-seq $\log_2$ (RPKM) value-6.64 (arrow A of Figure 3-10) were selected. These were coordinated with the top ten genes that were highly expressed in *Drosophila* Genome 2.0 Array to identify the gene set for further investigation. The results are listed in Table 3-1.

Further details regarding with the Affymetrix gene ID to search for these spots were obtained by using NetAffx (www.affymetrix.com) description of these genes. Using this search, we found the probe and probe-set sequences, and the cross-hybridizing informations. We were also able to deposit the sequences and use the blast alignment tool in Ensembl genome browser to find the probe sets-gene match position to determine if these probes had been correctly designed by Affymetrix GeneChip. Problems with probes in microarrays are mainly caused by two reasons. One is cross-hybridization (false-positive or false-negative results); the other one is miss target transcripts (false-negative results) (Zhang et al., 2005). The search focused on these two reasons. We found that four of these genes, which were located at the position of the gene families (Lcp65Ag2, Lcp65Ag3; TWDIN, TWDIM), could cross-hybridize within the gene families, and some genes were cross-hybridized to several other genes within the genome (details in discussion).

Secondly, the Affymetrix *Drosophila* Genome 2.0 Array data was sorted. These genes with $\log_2$ (signal intensity) less than 2.3 and absent from Dros 2 were coordinated with the top ten highly expressed genes from RNA-seq and selected for further investigation as listed in Table 3-2.

Table 3-2 Example of ten highly expressed genes were detected by RNA-seq, but with very low expression in *Drosophila* expression microarrays.

Gene ID	Drosophila Genome 2		RNA-seq	
	$\log_2$(Signal Intensity)	Present call	Log2(RPKM)	Reads Present
CG34212	2.10397	A	10.0681	P
CG11042	2.235	A	7.89325	P
CG31210	2.00192	A	7.35785	P
CG32212	1.5825	A	5.311	P
CG14309	2.07286	A	6.2526	P
CG31804	2.3	A	6.41558	P
Dro	2.5909	A	6.22008	P
CG32212	1.5825	A	5.311	P
vas	2.23102	A	5.70787	P
CG3740	2.05629	A	5.13296	P

This table lists the top ten genes that were absent on Dros 2 but highly expressed on RNA-seq. A present call of “A” indicates absent on Dros 2, mainly indicated by the array signal intensity <5. A real call “P” indicates detectable reads by RNA-seq, mainly indicated by $\log_2$ (RPKM) >2 in this study.

To further investigate of the genes in Table 3-2, we blasted the probe-set sequences using the Ensembl (http://www.ensembl.org) to identify the probe-gene match to determine whether the probe or probe-set targeted the genes or any miss target effect. Meanwhile, the RNA-seq data was inspected in the Tablet (version 1.12.03.26) genome brower by viewing the bam files for whole fly data as generated by Bowtie of to check if the Affymetrix probe-set was designed in the right place.

The main reasons affecting microarray performance were missed target (CG14309) or part missed target (CG34212). Wrong annotation of the genes also affected the microarray performance (CG3212, CR31084, and CG11042). RNA-seq had more ability to detect the lowly expressed genes such as vas. The cross-hybridization affect was the major reason for false positive or false negative signals in the microarray platform (CG32212, CG3740). Details of the genes that were affected by the platform are also discussed further in section 3.5.

3.4.1.6 Biological replicate samples of RNA-seq *versus Drosophila* tiling microarrays

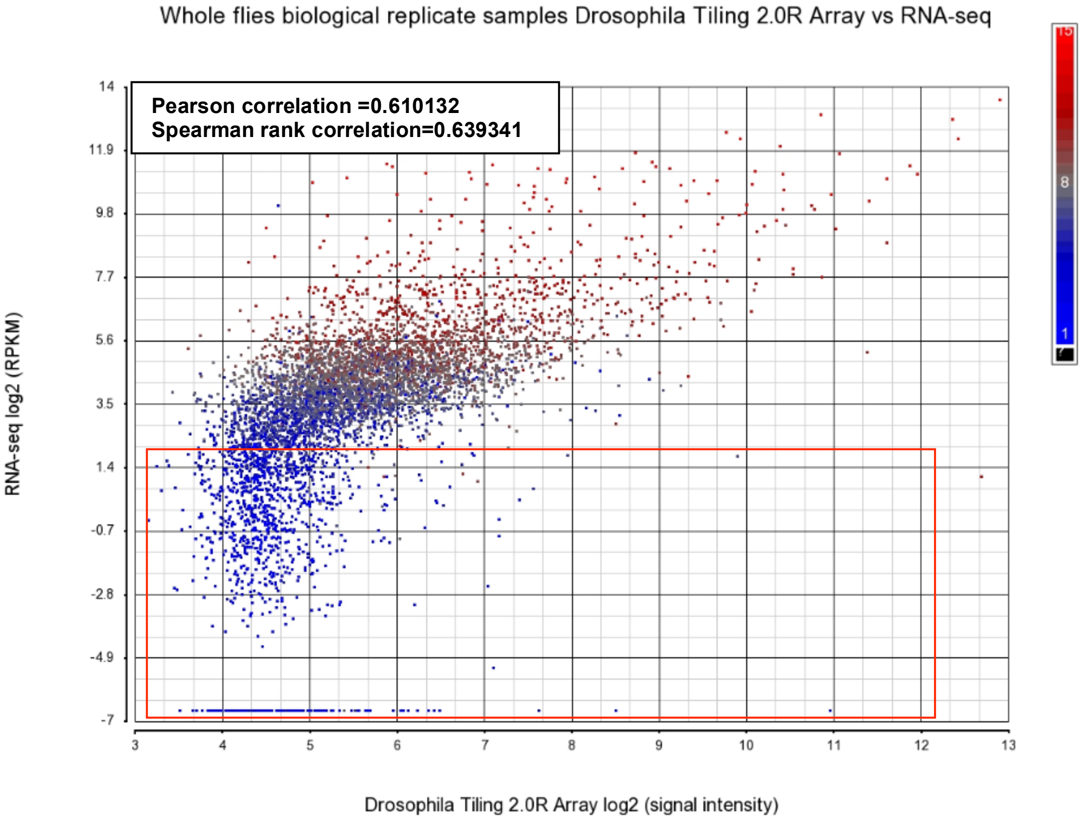


Figure 3-11 Scatter plot of whole fly biological replicate samples of RNA-seq compared with *Drosophila* tiling microarrays.

The scatter plot of *Drosophila* Tiling 2.0R array $\log_2$ (signal intensity) compared with 76bp Illumina RNA-seq $\log_2$ (RPKM) data showed good nonlinear correlation between the two measurements with Pearson correlation coefficient 0.610132 and Spearman rank correlation 0.639341. The dynamic range (ratio of largest observable value to apparent background value) of the RNA-seq data is clearly larger than that of tiling microarray data. The microarray data appears to be slight compression at the top. The colour of the spots denote the expression value of both platforms, the value ranging from low (<8) to high (>8), the colour changed from blue to red respectively. Genes with only one transcript, and the unequivocal annotation were chosen for both platforms. The low end of tiling microarrays (the row region corresponds to Tiling 2 $\log_2$ signals in the range >4 and corresponds to RNA-seq $\log_2$ RPKM <4 as indicated in the red square) that are those genes expressed in tiling microarrays and were undetectable by RNA-seq showed the high false-positive rates of tiling arrays than RNA-seq, and genes could not detect by both platforms (far bottom left quadrant) indicated much noise are from the lowly expressed genes from two platforms. The compressed signal at the top of tiling microarrays indicated the saturation of the signals.

Biological replicate samples comparison of RNA-seq and tiling microarrays displayed the same shape as the technical replicate samples comparison but with more noise at the lower expression genes indicating the known issue of cross-hybridization and the high false-positive rates of tiling arrays.

Table 3-3 Examples of false-positive signals in tiling microarrays but undetected by RNA-seq with the corresponding signals from Dros 2

Gene ID	Tiling microarrays $\log_2$ (signal intensity)	Expression microarrays Log_2 (signal intensity)	RNA-seq Log_2 (RPKM)
CG10102	10.95	2.1	-6.64
PPk7	8.5	2.1	-6.64
CG15212	7.6	4.7	-6.64
CG5195	7.17	2.2	-0.31
Aly	7.16	2.5	-0.87
CG15335	7.0	2.8	-2.5
CG32580	7.09	3.1	-5
Hdm	6.6	2.1	-6.64
Beat-IIIa	6.3	2.07	-0.56
CG12964	6.19	2.1	-3.14

This table lists ten genes that were highly expressed on tiling microarrays but were undetectable by RNA-seq and the corresponding signals of Dros 2. For Dros 2 signal intensity <5 or a $\log_2$ (signal intensity)<2.3 indicated that the signals were not detected. For RNA-seq, a $\log_2$ (RPKM) <2 indicated the genes were undetected in this study. The table showed that the tiling microarrays detected false-positive signals that were not detected by Dros 2 and RNA-seq..

Ten genes were selected as examples from the low-end of tiling microarrays (red square) in Table 3-3 that demonstrated the high false-positive signals detected by tiling microarrays data but since the both the RNA-seq and Dros2 data disagreed with the tiling microarrays data. Figure 3-16 also revealed the low dynamic range of tiling microarrays since the top signals of tiling microarrays are compressed. However, RNA-seq showed the wide dynamic range and less false-positive signals, but RNA-seq may need more reads to detect the lower expression genes.

3.4.1.7 Biological replicate samples of *Drosophila* expression microarrays versus *Drosophila* tiling microarrays

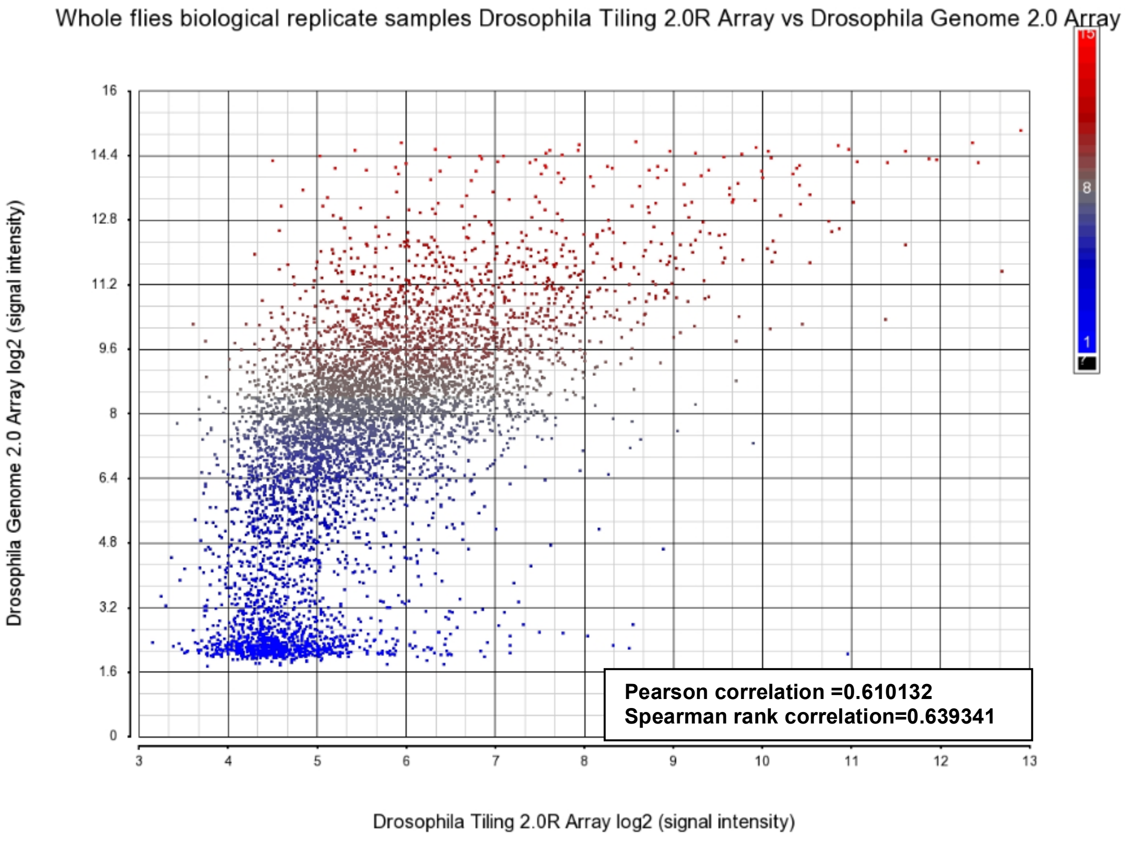


Figure 3-12 Scatter plot of whole flies biological replicate samples of *Drosophila* expression microarrays compared with *Drosophila* tiling microarrays.

The scatter plot showed good nonlinear correlation between the two measurements with Pearson correlation coefficient 0.610132 and Spearman rank correlation 0.639341. Genes with only one transcript, and the unequivocal annotation were chosen for both platforms. The colour of the spots are according the expression values from both platforms, the value from low to high, the colour changed from blue to red. The low end of tiling arrays and Dros 2 that were genes expressed in tiling arrays and were undetectable by Dros 2 array suggested the high-positive rates of tiling arrays, and genes could not detect by both platforms showed both platforms have low ability to detect the lowly expressed genes.

Biological replicate samples comparison of *Drosophila* expression microarrays and *Drosophila* tiling microarrays displayed the same shape of technical replicate samples comparison but with more noise at the lower expression genes, indicating the known issue of cross-hybridization of microarray and the high false-positive rates of tiling microarrays (Figure 3-12).

3.4.2 *Relative gene expression level comparison*

3.4.2.1 **Biological replicate samples of *Drosophila* tubule/whole fly expression microarrays *versus* tubules/whole fly RNA-seq**

These comparisons are between tubule/whole fly fold changes of Drosophila expression microarray and tubule/whole fly fold changes of RNA-seq. Four biological replicate samples of whole fly and tubule from Canton S flies were put on the *Drosophila* Genome 2.0 Arrays. Tubule and whole fly of *Drosophila* Genome 2.0 Array .CEL files were imported into Partek Genomic Suite with annotation from Affymetrix Drosophila_2. GCRMA normalization in Partek was applied, and a one-way ANOVA was generated for tubule/whole fly fold changes with FDR<0.01. Two biological replicates samples of tubule and whole fly were run by Illumina GAIIx using single-end, 54bp RNA-seq which produced 6.3 million reads. Reads of RNA-seq samples were aligned by Bowtie, producing bam files that were imported into Partek Genomic Suite and normalized with annotation from Dm3 RefFlat. One-way ANOVA was used for differential gene expressions (DGEs) analysis of tubule/whole fly with FDR<0.01. An in-house Perl script was used to change the Affymetrix oligo ID to gene ID. Two data sheets were joined together in Partek Genomic suite, and a scatter plot was generated using Partek (Figure 3-13) for a total of 5593 genes.

The top 30 genes from tubule/whole fly (*Drosophila* Genome 2.0 Array) compared with *Drosophila* RNA-seq tubules/whole fly top 30 genes are listed in Appendix VII.

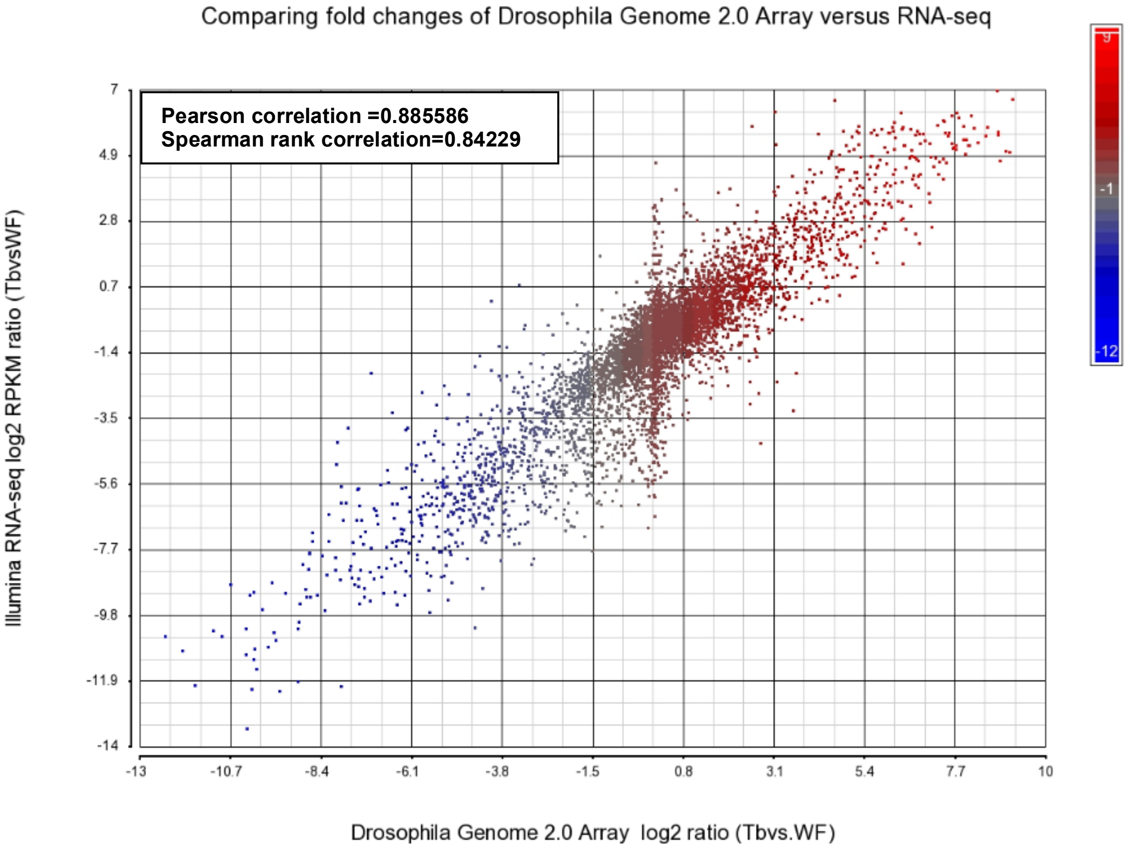


Figure 3-13 Comparison of estimated $\log_2$(folds change) for *Drosophila* tubule/whole fly from RNA-seq and *Drosophila* expression microarrays.

Only genes interrogated using both platforms were plotted. Genes with only one transcript, with unique mapped reads, and the unequivocal annotation were chosen for both platforms. The colour indicated the $\log_2$ (ratio) value, if $\log_2$ (ratio) from both platforms is greater than -12 and less than -1, the colour is blue. If $\log_2$ (ratio) from both platforms is greater than -1, the colour is red. The tubules *versus* whole fly were generated by Partek analysis, using a one-way ANOVA with FDR<0.01 and *Drosophila* Genome 2.0 annotation. The tubules and whole fly RNA-seq reads were aligned by Bowtie, and run in Partek analysis, using one-way ANOVA with FDR<0.01 and Dm3 RefFlat annotation. The middle red line of dots came from the very low read genes of RNA-seq (RPKM<4). The two platforms showed very strong correlation with a Pearson correlation coefficient 0.885586 and Spearman rank correlation 0.84229.

The scatter plot showed a strong but nonlinear correlation of the relative expression level of RNA-seq and Dros 2 with a Pearson correlation coefficient 0.885586 and Spearman rank correlation 0.84229. The middle red line of dots with high fold change in RNA-seq and low fold change in Dros 2 seem to be those genes with fewer reads. The strongest correlation came from those genes with highly mapped reads and high signals from Dros 2 Arrays.

24% of DGEs (folds change ≥ 3) were detected by both platforms from a total of 5593 genes. 44% of DGEs were detected by RNA-seq but not by expression microarrays, whilst 36% of DGEs were detected by expression microarrays but not by RNA-seq. RNA-seq detected more DGEs than arrays.

3.4.2.2 **Biological replicate samples of tubule/whole fly RNA-seq *versus Drosophila* tubule/whole fly tiling microarrays**

The tiling microarrays of *Drosophila* tubule/whole fly differential gene expression were analyzed in Partek Genomic Suite using a one-way ANOVA after the expression values were generated (see section 3.4.1.2). The corrected q-value is less than 0.001 for a gene to be called differentially expressed. The q-value cut is stricter than the differential expression analysis of RNA-seq and expression microarrays comparison because of the high false positive signals from tiling microarrays.

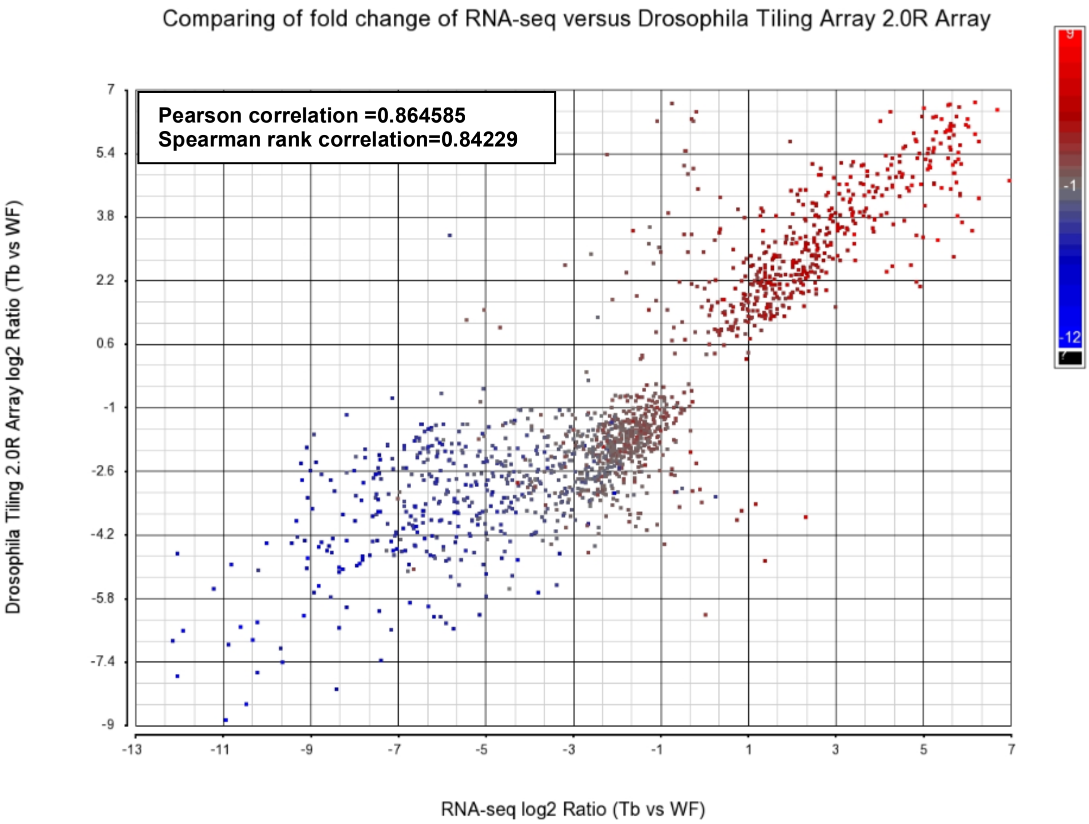


Figure 3-14 Comparison of estimated log$_2$ (folds change) *Drosophila* tubule/whole fly from RNA-seq and *Drosophila* tiling microarrays.

Comparison of both platforms was performed in Partek analysis using a one-way ANOVA analysis. Dros 2 (q≤0.01) with total 5593 genes and Tiling 2 (q≤0.001) with 3483 genes were applied for the analysis. The scatter plot showed a strong and linear correlation between the two platforms. The colour indicates the value of expression from both platforms. Higher values are shown in red, lower values are in blue. The black colour indicates the genes were present in one platform but absent in another platform because of the different cutting criteria in the comparisons.

The differential gene expression lists of RNA-seq (5593 differential expression genes) and tiling microarrays (3483 differential expression genes) were joined together by gene ID. The scatter plot was generated (Figure 3-14). The scatter plot of RNA-seq and Tiling 2 differential gene expressions between *Drosophila* tubule and *Drosophila* whole fly showed the strong and almost linear correlation between the two measurements with a Pearson correlation coefficient 0.864585 and Spearman rank correlation 0.84229 indicating that the ability for detecting the DGEs of the two platforms is similar but RNA-seq has a larger dynamic range than tiling microarrays. There was noise in the lower expression genes due to the cross-hybridization of tiling microarrays producing high-false positive signals. This also showed both platforms were limited in terms of detecting the lowly expressed genes. The higher expression genes were detected more by RNA-seq than tiling array due to the low dynamic range of tiling and saturation of the signal in tiling arrays.

3.4.2.3 **Biological replicate samples of *Drosophila* tubule/whole expression microarrays *versus Drosophila* tubule/whole fly tiling microarrays**

The tiling microarrays differential gene expression of *Drosophila* tubule and *Drosophila* whole fly were analyzed in Partek Genomic Suite using a one-way ANOVA after the expression values were generated (see section 3.4.1.2). The corrected q-value is less than 0.001 for a gene to be called differentially expressed. The q-value cut is stricter than the differential expression analysis of RNA-seq and Dros 2 comparison because of the high false positive signals from tiling microarrays. Expression microarrays tubule/whole fly fold changes were generated in Partek Genomic Suite (detailed in Section 3.4.2.1). An in-house Perl script was used to change the Affymetrix oligo ID to gene ID. Two data sheets of tubule/whole fly differential expression genes of tiling microarrays (3483 differential expression genes) and tubule/whole fly (5593 differential expression genes) of expression microarray were merged together in Partek Genomic Suite and a scatter plot was generated using Partek Genomic Suite (Figure 3-15).

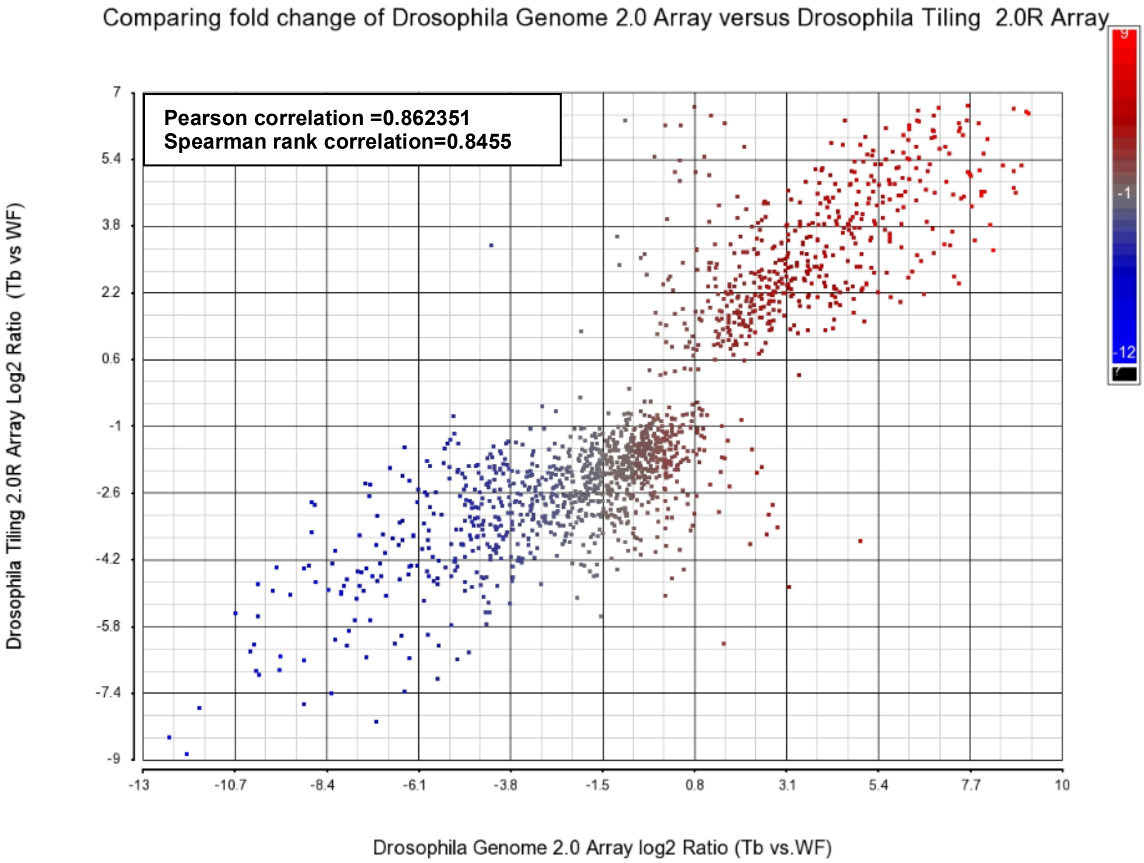


Figure 3-15 Comparison of estimated $\log_2$ (folds change) *Drosophila* tubule/whole fly from *Drosophila* expression microarrays and *Drosophila* tiling microarrays.

Comparison of both platforms was performed in Partek Genomic Suite using a one-way ANOVA analysis. Dros 2 ($q \leq 0.01$) with total 5593 genes and Tiling 2 ($q \leq 0.001$) with 3483 genes were applied for the analysis. The scatter plot showed a strong and linear correlation between the two platforms with a Pearson correlation 0.862351 and Spearman rank correlation 0.8455. The colour indicates the value of expression from both platforms with red indicating higher values and blue indicating lower values. The black colour indicates the genes were present in one platform but absent in another platform because of the different cutting criteria used in the comparisons.

The scatter plot of Dros 2 and Tiling 2 differential genes expression between *Drosophila* tubule and *Drosophila* whole fly showed a strong and almost linear correlation between the two measurements with a Pearson correlation coefficient 0.862351 and Spearman rank correlation 0.8455, indicating the ability to detect the DGEs of the two platforms is similar. There was noise for the higher expression genes and lower expression genes due to the cross-hybridization of both platforms and the signal saturation of both platforms.

3.4.2.4 **Venn diagram of *Drosophila* RNA-seq *versus Drosophila* tiling microarrays *versus Drosophila* expression microarrays**

Three differential expression genes (tubule versus whole fly) data sets from three platforms were imported into Partek Genomic Suite. Venn diagram was generated for the top 1000 differentially expressed genes between *Drosophila* RNA-seq, *Drosophila* tiling microarrays and *Drosophila* expression microarrays as shown in Figure 3-16.

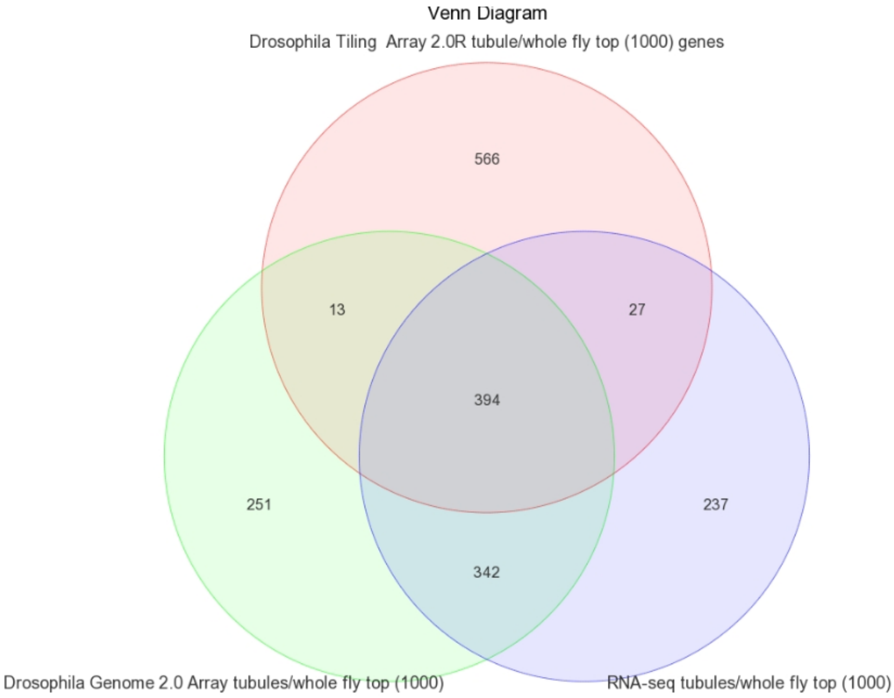


Figure 3-16 Venn diagram of top 1000 folds change genes.

Top 1000 genes were called differentially expressed by each platform (RNA-seq, *Drosophila* expression microarrays and *Drosophila* tiling microarrays). There was significant overlap (39.4%) between the three platforms, but more genes overlapped between RNA-seq and Dros 2 (73.6%) than between both RNA-seq and tiling arrays (40.7%), and between tiling microarrays and Dros 2 (42.1%). This likely reflects that RNA-seq and Dros 2 were more accurate for measuring the differential expression genes than tiling microarrays, and RNA-seq had greater dynamic range than tiling microarray and Dros 2.

The Venn diagram shows that 39.4% of those called as differentially expressed genes were detected by the three platforms. However, 59.3% genes of those called by tiling microarrays were not detected as differentially expressed genes by Dros 2; 57.9% of those called by tiling microarrays were not detected by RNA-seq as differentially expressed genes. RNA-seq and Dros 2 again showed strong agreement for detecting the differentially expressed genes, with 73.6% genes of those called differentially expressed genes detected by both platforms. However, still 26.4 % of those called by RNA-seq were not detected as differentially expression by Dros 2; 26.4% of those called by Dros 2 were not detected as differentially expressed by RNA-seq. Tiling microarrays behaved differently when detecting the differentially expressed genes than the other two platforms due to the high false-positive rate and the continuous probes measurement.

3.5 **Discussion**

3.5.1 ***RNA-seq compared with Drosophila expression microarrays***

Microarrays and RNA-seq are two popular methods to measure gene expression at the whole transcriptome level. Both technologies have their own merits and drawbacks. Through the comparison of biological replicate samples of *Drosophila* Genome 2.0 Array (Dros 2, three-prime expression microarrays) *versus* RNA-seq, we found groups of genes highly expressed in Dros 2, but undetectable by RNA-seq (table 3-1). In theory, RNA-seq is more sensitive than microarrays. 6-8 million mapped reads of RNA-seq provide adequate coverage to accurately estimate roughly 80-90% of the head transcriptome in flies (Malone and Oliver, 2011). The RNA-seq from this comparison had 6.3 million mapped reads, so it was enough to cover the expression of most genes. However, three-prime expression microarrays can demonstrate false-positive expression signals for several reasons.

Firstly, nonspecific probes could cross-hybridize to multiple genes (within a gene family or other similar sequence genes). If a genes shared 19bp or more in sequence identity, multiple genes might cross-hybridize with that set of genes (Zhang et al., 2005). We found probe-sets of four genes were cross-hybridized to multiple genes within the gene family in Table 3-1. For example, gene *Lcp65Ag2* (1640975_at), the blast search for the probe-set sequences in Ensembl genome browser, found that part of the probe-set sequences matched to *Lcp65Ag1-RA, Lcp65Ag3-RA, Lcp65Ae-RA,* and *Lcp65Af-RA* within the area. Similarly for gene *TWdIM*, we found part of probe-set sequences matched to *TWdIN-RA, TWdIH-RA, TWdIJ-RB, TWdIP-RA, TWdIB-RA, TWdIL-RB, TWdIO-RA, TWdIK-RA,* and *TWdIR-RA* within these gene families. This is the typical of the cross-hybridization within gene families that results in detecting false-positive signals, also affecting the genes expression and generating a number of false-positive expression signals.

Secondly, nonspecific probes could cross-hybridize to multiple genes across the genome. Genes *CG13705, CG17290.CG31909, CG10598* and *Sdic3* from Table 3-1 all cross-hybridize to other genes within the genome. *CG13750 (1632527_at)* cross-hybridized to three genes*1623412a_at, 1624625a _at,* and *1628360_at (CG17150)*. *CG17290 (1625558a_at)* cross-hybridized to two other genes namely

CG30458-RA and *CG10953*. *CG31909 (1638603_at)* cross-hybridized to *CG31909-RB, CG31909-RC, CG43800-RA* and *CG43800-RB*. *CG10598 (1633409a_at)* cross-hybridized to *CG14191-RA*. In fact, *CG14191-RA* was highly expressed in whole flies but RNA-seq only detected CG10598, which was not expressed in whole flies. *Sdic3 (1635695_at)* cross-hybridized to 40 transcripts of which the main ones are *1632213a _at*, *1628129a_at* and *1631477a _at*.

Thirdly, the unmatched annotation files between microarray and RNA-seq are another reason for the signal mismatch between the two measurements. *Drosophila* Genome 2.0 Array was designed in 2006 using the annotation from FlyBase version 5.3, RNA-seq used the annotation from FlyBase version 5.34. For the top 10 genes in Table 3-1, there was no case of unmatched gene annotation between the two platforms.

Taken together, Affymetrix three-prime expression microarrays contain thousands of redundant probe sets that interrogate different regions of the same genes, which can lead to inaccurate inference about overall gene expression (Cui and Loraine, 2009). Cross-hybridization is also very common in three-prime expression microarrays; extra filtering may be needed to get the right information about the gene expression for this type arrays.

RNA-seq has a wider dynamic range to measure the gene expression than microarrays. RNA-seq and three-prime expression microarrays are correlated very well. However, we still found a group of genes that have high signals in RNA-seq but were undetected by microarrays in Figure 3-13 (low end of RNA-seq) and Table 3-2. There are multiple reasons for this:

The first of these reasons is, missing the target transcript sequences on the Affymetrix GeneChip (Zhang et al., 2005). This is mainly caused by inaccuracy in annotation when the GeneChips were designed. In addition, there is a three-prime design bias for Affymetrix expression microarrays. For example in Table 3-2, probes sequences of *CG31084 (1627438_at)* was blasted in Ensembl, and these probes were matched to Chr3R 22253684-22253881 and Chr3R 22253974-22254091. Viewing *CG31084* in RNA-seq, we found the reads were mapped to Chr3R 22251024-22251483. Therefore, the Affymetrix expression microarrays probes for *CG31084* missed the target due to the wrong annotation of this gene.

Gene *CG34212 (1631103_at),* provides another example where half probe missed the gene target. The Affymetrix probes of *CG34212* matched two places of the genome Chr2R 1938179-1938515 and Chr2R 1937519-1937724. Only a small part of the probe Chr2R 1937519-1937724 mapped the gene *CG34212*. In RNA-seq, the reads mapped to Chr2R 1937431-1937645. The partial probes of arrays measured the gene expression, which caused the low signals of arrays. In a third example, gene *CG14309 (1631866_at),* the probes were designed on arrays missing the gene target. The probes of *CG14309* were designed between Chr3R 14206882-14207393, but the huge reads detected by RNA-seq were between Chr3R 14207636-14210725. The tiling microarrays also detected the high signal for the gene *CG14309* as $\log_2$ (signal intensity) 7.16421, further confirming that Dros 2 has a three-prime bias design. The probes on the arrays for this gene completely missed the transcribed region as a result of no expression being detected by arrays. These are typical examples of three-prime bias of three-prime expression arrays. For RNA-sequence however, mapping the genome doesn't depend on the annotation, and also RNA-seq has the potential ability to reannotate the genes structure to instruct accurate design of the probes for microarrays in the future. The new generation of gene arrays tried to make probes across the whole exons of the genes avoiding the three-prime bias, however these types of arrays are only currently available for human, mouse, rat and *Arabidopsis* but not for *Drosophila.*

Secondly, cross-hybridization problems in arrays affect the expression signals. Cross-hybridization can generate both false-positive expression signals and false-negative expression signals. Some probes on GeneChip were designed according to ESTs information, where by one gene maybe represented by several ESTs. Therefore two or more different probe sets are sometimes assumed to target the same genes or transcripts, leading to another cause of cross-hybridization (Bellis, 2013; Cui et al., 2010; Cui and Loraine, 2009). For example, gene *CG32212 (1641330-at)* was cross-hybridized to 18 transcripts. One of the transcripts *CR42842-RA* was a pseudogene, and *CG32212* also cross-hybridized to *CG12519-RA, CG12519-RB, 825-Oak-RB* and *CG18294-RA*. Another example, *CG3740 (1624552_at)* was cross-hybridized to the gene "*dor*" *(1623139_at).*

Thirdly, the wrong annotation maybe continued in FlyBase. Gene *CG31210 (1628546_at)* and *CG11042 (1641149_at)* had been withdrawn from the FlyBase

5.54. This might be another reason contributing to two measurements not being matched.

Fourthly, RNA-seq supports a wider dynamic range than microarrays for measuring genes at low and high expression levels. Microarrays have a more limited dynamic range in terms of gene expression level (Malone and Oliver, 2011; Marioni et al., 2008). For genes at low expression, microarrays suffered from background noise that affected the detection of capability. For highly expressed genes, microarrays suffered from signal saturation of the scanner. For example, the gene *Vas (1624413_at)* was a lowly expressed gene which was not detected by microarrays but was detected by RNA-seq.

Lastly, microarray probes were designed at the gene level, which was not suitable to measure the individual transcript expression (Bellis, 2013) whereas RNA-seq measured all transcripts and the averaged signals at the gene level. So the difference in measurement will cause the different gene expression level for some genes.

Taken together, RNA-seq has a number of advantages over microarrays. RNA-seq doesn’t require the genome information, and supports detection at the gene expression at a single-base resolution (Wilhelm et al., 2008); RNA-seq has a wider dynamic range to detect gene expression from low to high levels; RNA-seq can detect alternative splicing and novel transcripts (Wang et al., 2009; Young et al., 2012). Nevertheless, microarrays still remain useful and complementary (Kogenaru et al., 2012; Malone and Oliver, 2011) to RNA-seq to measure the gene expression at the transcriptome level.

3.5.2 *RNA-seq compared with Drosophila tiling microarrays*

Both RNA-seq and tiling microarrays are unbiased, high-throughput analytical tools for identifying novel RNAs, discerning alternative splicing isoforms, and determining gene expression level (Agarwal et al., 2010). From the comparison of RNA-seq and tiling microarrays Figure 3-14, the correlation between the two platforms is reasonably good. However, RNA-seq has distinct advantages over tiling microarrays for detecting the highly and lowly expressed genes.

Firstly, tiling microarrays suffered from cross-hybridization problems and hybridization background noise; it is not a suitable tool to detect genes at low expression levels. However, Figure 3-14 and Table 3-3 showed a group of signals that were highly expressed on tiling microarrays but were poorly expressed by RNA-seq. This is a problem in the cross-hybridization that is a known issue in tiling microarrays producing a lot of false-positive signals. Tiling microarrays have a higher false-positive rate than any other microarrays. This is the case because firstly, there are a number of pseudogenes within the genome. Duplicated pseudogenes arise when a genomic region containing a gene is copied and a copy is subsequently disabled. However, although pseudogenes are not transcribed, the pseudogenes and their parents have high sequence similarity (Agarwal et al., 2010). Furthermore, a number of probes on tiling microarrays that are highly similar to their nearest neighbours tend to be called as expressed by tiling microarrays and not detected by RNA-seq, providing strong evidence of cross-hybridization (Agarwal et al., 2010). Tiling microarrays also considerably overestimated the proportion of “dark matter” transcripts over RNA-seq due to the high false-positive rate in the detection of expression (van Bakel et al., 2010).

Secondly, RNA-seq has a wider dynamic range of detection than tiling microarrays which suffered from signal saturation by the scanner. That is illustrated, for example in Figure 3-14, where the top signals of tiling microarrays are compressed.

Thirdly, RNA-seq can detect the exon boundaries; therefore RNA-seq can clearly detect the alternative splicing signals. By comparing, tiling microarrays can detect the novel genes but have difficulty in detecting the fine structure of the genes. The novel genes that were found by tiling microarrays were difficult to confirm, as discussed for example in section 3.4.3.2.

Taken together, the comparison of RNA-seq and tiling microarrays demonstrated that most gene expression levels are well correlated between RNA-seq and tiling microarrays. RNA-seq clearly has advantages over tiling microarrays in detect exon boundaries, detecting alternatively splicing, and also has a wider dynamic range for measuring gene expression with a low false-positive rate. However, tiling microarrays remain cost effective for many species, and perform reasonably well with respect to expression levels (Agarwal et al., 2010), and can function as a tool to verify the RNA-seq results at the global level.

3.5.3 *Drosophila expression microarrays compared with Drosophila tiling microarrays*

Tiling microarrays and three-prime expression arrays are reasonable correlated when measuring the gene expression. However, the two data are fundamentally different. The tiling microarrays data are continuous whilst the three-prime expression microarrays data are discrete. This made the comparison a challenging task (Sasidharan et al., 2009), and may also result in the comparison not reflecting the real gene expression due to the basis for measurement.

Figure 3-15 showed the both platforms are not correlated for the highly and lowly expressed genes. There may be multiple reasons for this. Firstly, both platforms suffered from the hybridization background and the cross-hybridization affect. Secondly, both platforms suffered from signal saturation when measuring the highly expressed signals. Thirdly, tiling microarrays included probes for unannotated genes while three-prime expression arrays are only for detecting annotated genes.

Taken together, the comparison of tiling microarrays and three-prime expression microarrays demonstrated that most gene expression levels are well correlated. Three-prime expression arrays have advantages in detecting expression of known gene and tiling microarrays can detect novel genes and the transcription of “dark matters” (van Bakel et al., 2010) in the genome.

This thesis first demonstrated the three-way comparison of RNA-seq, tiling microarrays, expression microarrays that will be a valuable guide for the researcher in choosing suitable platforms for detecting gene expression in the future.

3.6 Conclusions

RNA-seq, expression microarrays and tiling microarrays are three popular methods to measure gene expression at the high-throughput whole transcriptome level. RNA-seq has more agreement with expression arrays than tiling microarrays. RNA-seq has a number of advantanges over microarrays; RNA-seq can measure gene expression without the genome reference; RNA-seq can detect the exon boundaries and detect the novel alternative splicing isoforms; RNA-seq can discover novel genes; RNA-seq has a wider dynamic range to for the measurement of gene expression. Expression arrays are still valuable for detecting the expression of known gene, and can be used to complement RNA-seq. Tiling microarrays are also a gene discovery tool but suffered from high false-positive rates. Data from tiling microarrays must be strictly cut in order to reduce the false-positive expression during the analysis, and so the data must be interpreted with cautions. Therefore, RNA-seq will be chosen as a tool to search the novel genes in *Drosophila* tubules in this thesis.

4. RNA-seq and directional RNA-seq for novel gene discovery in the transcriptome of *Drosophila* tubules

Summary

The next generation sequencing is a recently developed high-throughput method, which overcomes the limitations of previous sequencing methods and has the ability to produce an enormous volume of data cheaply (Franzen et al., 2013). The RNA-seq approach of next-generation sequencing avoids the need for bacterial cloning, can sequence the genome to a resolution of one base, and measures transcript expression by counting the reads corresponding to the RNA from each known exon, splice event or new candidate gene (Mortazavi et al., 2008). Thus RNA-seq method revolutionises the whole process of discovering novel genes and their variants at the genome level. This chapter describes the application of RNA-seq and strand-specific RNA-seq technologies to the discovery of tubule enriched novel genes in *Drosophila* and also confirmation of the novel gene by RT-PCR.

4.1 Introduction

4.1.1 *RNA-seq and novel gene discovery*

RNA-seq is a recently developed approach to transcriptome profiling that uses deep-sequencing technologies. Studies using this method have revealed a far greater complexity of eukaryotic transcriptome than was previously appreciated (Graveley et al., 2011). RNA-seq also provides a more precise measurement of the expression levels of transcripts and their isoforms, discovering the activities of novel coding and noncoding genes during transcription (Wang et al., 2009; Wilhelm et al., 2008).

There are two methods to produce the RNA-seq data for novel gene discovery.

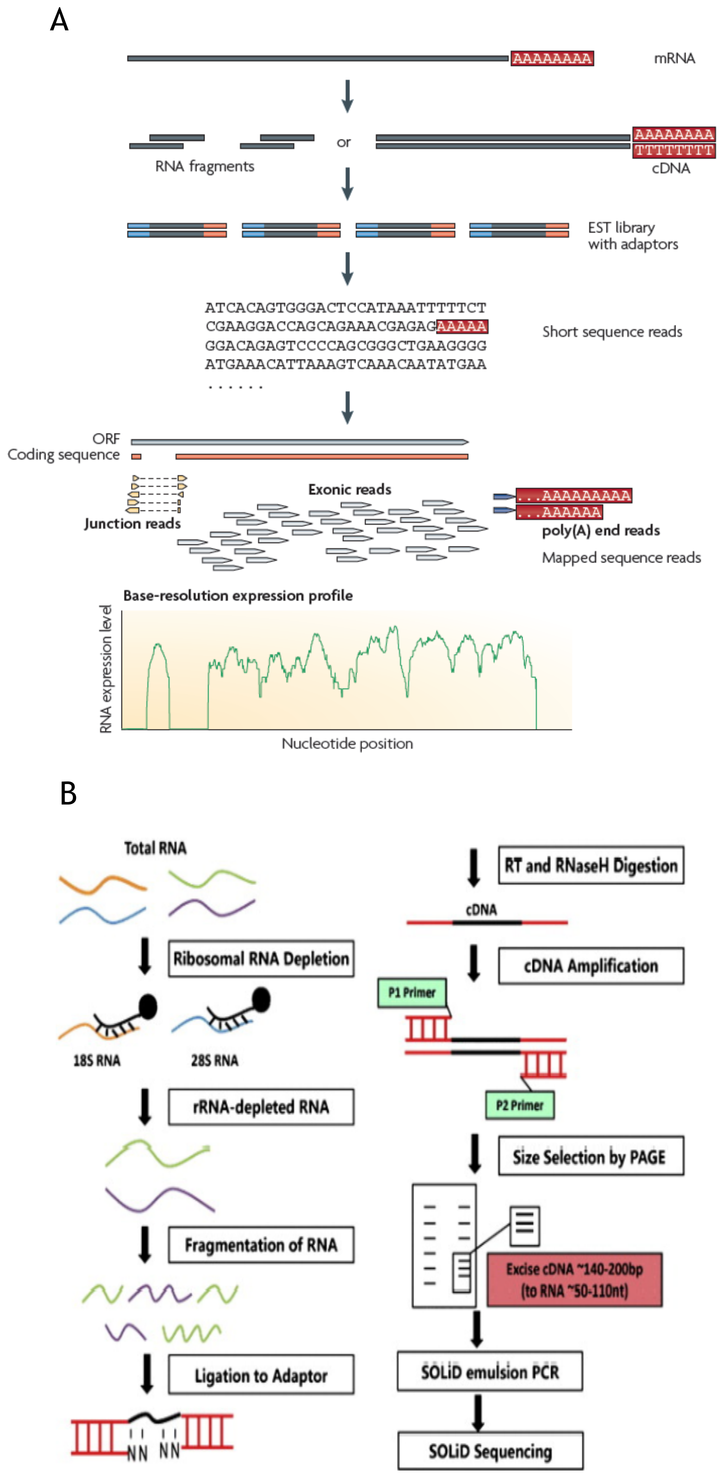


Figure 4-1 A typical RNA-seq experiment.

(A) Poly(A) selection for RNA-seq. mRNA is selected using oligo (dT) beads. RNA is then first converted into a library of cDNA fragments through either RNA or DNA fragmentation. Sequencing adaptors (blue) are subsequently added to each cDNA fragment and a short sequence is obtained from each cDNA using high-throughput sequencing technology. Picture adapted from (Wang et al., 2009). **(B) A flowchart of rmRNA-seq protocol for SOLid**. Ribosomal RNAs (coloured in orange and blue) are depleted with sequence-specific biotin-labelled probes and the remaining mRNA-rich fraction (green and violet) is fragmented with RNase III. After ligation to adaptors (red; NN stands for random oligonucleotide hexamers), the fragments in a size range of ~ 50 bp are collected and reverse-transcribed into a single-stranded cDNA library. The library is subsequently amplified, size-selected (140 to 200 bp), and sequenced in high coverage. Picture adapted from (Cui et al., 2010).

The first of these methods, which is illustrated in Figure 4.1 A, involves poly (A) selection followed by random priming to convert the RNA of interest to a cDNA library for high-throughput deep sequencing (Graveley et al., 2011; Vidal et al., 2013).

The second of these methods, which is illustrated in Figure 4.1 B, involves ribosome reduction of the total RNA followed by random priming to convert the RNA of interest to a cDNA library for further deep sequencing (Liu et al., 2011b).

The first method focuses on discovery of the novel RNAs with a poly(A) tail, such as coding RNA and mRNA-like non-coding RNA. The second method is suitable for discovery of all the coding and non-coding RNAs. However, different RNA sample preparations may result in significant variations in gene expression profiles (Cui et al., 2010; Tariq et al., 2011).

RNA-seq has clear advantages over previous methods. Firstly, RNA-seq is not limited to detecting transcripts that correspond to existing genomic sequence. The *de novo* assembly can sequence the genome and construct the transcriptome (Fan et al., 2013; Torales et al., 2013). This is very attractive in the case of non-model organisms for which the genome sequences are not determined, and for which no GeneChips are available. Secondly, RNA-seq measurement has a higher dynamic range of expression level. RNA-seq contains very low background signal, so it is able to measure genes with lower expression levels. RNA-seq also has no upper limit for quantification, and so can reveal the absolute level of gene expression (Graveley, 2008; Kogenaru et al., 2012; Wilhelm et al., 2008). Thirdly, RNA-seq can clearly detect the transcription boundaries, and so can detect both the novel junctions of novel isoforms and reveal completely new genes (Vidal et al., 2013; Wilhelm et al., 2008). This is a revolutionary tool for transcriptomics research (Mortazavi et al., 2008).

There are, however, still a number of challenges when using this technology (Wang et al., 2009). The library construction method currently requires fragmentation and amplification, which will introduce bias and artefacts into the system. Bioinformatic challenges include the need to store the large amount of data produced, the need for algorithms to identify high-quality reads, and the provision software for read mapping, read construction and expression quantification (Hitzemann et al., 2013).

4.1.2 *Directional RNA-seq and novel gene discovery*

Strand-specific, massively parallel cDNA sequencing (ssRNA-seq), also called directional RNA-seq, and is a powerful tool for transcript discovery, genome annotation and expression profiling (Franzen et al., 2013; Levin et al., 2010). The standard RNA-seq method does not provide information about which strand was originally transcribed, and therefore cannot distinguish overlapping transcription from two strands. Strand-specific RNA-seq is uniquely suited for novel gene discovery, especially for noncoding RNA discovery (Yassour et al., 2010; Zhu et al., 2013b). Studies reveal that most antisense transcripts may result from promiscuous bi-directional transcription in a dense genome, so strand-specific RNA-seq provides the opportunity to discover, for example, the long noncoding antisense RNAs that may not be detected using previous methods (Young et al., 2012).

Methods to construct the strand-specific RNA-seq libraries can be categorized into many classes. Methods in the first class rely on attaching different adaptors in a known orientation relative to the 5' and 3' ends of the RNA transcripts as illustrated in Figure 4.2 A. Methods in the second class as illustrated in Figure 4.2 B, rely on making one strand by chemical modification, either on RNA itself by bisulfite treatment or during second-strand cDNA synthesis followed by degradation of the unmarked strand (Levin et al., 2010). The experimental protocol used for the work reported in this thesis applies the first method according to the recommended protocol from Illumina.

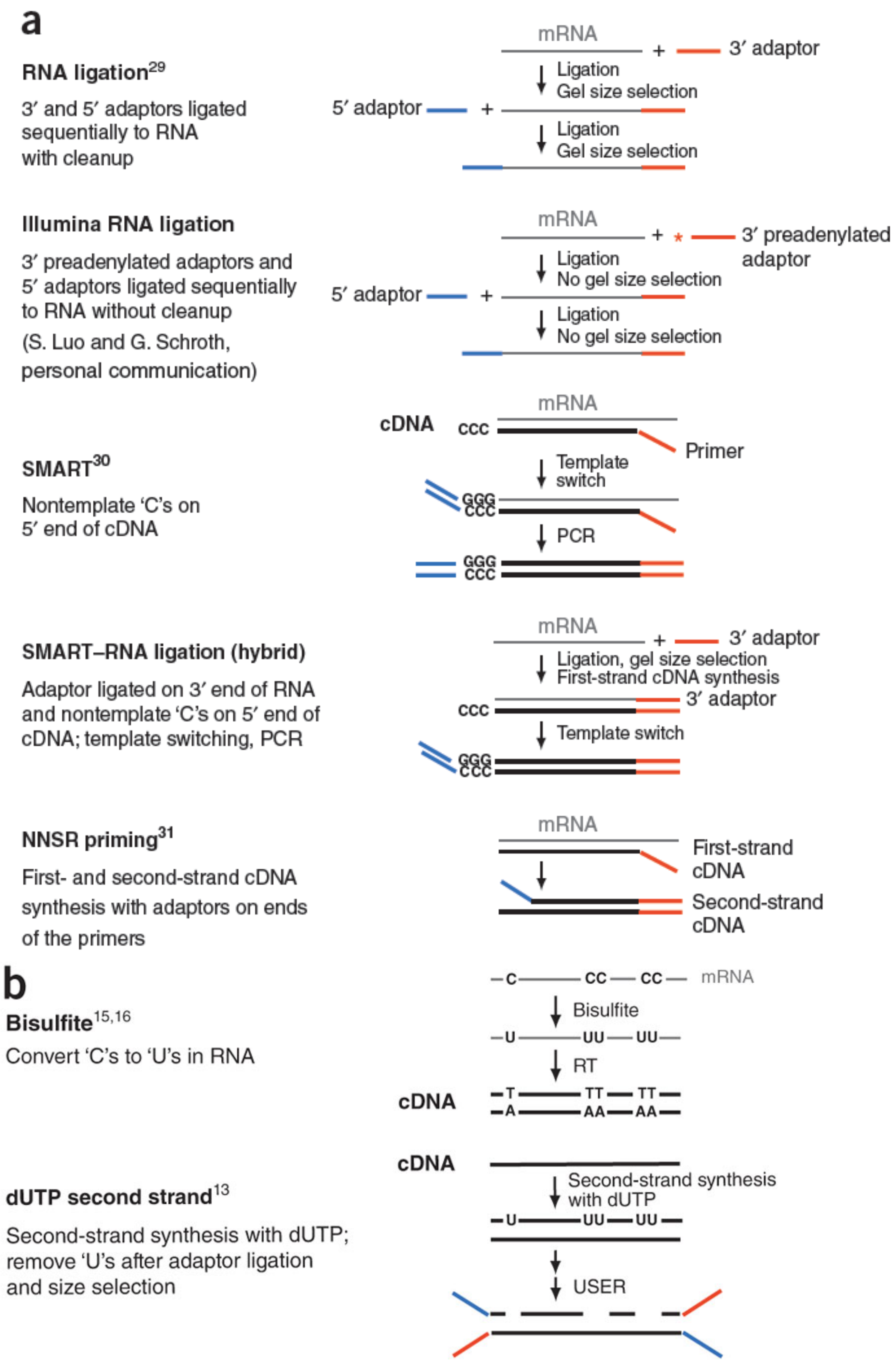

a
RNA ligation[29]
3′ and 5′ adaptors ligated sequentially to RNA with cleanup
mRNA
+
3′ adaptor
Ligation
Gel size selection
5′ adaptor
+
Ligation
Gel size selection
Illumina RNA ligation
3′ preadenylated adaptors and 5′ adaptors ligated sequentially to RNA without cleanup
(S. Luo and G. Schroth, personal communication)
mRNA
+ *
3′ preadenylated adaptor
Ligation
No gel size selection
5′ adaptor
+
Ligation
No gel size selection
SMART[30]
Nontemplate ‘C’s on 5′ end of cDNA
mRNA
cDNA
CCC
Primer
Template switch
GGG
CCC
PCR
GGG
CCC
SMART–RNA ligation (hybrid)
Adaptor ligated on 3′ end of RNA and nontemplate ‘C’s on 5′ end of cDNA; template switching, PCR
mRNA
+
3′ adaptor
Ligation, gel size selection
First-strand cDNA synthesis
CCC
3′ adaptor
Template switch
GGG
CCC
NNSR priming[31]
First- and second-strand cDNA synthesis with adaptors on ends of the primers
mRNA
First-strand cDNA
Second-strand cDNA
b
Bisulfite[15,16]
Convert ‘C’s to ‘U’s in RNA
C
CC
CC
mRNA
Bisulfite
U
UU
UU
RT
cDNA
T
TT
TT
A
AA
AA
dUTP second strand[13]
Second-strand synthesis with dUTP; remove ‘U’s after adaptor ligation and size selection
cDNA
Second-strand synthesis with dUTP
U
UU
UU
USER

Figure 4-2 Strand-specific RNA-Seq.

Strand-specific RNA-seq showing differential adaptor methods (A) and differential marking methods (B) mRNA is shown in grey, and cDNA in black. For differential adaptor methods, 5’ adaptors are shown in blue, and 3’ adaptors in red (Levin et al., 2010).

4.1.3 *Advantages and disadvantages of RNA-seq compared to directional RNA-seq*

RNA-seq allows analysis of all expressed transcripts, with three key goals in terms of structure annotation, expression quantification, and characterizing alternative splicing. Strand-specific RNA-seq offers improvements on standard RNA-seq with respect to these three goals as highlighted in Table 4-1, but incurs higher costs.

Table 4-1 Comparison of RNA-seq and strand-specific RNA-seq

	RNA-seq	**Strand-specific RNA-seq**
Advantages	Annotating the structures of all transcribed genes including their 5′ and 3′ ends and all splice junctions	Accurately identifying antisense transcripts
	Quantifying expression of each transcript	Determining the transcribed strand of non-coding RNAs (lincRNAs)
	Measuring the extent of alternative splicing	Demarcating the boundaries of closely situated or overlapping genes
Disadvantages	Cannot determine the polarity of RNA transcription	Costs much more than RNA-seq

4.1.4 *Units of expression measurement*

Quantifying the results of RNA-seq is much more complicated than doing so for microarrays. The sensitivity of RNA-seq will be a function of both molar concentration and transcript length. To take this into account, the unit measure of read density reflects the molar concentration of a transcript in the starting sample by normalizing for RNA length and for the total read number in the measurement (Mortazavi et al., 2008). The normalization method will facilitate transparent comparison of transcript levels both within and between samples (Mortazavi et al., 2008).

Two specific measures of reads density that are commonly used are RPKM and FPKM. RPKM indicates the Reads Per Kilobase of exon model per Million mapped reads. For example, a 1kb transcript with 1000 alignments in a sample of 10 million reads (out of which 8 million reads can be mapped) will have a RPKM=1000/ (1*8) =125. FPKM is used for pair-end sequencing, and indicates the Fragments Per Kilobase of exon per Million fragments mapped. A pair of reads constitutes one fragment.

4.1.5 *Analysis tools for RNA-seq*

A range of software analysis tools are available for use with RNA-seq data. The main software used for RNA-seq analysis is:

- CLC Genomic workbench (CLC bio), which can be used to discover the novel exons but offers limited functionality for discovering novel alternative splicing.

- Partek Genomics Suite, which can discover ‘unexpected regions’ including the novel genes, 3’and 5’ extensions and the splicing variants between the tissues. However this software offers limited functionality for discovery of the novel alternative splicing isoforms and transcripts discovery.

- TopHat and Cufflinks is freely available public domain software. This software can be used to generate pipelines that represent the best option for novel gene and novel alternative splicing discovery.

Many mapping tools have been developed since the introduction of RNA-seq, with the TopHat being among the most popular ones. TopHat aligns RNA-seq reads using the ultra high-throughput short reads aligner Bowtie, and then analyses the mapping results to identify known and novel splice junctions between exons as illustrated in Figure 4.3 A (Trapnell et al., 2009). For transcriptome reconstruction, the most commonly used tools are Cufflinks

(Trapnell et al., 2010) and Scripture (Guttman et al., 2010), both of which reconstruct a set of transcripts using reads mapped with TopHat. As shown in Figure 4.3 B, the Cufflinks package makes use of the components Cuffcompare and Cuffdiff for different gene expression detection and discovering novel genes (Trapnell et al., 2012; Trapnell et al., 2010).

Although other mappers, such as GSNAP (Wu and Nacu, 2010), have been described as more accurate than TopHat, they have never been used in combination with the transcriptome reconstruction tools. It is these transcriptome reconstruction tools that not only assemble the transcripts and estimate their abundances, but also search for the difference between assemblies and references in order to discover novel genes and test for differential expression and regulation in RNA-seq samples as mentioned previously (Palmieri et al., 2012). In this sense, the TopHat and Cufflinks pipeline is the only pipeline so far that performs all the analysis together.

However, TopHat and Cufflinks do not address all applications of RNA-seq, nor they are only tools for RNA-seq analysis. TopHat and Cufflinks require a sequenced genome reference. The software has been designed for use specifically with data formatted from either Illumina or SOLiD sequencing machines. In addition, it can be difficult to distinguish full-length novel transcripts from partial fragments using RNA-seq alone. As a consequence, the results obtained need to be validated by traditional cloning and PCR-based techniques, or validation of transcript ends by rapid amplification of cDNA ends (RACE) to rule out incomplete reconstruction due to gaps in sequencing coverage (Trapnell et al., 2012).

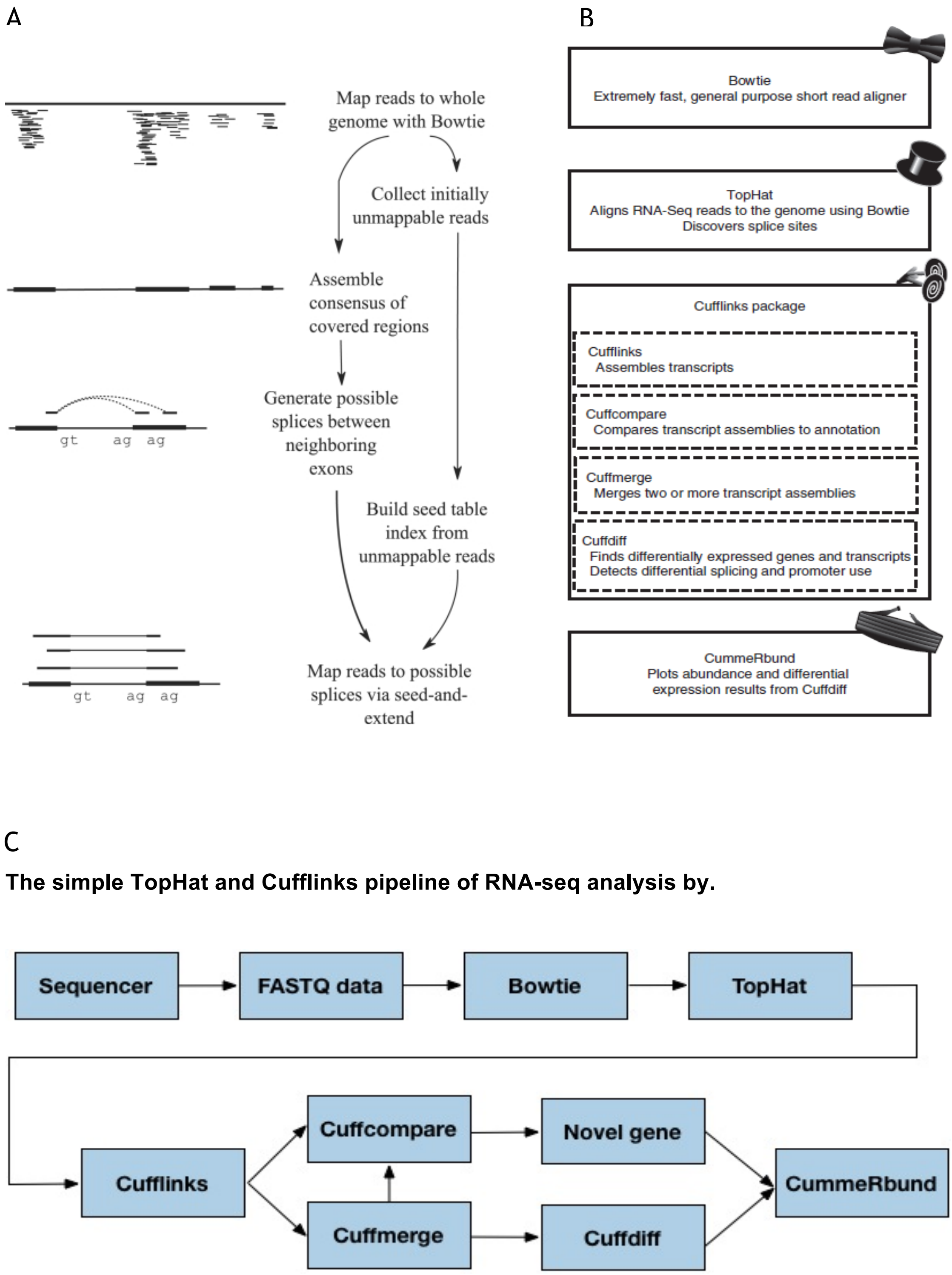

A
Map reads to whole genome with Bowtie
Collect initially unmappable reads
Assemble consensus of covered regions
Generate possible splices between neighboring exons
Build seed table index from unmappable reads
Map reads to possible splices via seed-and-extend
gt ag ag
gt ag ag
B
Bowtie
Extremely fast, general purpose short read aligner
TopHat
Aligns RNA-Seq reads to the genome using Bowtie
Discovers splice sites
Cufflinks package
Cufflinks
Assembles transcripts
Cuffcompare
Compares transcript assemblies to annotation
Cuffmerge
Merges two or more transcript assemblies
Cuffdiff
Finds differentially expressed genes and transcripts
Detects differential splicing and promoter use
CummeRbund
Plots abundance and differential expression results from Cuffdiff
C
The simple TopHat and Cufflinks pipeline of RNA-seq analysis by.
Sequencer
FASTQ data
Bowtie
TopHat
Cufflinks
Cuffcompare
Cuffmerge
Novel gene
Cuffdiff
CummeRbund

Figure 4-3 The TopHat and Cufflinks pipeline.

(A) The TopHat pipeline. RNA-Seq reads are mapped against the whole reference genome, and those reads that do not map are set aside. An initial consensus of mapped regions is computed by Maq. Sequences flanking potential donor/acceptor splice sites within neighbouring regions are joined to form potential splice junctions. Picture adapted from (Trapnell et al., 2009) **(B) The TopHat and Cufflinks pipeline**. TopHat uses Bowtie to align the reads and TopHat discovers splice sites. Cufflinks assembles the transcripts. CummeRbund views the image. Picture adapted from (Trapnell et al., 2012). **(C) A simple TopHat and Cufflinks workflow for RNA-seq analysis**.

4.1.6 ***RNA-seq in Drosophila research***

The rapid development of RNA-seq has led to several research groups using this technology to investigate the transcriptome of *Drosophila*. Whilst this has included looking for the novel transcripts, novel alternative splicing but most research focused on development in *Drosophila*.

One of the key research projects at present is the National Human Genome Research Institute (NHGRI) model organism ENCyclopedia Of DNA Elements (modENCODE). A principal goal of this project is to go beyond the annotations and identify previously unannotated transcripts in *Drosophila*. Two papers have been published recently by modENCODE. One of these papers investigated the transcriptome of 27 distinct stages of development of *Drosophila melanogaster*. This project identified 1,938 new transcribed regions not linked to any annotated gene model (Graveley et al., 2011). The second paper reported on an investigation into transcriptional diversity in 25 *Drosophila* cell lines. This second project identified 1,405 novel transcribed regions; 684 of these appear to be new exons of neighbouring, often distant, genes. Another *Drosophila* RNA-seq research project investigated 10 *Drosophila* developmental stages by using paired-end RNA sequencing (Daines et al., 2011). In this study, a total of 319 novel transcripts were identified, representing a 2% increase over the current level of annotation. Yet another group reanalysed modENCODE data by redeveloping the TopHat program, subsequently identifying 1,119 lincRNA loci in the *Drosophila melanogaster* genome (Young et al., 2012).

Given their distinct focus on development rather than differentiated adult tissues, these recently publications can form a useful complement to the work presented in this thesis. Our research principally focuses on the discovery of novel transcripts, genes, exons and alternative splicing in specific adult tissues of *Drosophila melanogaster*. We have generated the RNA-seq data from heads, testes, whole flies and tubules. In particular, my project is concerned mainly with studies of tubules. We can use the data from modENCODE to help us confirm our results, and our results in turn can be integrated with those of the modENCODE project for the benefit of the wider research community.

4.2 RNA-seq and Strand –specific RNA-seq (ssRNA-seq) experiment design

The aim of this experiment using RNA-seq analysis was to identify novel genes in *Drosophila* tubules. Four biological replicates of whole flies (Canton S), three biological replicates of tubules, three biological replicates of heads, three biological replicates of testes of Canton S flies were sequenced by the Illumina Genome Analysis System (GAIIx) using RNA-seq technology in order to obtain the fold-change of tubule/whole flies, testes/whole flies and heads/whole flies and find out the tubule-enriched, testes-enriched and head-enriched novel genes of *Drosophila* melanogaster. Because the RNA-seq was just started when this project was set up, so the ‘pilot’ experiment was from one sample of each tissue (whole flies, testes, tubules and heads) generated by single-end reads of 54bp. Further replicates of the four tissues were generated by single-end reads of 76bp as technology developed between the two experiments.

The aim of the experiment using strand-specific RNA-seq was to identify the direction of novel transcripts and thereby increase confidence in the findings of the RNA-seq results. Due to the disadvantages of RNA-seq which can’t identify the direction of the strand, however the information of the strand is important to identify novel genes and this project focused on searching novel genes in tubules, so one sample of tubule (Canton S flies) was analyzed by the strand-specific RNA-seq technology to help verifying the novel tubule-enriched genes which found by RNA-seq technology.

4.3 Results

4.3.1 *Transcript Classification Scheme in Cufflinks*

The ‘class_code’ is a classification alphabet designed for Cufflinks pipeline to represent the status of transcripts compared with the reference genome. The set of class_codes and their representation in Cufflinks is listed in table 4-2. These classifications were used in the analysis pipeline developed for this project.

Table 4-2 Class_code and its representation in Cufflinks

Priority	Code	Description
1	=	Complete match of intron chain
2	c	Contained
3	j	Potentially novel isoform (fragment): at least one splice junction is shared with a reference transcript
4	e	Single exon transfrag overlapping a reference exon and at least 10 bp of a reference intron, indicating a possible pre-mRNA fragment.
5	i	A transfrag falling entirely within a reference intron
6	o	Generic exonic overlap with a reference transcript
7	p	Possible polymerase run-on fragment (within 2Kbases of a reference transcript)
8	r	Repeat. Currently determined by looking at the soft-masked reference sequence and applied to transcripts where at least 50% of the bases are lower case
9	u	Unknown, intergenic transcript
10	x	Exonic overlap with reference on the opposite strand
11	s	An intron of the transfrag overlaps a reference intron on the opposite strand (likely due to read mapping errors)
12	.	(.tracking file only, indicates multiple classifications)

Table adapted from the online Cufflinks user manual available at http://cufflinks.cbcb.umd.edu/manual.html. Each ‘class_code’ is characterized in terms of its **priority** meaning the ordering used to assign class codes in the case when multiple classifications are possible (low number indicate higher priority), its assigned **code** symbol, and a textual **description**.

Novel tissue-enriched genes were identified using the data analysis pipeline described in Figure 4-4.

4.3.2 *Analysis pipeline to find tissue-enriched genes*

The single-end 54bp (one replicate) and 76bp (two-replicates) sequence samples from the Illumina GAIIx were cleaned using scripts developed in-house to remove the adaptors, then the FASTQ files were exported for analysis in TopHat. TopHat aligned the reads to the genome using Bowtie (Langmead et al., 2009), and TopHat (Trapnell et al., 2009) also generated a database of the possible splicing junctions, and then mapped the reads against these junctions to confirm them. Next, the unmapped reads were splitted and remapped to the database of the splicing junctions. The known splicing junctions and novel splicing junctions were found.

Cufflinks was used to assemble all the possible transcripts and build new combined annotations. Cuffcompare was then used to compare the resulting combined annotations with the FlyBase reference annotation to identify the novel isoforms (class_code ‘j’) and novel gene (class_code ‘u’).

Cuffmerge was used to merge all the transcripts involved in the comparison in the project. Cuffcompare could also take the entire merged file to compare the reference annotation files to find all the novel genes and isoforms in the entire merged files.

Cuffdiff was run to perform a comparison across the merged files to find the enriched genes in specific tissues that met the criteria of a fold change greater than 3 and p-value less than 0.05, and a false positive rate value (q value) less than 0.05 (statistic as detailed in Section 4.3.7.1). The types of tissue-specific enriched genes were indicated by different class codes. The Venn diagram helped to identify the overlapped genes in heads, testes and tubules, and then find the possible tissue specific novel genes.

In order to eliminate potential for genomic contamination, those candidates in the tissue-specific novel gene list that have a single exon were not considered as novel genes unless they were conserved in other species. The candidate novel genes with two exons are more likely to be novel genes and not genomic contamination. In addition, the entire candidate novel genes had to be supported by strand-specific RNA-seq results (ssRNA-seq) to make the final list of novel genes due to the fact that the RNA-seq was not direction-aware, and the information from RNA-seq came from both forward and reverse strands. Some of this information may be inaccurate, especially in the case of two genes overlapped from different directions or partially overlapped genes from both strands.

The process of how to select tissue-specific novel genes using TopHat and Cufflinks pipeline was also listed in Figure 4-4.

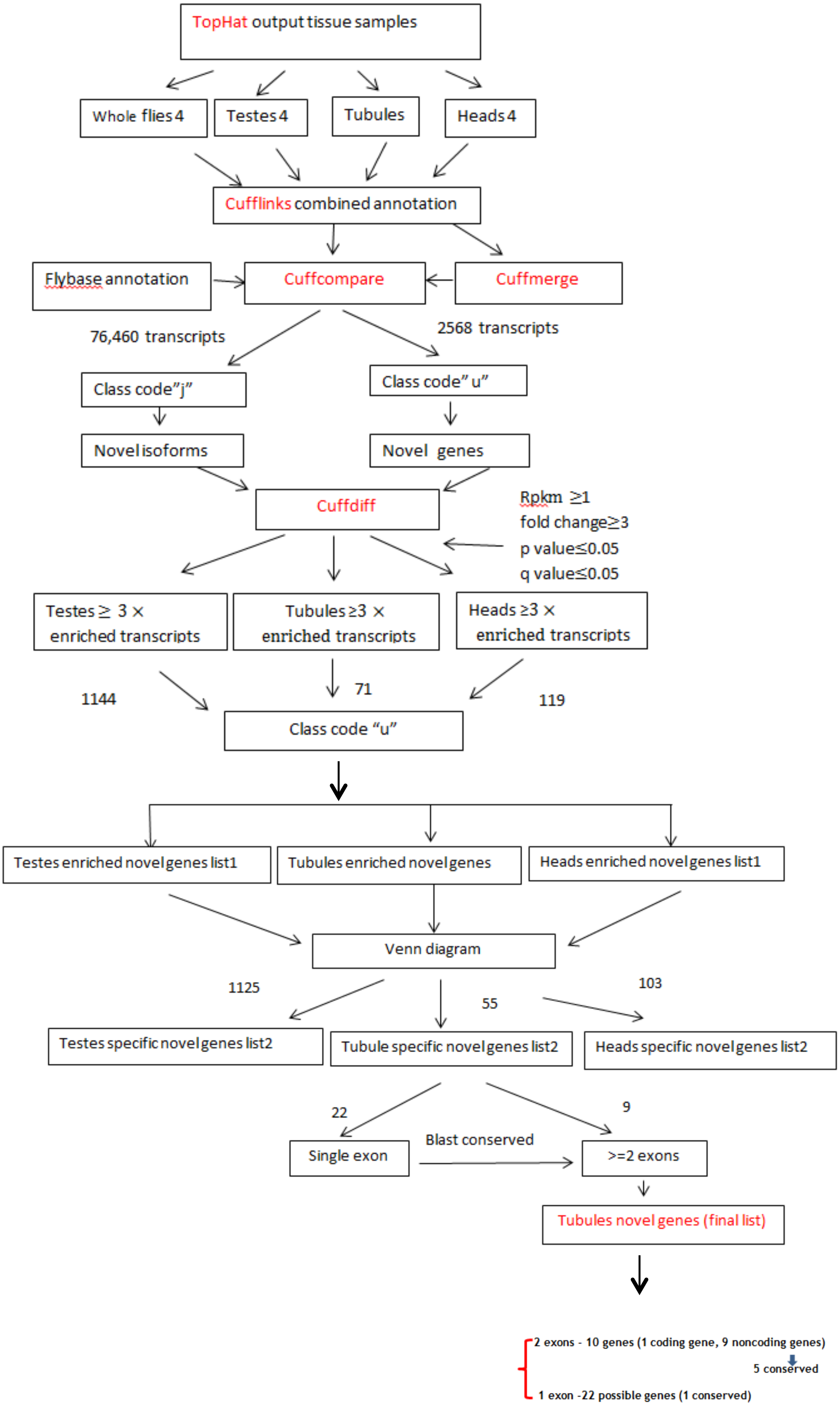

TopHat output tissue samples
Whole flies 4
Testes 4
Tubules
Heads 4
Cufflinks combined annotation
Flybase annotation
Cuffcompare
Cuffmerge
76,460 transcripts
2568 transcripts
Class code"j"
Class code" u"
Novel isoforms
Novel genes
Cuffdiff
Rpkm ≥1
fold change≥3
p value≤0.05
q value≤0.05
Testes ≥ 3 × enriched transcripts
Tubules ≥3 × enriched transcripts
Heads ≥3 × enriched transcripts
1144
71
119
Class code "u"
Testes enriched novel genes list1
Tubules enriched novel genes
Heads enriched novel genes list1
Venn diagram
1125
55
103
Testes specific novel genes list2
Tubule specific novel genes list2
Heads specific novel genes list2
22
9
Single exon
Blast conserved
>=2 exons
Tubules novel genes (final list)
2 exons - 10 genes (1 coding gene, 9 noncoding genes)
5 conserved
1 exon -22 possible genes (1 conserved)

Figure 4-4 The analysis pipeline using TopHat and Cufflinks to find novel tissue enriched genes.

Tubules, testes, heads and whole fly single-end 54bp and 76bp RNA-seq results imported into TopHat. TopHat aligned the reads to exons and splice junctions using Bowtie. TopHat results were assembled into transcripts by Cufflinks. Cuffcompare used cufflinks results to compare with FlyBase annotation to find new isoforms (76,640), new genes located intergenic regions (2,568). Cuffdiff used Cufflinks results to find the novel tissue enriched genes (1,144 in testes, 71 in tubules, and 119 in heads), the filter was also applied as gene expression level RPKM>1, folds-change ≥3, p value<0.05, q value <0.05. Venn diagram to choose the nonoverlapping tissue specific novel genes (1,124 in testes, 55 in tubules, 103 in heads) with more than two exons were considered novel genes (10 genes), 22 single exons may be possible novel genes, they are conserved in *Drosophila* species. Note that new isoforms are assigned the class_code 'j', and new genes are assigned the class_code 'u'.

4.3.3 *Filtering of reads produced by Illumina GAIIx*

The samples were run on Illumina GAIIx, and then the proprietary software within GAIIx was used to filter out high quality reads. A script was developed in house to remove the adaptors, clean the reads producing Fastq files suitable for importing into TopHat.

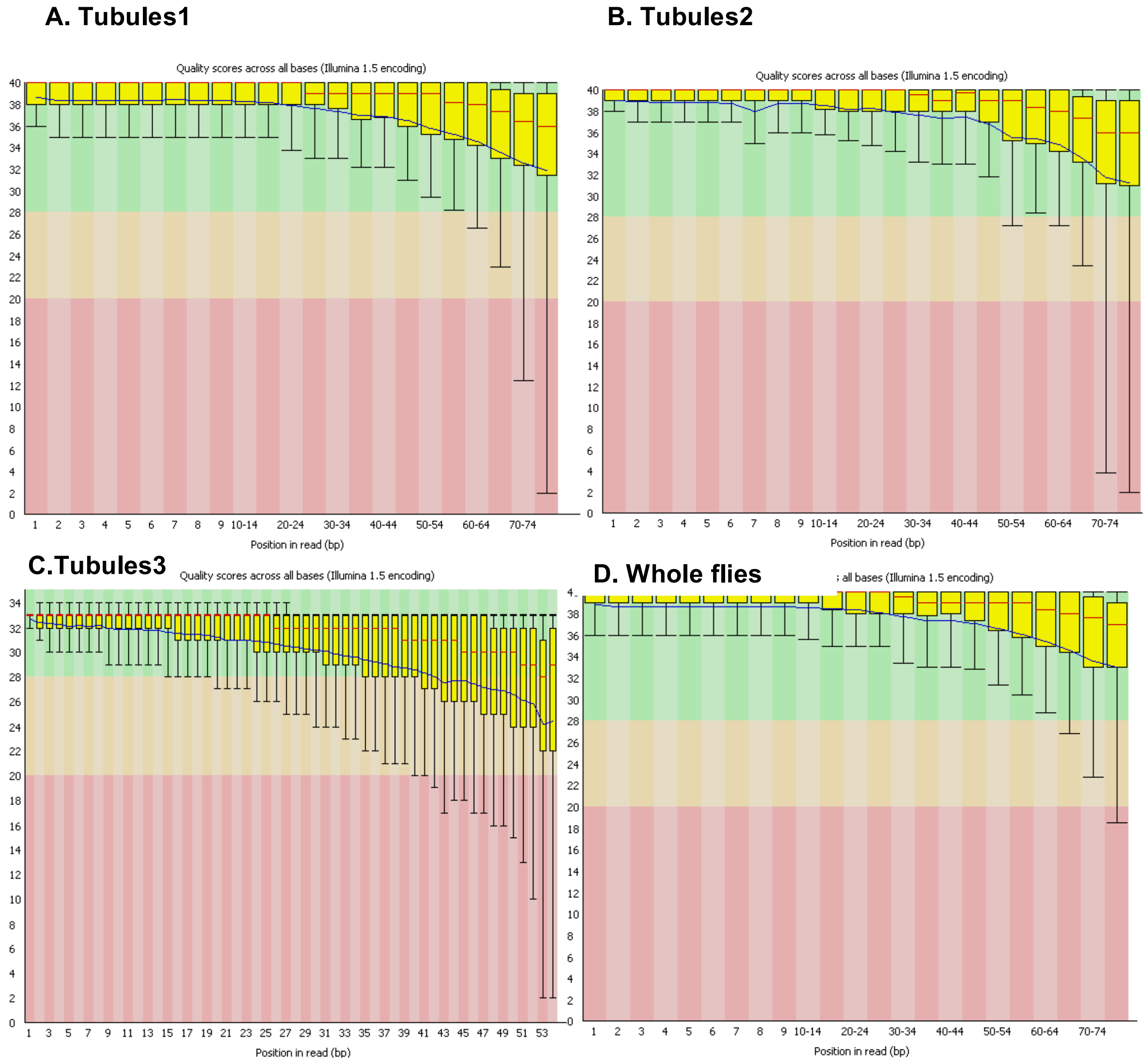
Table 4-3 Filtered reads produced by Illumina GAIIx

Flow cell	Lane	sample	Read length	Number of PF reads	Number of Clean reads	% clean
FC058	S2	Tb1	76	32,553,488	32,057,244	98.48
FC058	S3	Tb2	76	26,528,842	26,232,304	98.88
FC006	S1	Tb3	54	5,473,207	5,449,066	99.56
FC058	S4	Hd1	76	29,509,968	29,026,152	98.36
FC058	S5	Hd2	76	30,700,445	30,346,171	98.85
FC006	S4	Hd3	54	6,092,923	6,077,917	99.75
FC063	S7	Ts1	76	29,486,131	29,406,000	99.73
FC055	S4	Ts2	76	30,146,737	29,951,494	99.35
FC006	S3	Ts3	54	6,017,042	5,988,635	99.53
FC058	S6	Wf1	76	30,299,951	29,908,345	98.71
FC058	S7	Wf2	76	30,395,145	30,054,672	98.88
FC006	S2	Wf3	54	63,155,32	6,177,362	97.81
FC015	S4	Wf4	76	13,687,600	13,567,308	99.12
FC053	S5	Tb3 direct	76	27,076,624	26,928,326	99.45

Tb tubules **Hd** heads **Wf** whole flies **Ts** testes **Tb direct** tubules strand-specific RNA-seq. **Flow cell** tells which flow cell the samples amplified on. **Lane** indicates the position of the samples on the flow cell. **Read length** indicates the sequence length. **Number of PF reads** means the number of the reads past the filter of GAIIx proprietary software. **Number of cleaned reads** means the reads generated by house-made script after removed the adaptors. **% clean** the percentage of the clean to pass filter reads.

Table 4-3 lists the set of filtered reads produced by the Illumina GAIIx for samples used in this project.

4.3.4 *Checking quality of RNA-seq data using FastQC*


A. Tubules1

B. Tubules2

C.Tubules3

D. Whole flies

Figure 4-5 FastQC quality control of RNA-seq samples.

The central red line represents the median value. The yellow box represents the inter-quartile range (25-75%). The upper and lower whiskers represent the 10% and 90% points. The blue line represents the mean quality. The y-axis on the graph shows the quality scores also called the Phred quality scores. The background of the graph is divided along the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). (A) (B) (D) Tubule and whole fly sequences of read length 76bp. From base 1 to 76, the mean Phred quality score is above 28 indicating probability of an incorrect base call is 1 in 1000. The base call accuracy is 99.9%. Graph showed most base calls are very good quality (in green) (C) Tubule sequences of read length 54bp. From base 1 to 40, the mean Phred quality score is above 28 indicating the probability of an incorrect base call is 1 in 1000. The base call accuracy is 99.9% and calls are very good quality (green), from base 42 to 54, the mean Phred quality score drops to 24 indicating the probability of an incorrect base call is 1 in 100. The base call accuracy is 99 % and calls are of reasonable quality (orange).

FastQC offers a simple quality control checks aimed at providing a QC report that can spot problems which originated either in the sequencer, or in the starting library material.

The FastQC analysis was performed by a series of analysis modules. Figure 4.5 presents the results of the Per Base Sequence Quality mode. This view shows the range of quality values across all bases at each position in the FastQ file using a Box Whisker type of plot.

The quality scores also called Phred quality scores which were originally developed by the program Phred to help in the automation of DNA sequencing in the Human Genome Project (Ewing and Green, 1998; Ewing et al., 1998). Phred quality scores have become widely accepted to characterize the quality of DNA sequences, and can be used to compare the efficacy of different sequencing methods. Phred quality can be used to automatically determine accurate, quality-based consensus sequences. Higher Phred scores indicate better base call as listed in Table 4-4. The maximum Phred quality is 40 in Illumina.

Table 4-4 Phred quality scores and their interpretation

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

Phred quality scores are logarithmically linked to error probabilities. Phred quality score 10 indicates of an incorrect base call is 1 in 10 and the base call accuracy is 90%. Phred quality score 20 indicates of an incorrect base call is 1 in 100 and the base call accuracy is 99%. Phred quality score 30 indicates of an incorrect base call is 1 in 1000 and the base call accuracy is 99.9%. Phred quality score 40 indicates of an incorrect base call is 1 in 10000 and the base call accuracy is 99.99%. Phred quality score 50 indicates of an incorrect base call is 1 in 100000 and the base call accuracy is 99.999%. The higher score indicates the better quality of the sequence.

The background of the graph shown in Figure 4-5 divided along y axis in order to highlight very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). The quality of calls on most platforms degrades as the run progresses, so it was common to see base calls falling into the orange area towards the end of a read.

4.3.5 *Aligning reads using TopHat and Bowtie*

The Glasgow University Polyomics built TopHat and Cufflinks pipeline was used in this analysis.

TopHat (version 1.3.0) was firstly to use the clean Fastq reads files (from Table 4-3 Number of clean reads) to align the reads to the exons using Bowtie which assigned at least 12bp bases on each side of the junction by this project. Table 4-5 summarizes the results of this alignment.

Table 4-5 Reads mapped to exons and junctions, produced by TopHat

Sample	Reads processed	Reads mapped to the exons	Reads mapped to the junction
Tb1	32,006,864	25,247,267	3,191,986
Tb2	26,194,960	20,676,854	2,831,559
Tb3	5,445,631	4,5694,00	401,047
Hd1	28,973,336	22,811,201	3,205,868
Hd2	30,297,540	23,889,196	3,150,896
Hd3	6,075,507	5,099,139	466,779
Ts1	29,124,920	22,072,809	2,636,754
Ts2	29,926,559	22,863,205	2,768,407
Ts3	5,984,409	5,050,273	355,182
Wf1	29,867,435	24,185,549	2,601,356
Wf2	30,003,785	24,162,772	2,844,382
Wf3	6,173,774	5,089,657	401,047
Wf4	13,508,858	90,118,51	1,341,704

Tb tubules **Hd** heads **Wf** whole flies **Ts** testes. **Reads processed** were the clean reads from GAIIx after being filtered. **Reads mapped to the exons** were the reads mapped to exons by TopHat. **Reads mapped to the junctions** were total reads mapped to the known and putative junctions

The unmapped reads were put aside. TopHat was next used to build the index of all possible spliced junctions including all the known and putative junctions recoded in the annotated data from Flybase 5.36 (www.flybase.org) by using the aligned exons. TopHat used all the discarded reads to map this newly built junction database (see Table 4-5 Reads mapped to the junctions), and then the novel junctions could then be found by a process of comparison after subtracting the known junctions. The final output of TopHat is an alignment file in BAM format (accept_hits.bam) and lists of splice junctions and indels in BED format ready for visualization in a genome brower.

4.3.6 *Discovering tissue-specific novel genes using Cufflinks*

Cufflinks (version 1.0.3) was run for each sample separately. Transcripts were assembled based on the TopHat alignment files (accepted_hits.bam), using existing genome annotations recodes in FlyBase 5.36 but also allowing for novel transcripts. Output: transcripts.gtf (Cufflinks assembled isoforms); Isoforms.fpkm_ tracking (isoform level expression value RPKM); genes.fpkm_ tracking (gene level expression value RPKM). This produced a list of all the generated transcripts.gtf which was stored in the file 'assemblies.txt' for every sample.

Cuffmerge was used to convert the input from files gtf to sam format and then merged Cufflinks generated transcripts .gtf files (also specified in the list as 'assemblies.txt' for each tissue) into a single merged.gtf fil**e.** The output files were: transcripts.gtf; isoform.fpkm_tracking; genes.fpkm_tracking for all tissue samples.

The transcripts.gtf file was compared against the reference genome annotation file by Cuffcompare and the final **merged.gtf** file was generated (examples of merged_gtf file is shown in Table 4-6). This merged file contained newly built gene_id (XLOC_...), transcript_id (TCONS_...), exon start and end point, and transcript class_codes indicating the possible type of the transcripts. Cuffcompare was used to discover the novel genes or transcripts in merged.gtf file that were located in intergenic regions (class_code "u") or fell into the intronic regions (class_code "i"). However, the novel genes discovered were not necessarily tissue-specific. Cuffdiff allowed the identification of tissue-specific novel genes. The example of a merged.gtf file with class_code is shown in Table 4-6. The summary of all the class_codes of the entire tissues is listed in Table 4-7.

Table 4-6 Example of an excerpt from a merged.gtf file with the reported class_code

chr2L	Cufflinks	gene_id "XLOC_000001"; transcript_id "TCONS_00000001"; exon_number "1"; class_code "j"; tss_id "TSS1";
chr2L	Cufflinks	gene_id "XLOC_000001"; transcript_id "TCONS_00000001"; exon_number "2""; class_code "j"; tss_id "TSS1";
chr2L	Cufflinks	gene_id "XLOC_000001"; transcript_id "TCONS_00000001"; exon_number "3"; class_code "j"; tss_id "TSS1";
chr2L	FlyBase	gene_id "XLOC_000001"; transcript_id "TCONS_00000003"; exon_number "1"; class_code "="; tss_id "TSS1";
chr2L	FlyBase	gene_id "XLOC_000001"; transcript_id "TCONS_00000003"; exon_number "2" class_code "="; tss_id "TSS1";
chr2L	Cufflinks	gene_id "XLOC_000002"; transcript_id "TCONS_00000004"; exon_number "1" class_code "s"; tss_id "TSS2";
chr2L	Cufflinks	gene_id "XLOC_000002"; transcript_id "TCONS_00000004"; exon_number "2" class_code "s"; tss_id "TSS2";
chr2L	Cufflinks	gene_id "XLOC_000014"; transcript_id "TCONS_00000051"; exon_number "1"; class_code "x";
chr2L	Cufflinks	gene_id "XLOC_000015"; transcript_id "TCONS_00000052"; exon_number "1"; class_code "o";
chr2L	Cufflinks	gene_id "XLOC_000806"; transcript_id "TCONS_00001671"; exon_number "1"; class_code "u"; tss_id "TSS984";
chr2L	Cufflinks	gene_id "XLOC_000806"; transcript_id "TCONS_00001671"; exon_number "2"; class_code "u"; tss_id "TSS984";
chr2L	Cufflinks	gene_id "XLOC_000831"; transcript_id "TCONS_00001706"; exon_number "1"; class_code "x"; tss_id "TSS1008";
chr2L	Cufflinks	gene_id "XLOC_000857"; transcript_id "TCONS_00001767"; class_code "u"; tss_id "TSS1049";
chr2L	Cufflinks	gene_id "XLOC_000857"; transcript_id "TCONS_00001767"; exon_number "2"; ; class_code "u"; tss_id "TSS1049";
chr2L	Cufflinks	gene_id "XLOC_000857"; transcript_id "TCONS_00001767"; exon_number "3"; class_code "u"; tss_id "TSS1049";
chr2L	Cufflinks	gene_id "XLOC_000867"; transcript_id "TCONS_00001805"; exon_number "1"; class_code "u"; tss_id "TSS1068";
chr2L	Cufflinks	gene_id "XLOC_000867"; transcript_id "TCONS_00001805"; exon_number "2"; class_code "u"; tss_id "TSS1068";
chr2L	Cufflinks	gene_id "XLOC_000886"; transcript_id "TCONS_00001843"; exon_number "1"; class_code "x";
chr2L	Cufflinks	gene_id "XLOC_000887"; transcript_id "TCONS_00001844"; exon_number "1"; class_code "j"; tss_id "TSS1091";
chr2L	Cufflinks	gene_id "XLOC_000887"; transcript_id "TCONS_00001844"; exon_number "2"; class_code "j"; tss_id "TSS1091";
chr2L	Cufflinks	gene_id "XLOC_000887"; transcript_id "TCONS_00001844"; exon_number "3"; class_code "j"; tss_id "TSS1091";
chr2L	Cufflinks	gene_id "XLOC_000887"; transcript_id "TCONS_00001844"; exon_number "4"; ; class_code "j"; tss_id "TSS1091";
chr2L	FlyBase	gene_id "XLOC_000901"; transcript_id "TCONS_00001877"; exon_number "1"; class_code "=";
chr2L	Cufflinks	gene_id "XLOC_000902"; transcript_id "TCONS_00001878"; exon_number "1"; class_code "x";
chr2R	Cufflinks	gene_id "XLOC_003782"; transcript_id "TCONS_00007807"; exon_number "1"; class_code "x"; tss_id "TSS4243";
chr2R	Cufflinks	gene_id "XLOC_003782"; transcript_id "TCONS_00007807"; exon_number "2"; class_code "x"; tss_id "TSS4243";
chr2R	FlyBase	gene_id "XLOC_003783"; transcript_id "TCONS_00007808"; exon_number "1"; class_code "="; tss_id "TSS4244";
chr2R	Cufflinks	gene_id "XLOC_003784"; transcript_id "TCONS_00007810"; exon_number "1"; class_code "j"; tss_id "TSS4246";
chr2R	Cufflinks	gene_id "XLOC_003784"; transcript_id "TCONS_00007810"; exon_number "2"; class_code "j"; tss_id "TSS4246";
chr2R	Cufflinks	gene_id "XLOC_003784"; transcript_id "TCONS_00007810"; exon_number "3"; class_code "j"; tss_id "TSS4246";
chr2R	Cufflinks	gene_id "XLOC_003784"; transcript_id "TCONS_00007810"; exon_number "4"; class_code "j"; tss_id "TSS4246";
chr2R	Cufflinks	gene_id "XLOC_004549"; transcript_id "TCONS_00009527"; exon_number "1";; class_code "u"; tss_id "TSS5130";
chr2R	Cufflinks	gene_id "XLOC_004549"; transcript_id "TCONS_00009527"; exon_number "2"; ; class_code "u"; tss_id "TSS5130";
chr2R	Cufflinks	gene_id "XLOC_004550"; transcript_id "TCONS_00009528"; exon_number "1"; class_code "u"; tss_id "TSS5131";
chr2R	Cufflinks	gene_id "XLOC_004550"; transcript_id "TCONS_00009528"; exon_number "2"; class_code "u"; tss_id "TSS5131";
chr3L	Cufflinks	gene_id "XLOC_006483"; transcript_id "TCONS_00013701"; exon_number "1"; class_code "j"; tss_id "TSS7243";
chr3L	Cufflinks	gene_id "XLOC_006483"; transcript_id "TCONS_00013701"; exon_number "2"; ; class_code "j"; tss_id "TSS7243";
chr3L	Cufflinks	gene_id "XLOC_006483"; transcript_id "TCONS_00013701"; exon_number "3"; class_code "j"; tss_id "TSS7243";
chr3L	Cufflinks	gene_id "XLOC_007340"; transcript_id "TCONS_00015452"; exon_number "1"; class_code "u"; tss_id "TSS8255";
chr3L	Cufflinks	gene_id "XLOC_007340"; transcript_id "TCONS_00015452"; exon_number "2"; ; class_code "u"; tss_id "TSS8255";
chr3L	Cufflinks	gene_id "XLOC_007340"; transcript_id "TCONS_00015453"; exon_number "1"; class_code "u"; tss_id "TSS8255";
chr3L	Cufflinks	gene_id "XLOC_007340"; transcript_id "TCONS_00015453"; exon_number "2"; class_code "u"; tss_id "TSS8255";
chr3L	Cufflinks	gene_id "XLOC_007340"; transcript_id "TCONS_00015453"; exon_number "3"; class_code "u"; tss_id "TSS8255";
chr3R	Cufflinks	gene_id "XLOC_011674"; transcript_id "TCONS_00024666"; exon_number "1"; class_code "u"; tss_id "TSS13385";
chr3R	Cufflinks	gene_id "XLOC_011674"; transcript_id "TCONS_00024666"; exon_number "2"; ; class_code "u"; tss_id "TSS13385";
chr3R	Cufflinks	gene_id "XLOC_011675"; transcript_id "TCONS_00024667"; exon_number "1"; class_code "x"; tss_id "TSS13386";
chr3R	Cufflinks	gene_id "XLOC_011675"; transcript_id "TCONS_00024667"; exon_number "2"; class_code "x"; tss_id "TSS13386";
chr3R	Cufflinks	gene_id "XLOC_011676"; transcript_id "TCONS_00024668"; exon_number "1"; class_code "="; tss_id "TSS13387";
chr3R	Cufflinks	gene_id "XLOC_011676"; transcript_id "TCONS_00024668"; exon_number "2"; class_code "="; tss_id "TSS13387";
chr3R	Cufflinks	gene_id "XLOC_011676"; transcript_id "TCONS_00024668"; exon_number "3"; class_code "="; tss_id "TSS13387";

Note that this is only an excerpt from merge.gtf file (the original file is too big to include in whole in the thesis). This file contains all the tissue samples. It shows the chromosome position, the gene prediction source (Cufflinks or FlyBase), gene identification number (XLOC_), transcripts identification number (TCONS_), exon numbers in the transcripts, and class_code. tss_id is the ID of this transcript's inferred start site. Determines which primary transcript this processed transcript is believed to come from.

Table 4-7 Summary of classified transcripts in the merged.gtf file for all tissues

Class_code id	Description	Total number of transcripts
=	Complete match of intron chain	114761
u	Unknown, intergenic transcript	2568
o	Generic exonic overlap with a reference transcript	1114
j	Potentially novel isoform (fragment): at least one splice junction is shared with a reference transcript	76460
x	Exonic overlap with reference on the opposite strand	1071
s	An intron of the transfrag overlaps a reference intron on the opposite strand (likely due to read mapping errors)	205
p	Possible polymerase run-on fragment (within 2Kbases of a reference transcript)	0
r	Repeat. Currently determined by looking at the soft-masked reference sequence and applied to transcripts where at least 50% of the bases are lower case.	0
e	Single exon transfrag overlapping a reference exon and at least 10 bp of a reference intron, indicating a possible pre-mRNA fragment.	0
i	A transfrag falling entirely within a reference intron	0
c	Contained	0
Total		196179

Note that this is the summary of all classified transcripts in the merged file produced by Cufflinks for heads, testes, tubules and whole flies. The class_code id is defined by Cufflinks.

4.3.7 *Calculating differential gene expression using Cuffdiff*

4.3.7.1 Differential expression in all tissues

Cuffdiff was run using the merged.gtf as input to find the differential expression in genes and isoforms so identifying tissue-enriched gene expression. The output files produced by this data analysis were gene_exp.diff including differential gene expression of tubules versus whole flies (Tb-Wf), heads versus whole flies (Hd-W), and testes versus whole flies (Ts-Wf) and isoform_exp.diff including transcript differential expression of tubules versus whole flies (Tb-Wf), heads versus whole flies (Hd-wf), and testes versus whole flies (Ts-Wf). From the data in isoform_exp.diff, entries with an undefined class_code marked as "_" corresponded to the class_code "u" in the merged.gtf file had been found. No class_code "i" had been found in the merged files. Selection criteria was applied to the resulting data to select the subset class_code which marked "u" in the isoform_exp.diff data and with $p<0.05$, $q<0.05$, fold change≥3 and RPKM≥1 in order to find the tubule-enriched genes. For the statistical testing, Cuffdiff fits a model of fragment count variance across replicates of each sample. The variance is estimated using either the negative binomial distribution when a gene has a single isoform, or the beta negative binomial distribution when a gene has multiple isoforms (Anders and Huber, 2010; Trapnell et al., 2013). For each gene, the $\log_2$-fold change between the FPKM values in two experimental conditions and their estimated variances produce a variable that is approximately normally distributed to which standard statistics can be applied (student's t test, two-tailed); *p*-values are then adjusted for multiple testing using Benjamini-Hochberg correction (Benjamini et al., 2001; Benjamini and Hochberg, 1995) and are reported as *q*-values (Bullard et al., 2010; Storey, 2003). Cuffdiff performs only pair-wise comparison for differential expression, so the comparisons were between tubules/whole flies, testes/whole flies and heads/wholes flies. To enhance accuracy of differential analysis, upper quantile normalization (--upper-quantile-norm") and multi-mapped read correction ("--multi-read-correction") were applied (Bullard et al., 2010; Dillies et al., 2013; Mortazavi et al., 2008). A summary of the novel exons and novel genes in different tissues (tubules, heads, testes and whole flies) generated by Cuffdiff is listed in Table 4-8.

Table 4-8 Summaries of novel single exons, multiexons (genes) in tubules, heads, testes and whole flies.

Tissue	Number of exons						
	1 exon	2 exons	3 exons	4 exons	5 exons	6 exons	Total
Testes	906	381	71	19	1	0	1378
Tubules	143	33	3	0	0	0	179
Heads	228	35	10	5	0	2	208
Wholes flies							2568

Note these are the results from Cuffdiff combined with merge.gtf results (RPKM≥1). In this study, we only consider multiexon or single exon which are conserved in multiple tissues as novel genes (Cabili et al., 2011; Graveley et al., 2011; Roberts et al., 2011).

Table 4-9 Summary of tissue-enriched novel exons and gene numbers.

Tissue	Class_code	Tissue vs whole flies	Total number of novel isoforms
Tubules	u	P<0.05, q<0.05,RPKM≥1, fold change≥3	71
Testes	u	P<0.05, q<0.05,RPKM≥1, fold change≥3	1144
Heads	u	P<0.05, q<0.05,RPKM≥1, fold change≥3	119

Note that this is the result from Cuffdiff combined with merge.gtf results (p<0.05, q<0.05, RPKM≥1, fold change≥3). The total number of novel isoforms including novel single exon and multiple exons isoforms (novel genes).

A summary of the tissue-enriched genes of tubules, testes, and heads from the Cuffdiff and merge.gtf results is shown in Table 4-9.

4.3.7.2 **Differential expression of tubule-specific genes**

The set of tubule-enriched genes in the annotated *Drosophila* genome contains more coding genes than noncoding genes (Figure 4-6). However, in the novel tubule-enriched gene list (Table 4-10 as generated by Cuffdiff); there are many more noncoding genes than coding genes. Most of the novel genes are more likely to be noncoding genes (Figure 4-6).

Coding genes are normally more highly expressed than noncoding genes. On analysing long noncoding RNA genes in *D. melanogaster* at 30 developmental time points using modENCODE whole transcriptome (RNA-seq) data, Daines and his colleague found that across the different samples, the total gene model expression was, on average, 253-fold higher than for long noncoding RNA loci. (Daines et al., 2011; Young et al., 2012). EST, SAGE and Tiling array technologies are mainly more biased towards the highly expressed genes (Alba et al., 2004; Graveley, 2008), so more coding genes were discovered than noncoding genes in the past.

Novel noncoding genes are often expressed in stage and/or sex-specific patterns (Young et al., 2012), some novel noncoding gene are expressed in tissue-specific patterns. As a consequence, these genes may be difficult to discover until the specific tissues or stages and/or sex are studied (Daines et al., 2011; Wang et al., 2004).

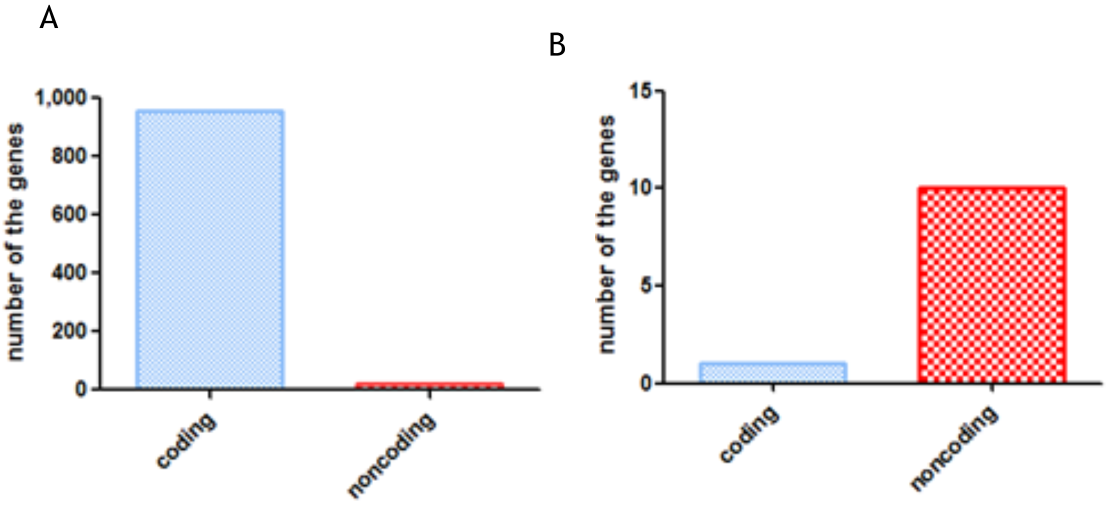


Figure 4-6 Tubule-enriched genes from the Cuffdiff.

Tubule-enriched genes list from Cuffdiff (tubules compared with whole flies, $p<0.05$, $q<0.05$, RPKM≥1, fold change≥3). **(A)The canonical genes list (**genes which annotated by FlyBase) indicated that there were more coding genes than noncoding genes in the annotated genes group. **(B) The noncanonical genes list** (novel gene lists generated from Table 4-10) indicated that there were more noncoding genes than coding genes in the novel gene group.

Table 4-10 Tubule-enriched novel exons and genes by Cuff_diff.

test_id	locus	whole flies RPKM	tubules RPKM	fold-change	p_value	q_value	class_code
TCONS_00006099	chr2L:13095624-13109468	0	7.42818	100000	0.005959	0.014623	u
TCONS_00006179	chr2L:19079233-19108108	0	2.69345	100000	0.020138	0.040711	u
TCONS_00006202	chr2L:21995616-21996517	0	1.5216	100000	0.005914	0.014529	u
TCONS_00006003	chr2L:3269436-3319912	0	2.15575	100000	0.000441	0.001518	u
TCONS_00012999	chr2R:18966948-18967217	0	21.2493	100000	0.001206	0.003687	u
TCONS_00012860	chr2R:2534767-2535342	0	2.17334	100000	0.009734	0.022123	u
TCONS_00027429	chr3R:1098380-1176536	0	0.175369	100000	0	0	u
TCONS_00027668	chr3R:16908757-16923621	0	0.020921	100000	5.04E-46	1.11E-44	u
TCONS_00027669	chr3R:17139066-17139474	0	2.75282	100000	0.02308	0.045647	u
TCONS_00027581	chr3R:6943963-6965634	0	0.024942	100000	0	0	u
TCONS_00034144	chrX:19217991-19218262	0	18.5448	100000	0.002074	0.005909	u
TCONS_00016194	chr3L:23777293-23780611	0.612365	86.539	141.319	0	0	u
TCONS_00027722	chr3R:23310324-23312528	0.213081	27.7166	130.0748	0	0	u
TCONS_00027738	chr3R:24393801-24396268	0.06886	7.21746	104.8132	3.56E-11	4.03E-10	u
TCONS_00034129	chrX:17350216-17425171	0.018634	1.81233	97.25827	0	0	u
TCONS_00027692	chr3R:18990380-18991802	0.057033	3.93441	68.98437	3.72E-05	0.000166	u
TCONS_00008658	chr2R:16189454-16190517	1.40763	76.2053	54.13712	8.88E-16	1.74E-14	u
TCONS_00001177	chr2L:8490771-8491444	0.109924	5.07965	46.21064	0.000216	0.000804	u
TCONS_00034123	chrX:16720640-16730794	0.166643	7.61787	45.71379	0	0	u
TCONS_00006097	chr2L:13095624-13109468	0.054874	1.82877	33.32707	5.22E-07	3.24E-06	u
TCONS_00020479	chr3R:5951869-5965492	0.158627	4.81691	30.3662	0.000555	0.001866	u
TCONS_00019349	chr3L:4852862-4853889	0.510701	15.0029	29.37723	8.36E-09	6.79E-08	u
TCONS_00027207	chr3R:26230174-26231091	0.118323	3.37127	28.4919	0.000275	0.001	u
TCONS_00027577	chr3R:6943963-6965634	0.174966	4.86959	27.83154	1.14E-05	5.59E-05	u
TCONS_00006065	chr2L:10048688-10049223	0.34109	9.19548	26.95908	0.000128	0.000503	u
TCONS_00027656	chr3R:15425392-15426143	0.135954	3.19778	23.52091	0.002792	0.007632	u
TCONS_00034121	chrX:16298524-16298875	0.52414	8.99951	17.17011	0.007586	0.017925	u
TCONS_00009515	chr2R:184503-184996	1.03831	17.7604	17.10516	5.84E-05	0.000249	u
TCONS_00027672	chr3R:17141001-17142046	0.134378	2.25207	16.75924	0.00136	0.004095	u
TCONS_00013882	chr3L:4629603-4687354	1.48425	23.9035	16.10484	3.76E-06	2.03E-05	u
TCONS_00006094	chr2L:13094768-13095402	0.24743	3.75432	15.17319	0.002173	0.006154	u
TCONS_00006027	chr2L:5339022-5365039	0.767296	11.4185	14.88152	2.11E-08	1.61E-07	u
TCONS_00004375	chr2L:9669700-9670824	0.175688	2.5663	14.60713	0.000551	0.001853	u
TCONS_00027315	chr3R:26956519-26957324	0.195563	2.83968	14.52048	0.004472	0.011489	u
TCONS_00022965	chr3R:24411661-24417150	0.18552	2.64739	14.27001	0.000729	0.002374	u
TCONS_00006095	chr2L:13095624-13109468	0.253019	3.60268	14.23879	0.001076	0.003337	u
TCONS_00012944	chr2R:14692364-14693007	0.150399	1.98069	13.16958	0.017534	0.036256	u
TCONS_00028338	chr4:77917-83616	0.101055	1.28196	12.68563	1.26E-06	7.36E-06	u
TCONS_00006096	chr2L:13095624-13109468	0.24878	3.14448	12.63954	0.003364	0.008989	u
TCONS_00034089	chrX:9474619-9475098	0.260195	3.15075	12.10925	0.020975	0.042118	u
TCONS_00006105	chr2L:13801919-13802826	0.236262	2.80555	11.8747	0.000738	0.002398	u
TCONS_00012972	chr2R:17369423-17385431	0.034267	0.392436	11.45241	0	0	u
TCONS_00027486	chr3R:6689076-6696542	42.3827	469.477	11.0771	5.12E-10	4.99E-09	u
TCONS_00001105	chr2L:8258615-8301072	0.102267	1.12577	11.00808	6.39E-05	0.000271	u
TCONS_00006092	chr2L:13092399-13093841	0.215135	2.32892	10.82534	0.001311	0.003971	u
TCONS_00006093	chr2L:13093910-13094690	0.158768	1.70519	10.74016	0.015536	0.032785	u
TCONS_00017431	chr3L:8569749-8571158	0.379966	3.99535	10.51508	7.15E-05	0.000299	u
TCONS_00012976	chr2R:17593556-17594031	0.300066	3.13174	10.43682	0.023019	0.045545	u
TCONS_00029066	chrX:3405353-3434285	0.272267	2.83816	10.4242	1.40E-06	8.11E-06	u
TCONS_00006008	chr2L:3608199-3608809	0.47313	4.90378	10.36454	0.003641	0.009635	u
TCONS_00027647	chr3R:14627098-14628610	0.345958	3.50511	10.13161	7.38E-05	0.000307	u
TCONS_00012938	chr2R:13365159-13365645	0.352813	3.45851	9.802623	0.016383	0.034253	u
TCONS_00013119	chr2RHet:198175-339379	6.82228	65.5755	9.611978	0.002614	0.007212	u
TCONS_00012877	chr2R:5378104-5378330	15.0949	143.12	9.481286	0.000199	0.00075	u
TCONS_00019659	chr3LHet:1181132-1430621	0.265998	2.40408	9.038019	0.001088	0.003368	u
TCONS_00006080	chr2L:12024039-12025463	0.350991	2.90838	8.286188	0.001364	0.004107	u
TCONS_00034124	chrX:16787650-16787877	3.24521	26.5153	8.170581	0.013434	0.029034	u
TCONS_00027646	chr3R:14624986-14627033	0.55695	4.50366	8.08629	4.22E-05	0.000186	u
TCONS_00012882	chr2R:6153841-6154026	7.97046	61.1122	7.667342	0.011247	0.024977	u
TCONS_00006098	chr2L:13095624-13109468	0.468943	3.52219	7.510935	9.14E-05	0.000371	u
TCONS_00002881	chr2L:21864959-21867977	0.024901	0.17567	7.054883	0.004542	0.011638	u
TCONS_00027621	chr3R:11833945-11834363	0.688453	4.71228	6.844737	0.019567	0.03974	u
TCONS_00006090	chr2L:13089866-13090523	1.36471	8.64992	6.338296	0.001616	0.004754	u
TCONS_00034071	chrX:3434369-3435823	0.413719	2.57905	6.233824	0.002858	0.007792	u
TCONS_00012858	chr2R:2324027-2325796	0.213649	1.26312	5.912094	0.007303	0.017365	u
TCONS_00027461	chr3R:3746122-3793296	0.707098	2.87983	4.072756	0.000871	0.002776	u
TCONS_00027673	chr3R:17142111-17143622	1.68963	6.72888	3.982461	0.003061	0.008276	u
TCONS_00023626	chr3R:1095875-1098297	4.99077	19.0402	3.815074	0.002094	0.00596	u
TCONS_00019387	chr3L:8685535-8686379	1.36052	5.03268	3.699065	0.01746	0.036118	u
TCONS_00019394	chr3L:9094737-9123692	27.4831	88.4202	3.21726	0.007442	0.017645	u
TCONS_00019393	chr3L:9094737-9123692	17.0489	52.8883	3.102164	0.010293	0.02316	u

This table is the differential expression analysis result of tubules versus whole flies from Cuffdiff (RPKM≥1, P<0.05, q<0.05 Fold Change≥3), showing **Test_id** (transcript id), **Locus** (chromosome position), **fold-change** (tubule versus whole fly ratio), **q value** (FDR adjusted p value). **Class-_code** “u” novel genes. The list was ranked by fold change enrichment in tubules. For row 1-12 transcripts have “0” RPKM in whole flies, and so are assigned an arbitrary enrichment of 1000000x

4.3.8 *Illustrating tissue specificity of novel genes using Venn diagram.*

Most of novel genes of *Drosophila* are found in testes and heads; so a Venn diagram was run between tubules, testes and heads in order to eliminate the overlap of novel gene between tissues and found the possible tissue specific novel genes. Using the differential expression genes list which was generated by Cuffdiff, the Venn diagram shows 1125 testes specific novel genes, 103 heads specific novel genes and 55 tubules specific novel genes, and that are listed in Table 4-11.

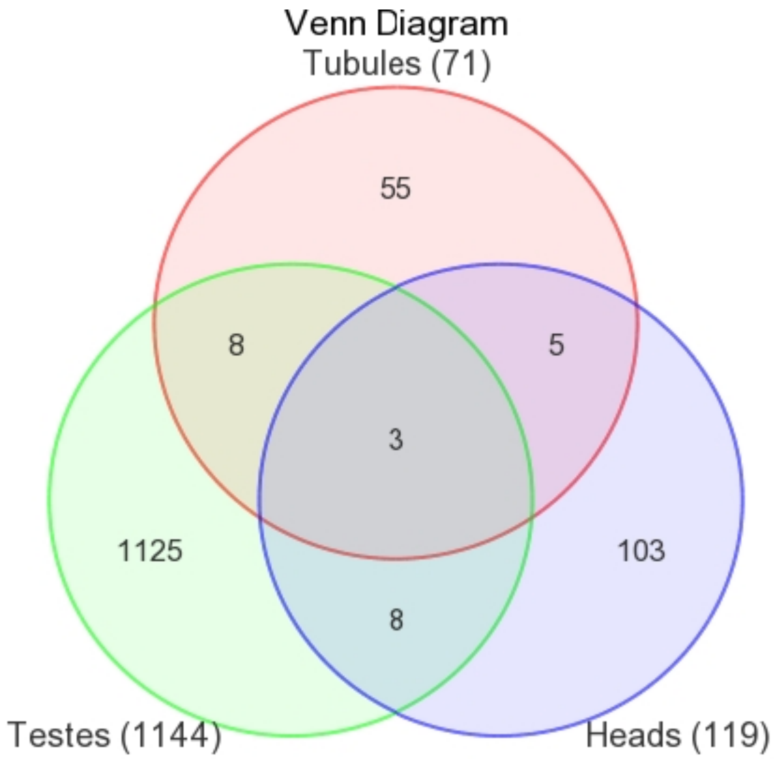


Figure 4-7 Identification of tissue-specific novel genes by Venn diagram.

The diagram is showing the number of novel genes in each subset corresponding to the intersection of tissue types. The total tissue-specific novel genes produced by Cuffdiff are 1144 in testes, 119 in heads, and 71 genes in tubules. The overlap between testes and heads are 11 genes, between heads and tubules are 8 genes, and between tubules and testes are 11 genes. This produced 55 tubule-specific genes, 1125 testes-specific genes and 119 heads-specific genes.

Table 4-11 55 Novel exons and genes in tubules generated by Venn diagram

test_id	locus	whole fly RPKM	tubules RPKM	fold-change	p_value	q_value	class_code
TCONS_00006003	chr2L:3269436-3319912	0	2.15575	100000	0.000441	0.001518	u
TCONS_00006179	chr2L:19079233-19108108	0	2.69345	100000	0.020138	0.040711	u
TCONS_00006202	chr2L:21995616-21996517	0	1.5216	100000	0.005914	0.014529	u
TCONS_00012860	chr2R:2534767-2535342	0	2.17334	100000	0.009734	0.022123	u
TCONS_00012999	chr2R:18966948-18967217	0	21.2493	100000	0.001206	0.003687	u
TCONS_00027669	chr3R:17139066-17139474	0	2.75282	100000	0.02308	0.045647	u
TCONS_00034144	chrX:19217991-19218262	0	18.5448	100000	0.002074	0.005909	u
TCONS_00016194	chr3L:23777293-23780611	0.612365	86.539	141.319	0	0	u
TCONS_00027722	chr3R:23310324-23312528	0.213081	27.7166	130.075	0	0	u
TCONS_00027738	chr3R:24393801-24396268	0.06886	7.21746	104.813	3.56E-11	4.03E-10	u
TCONS_00027692	chr3R:18990380-18991802	0.057033	3.93441	68.9844	3.72E-05	0.000166	u
TCONS_00008658	chr2R:16189454-16190517	1.40763	76.2053	54.1371	8.88E-16	1.74E-14	u
TCONS_00034123	chrX:16720640-16730794	0.166643	7.61787	45.7138	0	0	u
TCONS_00020479	chr3R:5951869-5965492	0.158627	4.81691	30.3662	0.000555	0.001866	u
TCONS_00019349	chr3L:4852862-4853889	0.510701	15.0029	29.3772	8.36E-09	6.79E-08	u
TCONS_00027207	chr3R:26230174-26231091	0.118323	3.37127	28.4919	0.000275	0.001	u
TCONS_00006065	chr2L:10048688-10049223	0.34109	9.19548	26.9591	0.000128	0.000503	u
TCONS_00027656	chr3R:15425392-15426143	0.135954	3.19778	23.5209	0.002792	0.007632	u
TCONS_00034121	chrX:16298524-16298875	0.52414	8.99951	17.1701	0.007586	0.017925	u
TCONS_00009515	chr2R:184503-184996	1.03831	17.7604	17.1052	5.84E-05	0.000249	u
TCONS_00013882	chr3L:4629603-4687354	1.48425	23.9035	16.1048	3.76E-06	2.03E-05	u
TCONS_00006094	chr2L:13094768-13095402	0.24743	3.75432	15.1732	0.002173	0.006154	u
TCONS_00006027	chr2L:5339022-5365039	0.767296	11.4185	14.8815	2.11E-08	1.61E-07	u
TCONS_00004375	chr2L:9669700-9670824	0.175688	2.5663	14.6071	0.000551	0.001853	u
TCONS_00027315	chr3R:26956519-26957324	0.195563	2.83968	14.5205	0.004472	0.011489	u
TCONS_00022965	chr3R:24411661-24417150	0.18552	2.64739	14.27	0.000729	0.002374	u
TCONS_00006095	chr2L:13095624-13109468	0.253019	3.60268	14.2388	0.001076	0.003337	u
TCONS_00012944	chr2R:14692364-14693007	0.150399	1.98069	13.1696	0.017534	0.036256	u
TCONS_00006096	chr2L:13095624-13109468	0.24878	3.14448	12.6395	0.003364	0.008989	u
TCONS_00034089	chrX:9474619-9475098	0.260195	3.15075	12.1092	0.020975	0.042118	u
TCONS_00006105	chr2L:13801919-13802826	0.236262	2.80555	11.8747	0.000738	0.002398	u
TCONS_00027486	chr3R:6689076-6696542	42.3827	469.477	11.0771	5.12E-10	4.99E-09	u
TCONS_00006092	chr2L:13092399-13093841	0.215135	2.32892	10.8253	0.001311	0.003971	u
TCONS_00006093	chr2L:13093910-13094690	0.158768	1.70519	10.7402	0.015536	0.032785	u
TCONS_00017431	chr3L:8569749-8571158	0.379966	3.99535	10.5151	7.15E-05	0.000299	u
TCONS_00012976	chr2R:17593556-17594031	0.300066	3.13174	10.4368	0.023019	0.045545	u
TCONS_00029066	chrX:3405353-3434285	0.272267	2.83816	10.4242	1.40E-06	8.11E-06	u
TCONS_00006008	chr2L:3608199-3608809	0.47313	4.90378	10.3645	0.003641	0.009635	u
TCONS_00012938	chr2R:13365159-13365645	0.352813	3.45851	9.80262	0.016383	0.034253	u
TCONS_00013119	chr2RHet:198175-339379	6.82228	65.5755	9.61198	0.002614	0.007212	u
TCONS_00012877	chr2R:5378104-5378330	15.0949	143.12	9.48129	0.000199	0.00075	u
TCONS_00019659	chr3LHet:1181132-1430621	0.265998	2.40408	9.03802	0.001088	0.003368	u
TCONS_00006080	chr2L:12024039-12025463	0.350991	2.90838	8.28619	0.001364	0.004107	u
TCONS_00034124	chrX:16787650-16787877	3.24521	26.5153	8.17058	0.013434	0.029034	u
TCONS_00012882	chr2R:6153841-6154026	7.97046	61.1122	7.66734	0.011247	0.024977	u
TCONS_00006098	chr2L:13095624-13109468	0.468943	3.52219	7.51093	9.14E-05	0.000371	u
TCONS_00027621	chr3R:11833945-11834363	0.688453	4.71228	6.84474	0.019567	0.03974	u
TCONS_00006090	chr2L:13089866-13090523	1.36471	8.64992	6.3383	0.001616	0.004754	u
TCONS_00034071	chrX:3434369-3435823	0.413719	2.57905	6.23382	0.002858	0.007792	u
TCONS_00027461	chr3R:3746122-3793296	0.707098	2.87983	4.07276	0.000871	0.002776	u
TCONS_00027673	chr3R:17142111-17143622	1.68963	6.72888	3.98246	0.003061	0.008276	u
TCONS_00023626	chr3R:1095875-1098297	4.99077	19.0402	3.81507	0.002094	0.00596	u
TCONS_00019387	chr3L:8685535-8686379	1.36052	5.03268	3.69907	0.01746	0.036118	u
TCONS_00019394	chr3L:9094737-9123692	27.4831	88.4202	3.21726	0.007442	0.017645	u
TCONS_00019393	chr3L:9094737-9123692	17.0489	52.8883	3.10216	0.010293	0.02316	u

Note this is the novel tubule-specific gene list generated by Cuff_diff and Venn diagram. **Test_id** (transcript id), **Locus** (chromosome position), **fold-change** (tubule vs whole fly ratio), **q_value** (FDR adjusted p value). **class-code "u"** novel genes. The list was ranked by fold change enriched in tubules. For row 1-7 transcripts have "0" RPKM in whole flies, and so are assigned an arbitrary enriched of 1000000x

Tubule-specific genes lists generated from RNA-seq need to be confirmed by strand-specific RNA-seq data. The polarity of the transcript is important for correct annotation of novel genes, because it provides essential information about the possible function of a gene (Parkhomchuk et al., 2009). RNA-seq from Illumina GAIIx can facilitate the discovery of novel transcripts, but most studies have not distinguished the transcribed strand (Nagalakshmi et al., 2008; Yassour et al., 2009). The transcripts detected by RNA-seq may be from both forward and reverse strands that were overlapped or partially overlapped. The antisense transcripts, which play an important role in gene regulation from bacteria to human, may be underestimated if only the RNA-seq method is used (Yassour et al., 2010). Most of the novel genes we detected may be noncoding genes which from the antisense strand our results suggest. It is necessary to use strand-specific tubule data to indentify every novel gene which is discovered by RNA-seq. Table 4-12 shows the RNA-seq data which has confirmation from strand-specific RNA-seq data. The genes were confirmed by RT-PCR (section 4.3.10) which informed using the RNA-seq data combined with tubule strand-specific RNA-seq data.

Table 4-12 Summary of 55 tubule-specific novel genes in the list

Exons of Transcripts	number	conservation	Supported by ssRNA-seq
I exon	45	1 conserved	22
2 exons	10	5 conserved	7

Note this table is the 55 tubules-specific novel genes that supported by ss-RNA-seq. 55 novel genes and exons checked manually by corresponding genes of ss-RNA-seq data in Tablet. **Conservation** means the novel genes or exons exist in multiple tissues.

4.3.9 *Finalizing the list of tubule-specific genes*

To ensure the results are not affected by genomic contamination in cDNA library, the novel genes are only considered such if the transcripts are longer than 200bp and have multiexons or single exons that are conserved in different species (Cabili et al., 2011; Graveley et al., 2011; Young et al., 2012). We also manually checked the novel transcripts with the ssRNA-seq data in Tablet (RNA-seq viewer version 1.12.03.26) to avoid false positive products. The final novel tubule-specific genes list only includes transcripts with multiexons longer than

200 bp and which are also supported by ssRNA-seq data (see Table 4-13). The RNA type was defined by CPC (coding potential calculator) which assess the protein-coding potential of a transcript based on six biologically meaningful sequence features (Kong et al., 2007) CPC is a user-friendly web-based interface of CPC at http://cpc.cbi.pku.edu.cn.

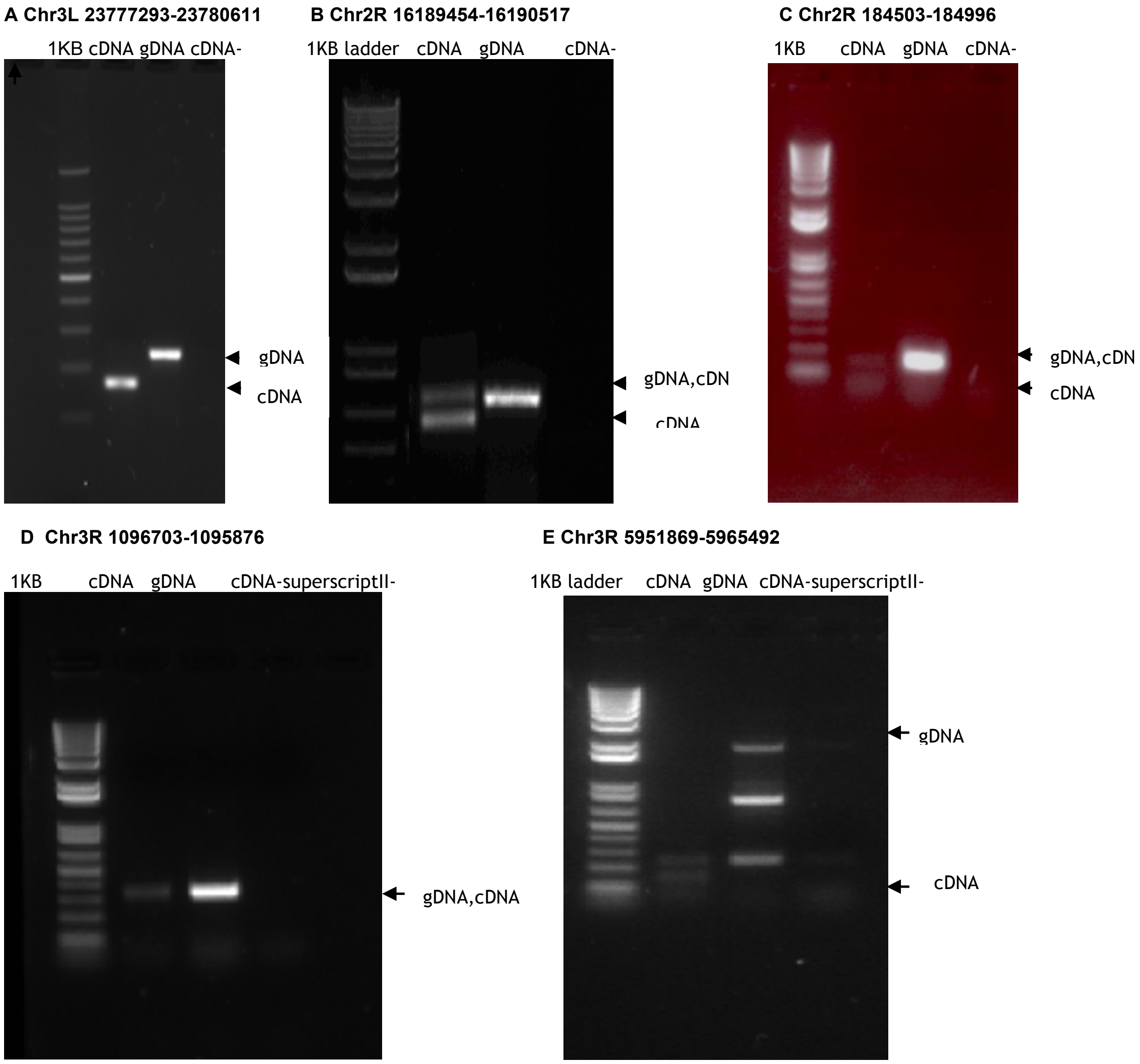
Table 4-13 Final tubule-specific gene list (only count multiexons)

Test_id	Locus	Whole flies RPKM	Tubules RPKM	Fold-change	RNA type	Conservation in Drosophila
TCONS_00016194	chr3L:23777293-23780611	0.612365	86.539	141.319	coding	conserved
TCONS_00008658	chr2R:16189454-16190517	1.40763	76.2053	54.1371	noncoding	conserved
TCONS_00023626	chr3R:1095875-1098297	4.99077	19.0402	3.81507	noncoding	Not conserved
TCONS_00009515	chr2R:184503-184996	1.03831	17.7604	17.1052	noncoding	conserved
TCONS_00020479	chr3R:5951869-5965492	0.158627	4.81691	30.3662	noncoding	conserved
TCONS_00022965	chr3R:24411661-24417150	0.18552	2.64739	14.27	mRNA-like noncoding	conserved
TCONS_00004375	chr2L: 9670824-9669700	0.175688	2.5663	14.6071	noncoding	Not conserved

This table is chosen from the Table 4-11 that is only considered transcripts have multiexons (all of these transcripts in Table 4-13 have two exons) and also supported by strand-specific RNA-seq data. **Test_id** is identification number defined by Cufflinks. **Whole flies value** (RPKM), **tubules value** (RPKM), **RNA type** from CPC (coding potential calculator) prediction score, **Conservation** from blastN search. The gene which is highlighted in red will be further investigated in Chapter 4. The RNA used in this table is from Canton S whole flies and Canton S tubules.

4.3.10 *RT-PCR validation of tubule-specific novel genes predicted by RNA-seq and supported by ssRNA-seq*

RT-PCR was performed on these predicted novel transcripts in order to confirm that these novel transcripts were real. Because all of these predicted novel transcripts contained two exons. The primers were designed on the two exons that the cDNA of PCR products would span an intron. If the predictions were correct, the PCR amplified genomic DNA and cDNA would result in two different sizes band, and the difference would be the spliced intron. Then the novel genes would be confirmed, and also genomic contamination would be eliminated. If the PCR amplified genomic DNA size is the same as the PCR amplified cDNA size indicating that the PCR amplified cDNA product may be genomic contamination, and in this case the Superscript minus control of cDNA will help to confirm if the amplified cDNA product is genomic contamination. If the PCR amplified Superscript minus control of cDNA has the same size band as PCR amplified cDNA that will confirm the transcript is not real, and only a genomic contamination. RT-PCR products and primers for novel genes are detailed in Appendix VI.



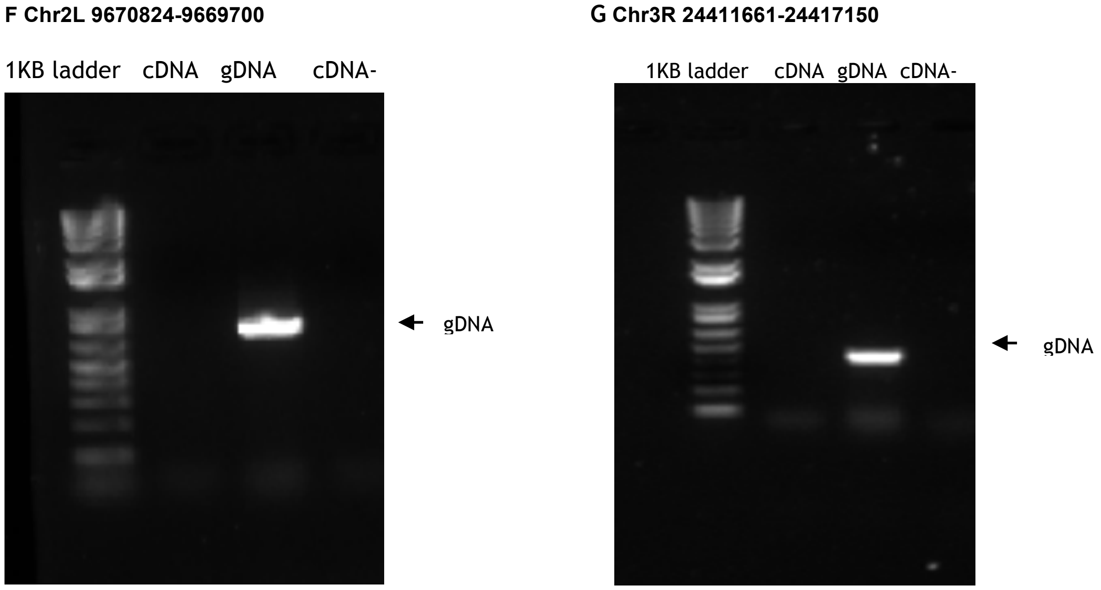


Figure 4-8 RT-PCR validations of tubule-specific novel genes predicted by RNA-seq.

(A) Chr3L 23777293-23780611 is a predicted novel coding transcript. PCR primers were designed to amplify a 241bp product from genomic DNA. The expected splice junction associated with this transcript, 3L: 23777433-23777506 (73bp) results in a 178bp RT-PCR product when amplified from tubule cDNA. **(B) Chr2R 16189454-16190517 is a predicted novel noncoding transcript (mlncRNA).** PCR primers were designed to amplify a 243bp product from genomic DNA. The expected splice junction associated with this transcript, 2R: 16190080-16190140 (60bp), results in a 183bp RT-PCR product and 243bp RT-PCR products indicating this transcript has two isoforms (246bp and 195bp). **(C) Chr2R 184996-194503 is a predicted noncoding transcript on minus strand**. PCR primers were designed to amplify a 161bp product from genomic DNA. The expected splice junction is Chr2R 184862-184775 (87bp). The PCR results show a 151bp and 64bp RT-PCR products from cDNA indicating this transcript has two isoforms. **(D) Chr3R 1096703-1095876 is a predicted novel noncoding transcript on reverse strand**. PCR primers were designed to amplify a 346bp product from genomic DNA. The expected splice junction associated with this transcript, 3R: 1096660-1096607 (53bp). The cDNA RT-PCR product is not detected for this junction but shows this transcript. **(E) Chr3R 5951869-5965492 is a predicted noncoding transcript on minus strand**. PCR primers were designed to amplify a 5870bp product from genomic DNA but this gDNA is only 3kb. The expected splice junction is Chr3R 5957141-5962859 (5718bp). The PCR results in a 152bp RT-PCR product (cDNA lower band) **(F) Chr2L 9669700-9670824 is a predicted novel noncoding transcript on reverse strand**. PCR primers were designed to amplify a 835bp product from genomic DNA. The expected splice junction associated with this transcript, 3R: 9670824-9669700 (113bp). No transcribed product has been detected with cDNA. **(G) Chr3R 24411661-24417150 is a predicted novel noncoding genes**. PCR primers were designed to amplify 428bp products from genomic DNA. Splice junction associated with this product is 3R 24416577-24416669 (92bp). No transcribed product has been detected with cDNA. The prediction may be wrong or the primer design may not be correct.

From Figure 4-8 (A) to (G), five out of seven (71%) novel genes which were supported by ssRNA-seq have also been confirmed by RT-PCR. (A) (B) (C) (E) showed the PCR products of genomic DNA and cDNA had different sizes indicating that the transcripts were spliced during the reverse transcription and the introns had been spliced out. The size difference between genomic DNA and cDNA were the spliced introns, and also the superscript minus corresponding cDNA control didn't show any band indicating there was no genomic contamination in the RNA samples and the transcripts were real. So Figure 4-8 (A) (B) (C) (E) had confirmed the splice junctions. These novel genes are more likely real. (D) Only confirmed a transcriptional expression not the splice junction. Because the genomic DNA and cDNA had the same size, but the cDNA superscript minus control was no amplified product. (F) and (G) did not detect PCR amplified product may be the expression was too low to be detected or the problems with primers design were not sure.

4.3.11 *Tubules specific novel transcripts predicted by RNA-seq but not supported by ssRNA-seq*

Table 4-14 were chosen from Table 4-11 which were the list of 55 predicted novel genes of tubules. All of these transcripts have two exons; however these transcripts were not shown on strand-specific RNA-seq data. We further investigated these novel transcripts by RT-PCR.

Table 4-14 Tubules specific novel transcripts predicted by RNA-seq but not supported by ssRNA-seq

test_id	Locus	Whole flies	Tubules	Fold-change	RNA type	Conservation
TCONS_00017431	chr3L:8569749-8571158	0.379966	3.99535	10.5151	noncoding	Not conserved
TCONS_00027207	chr3R:26230174-26231091	0.118323	3.37127	28.4919	noncoding	Not conserved
TCONS_00027315	chr3R:26956519-26957324	0.195563	2.83968	14.5205	noncoding	Not conserved

This table is chosen from the Table 4-11 that only considers transcripts with multiexons (all of these transcripts in Table 4-14 have two exons) but were not supported by strand-specific RNA-seq data. **Test_id** is identification number defined by Cufflinks. **Whole flies value** (RPKM), **tubules value** (RPKM), **RNA type** from CPC (coding potential calculator) prediction score, **Conservation** from blastN search.

RT-PCR for two samples

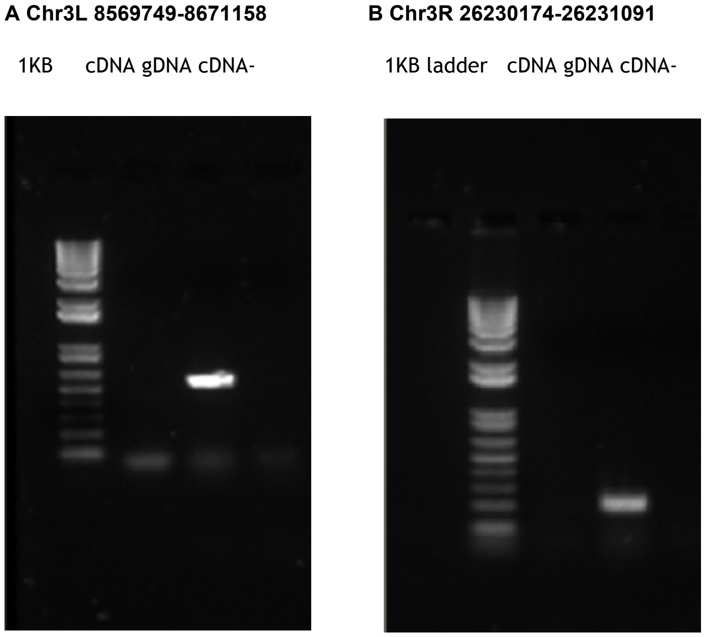


Figure 4-9 Examples by validating the novel genes predicated by RNA-seq but not supported by ssRNA-seq

(A) Chr3L 8571158-8569749 and (B) Chr3R 26230714-26231091 are predicted novel noncoding genes, PCP primers were designed to amplify 580bp and 213bp products from genomic DNA. Splice junctions associated with these two products are 3L: 8569846-8569784 (62bp); 3R26231050-26230965 (85bp). No transcribed products have been detected with cDNA. The prediction may be wrong because the transcripts were not supported by ssRNA-seq data.

Two of thee transcripts could not be validated by qPCR. So the prediction may be wrong, due give no support from strand-specific RNA-seq data.

4.4 Discussion

RNA-sequencing (RNA-seq) is a powerful method for discovering, annotating, and quantifying RNA transcripts in different organisms at the whole transcriptome level. RNA-seq can be used for discovery applications such as identifying alternative splicing events, gene fusions, allele-specific expression, and rare and novel transcripts.

RNA-seq is a robust technology for transcriptome profiling including the characterization of gene models. However, not all annotated genes are well represented by RNA-seq reads. Some genes are under represented by RNA-seq. This may have multiple causes. Firstly, a lack of gene coverage may result due to some reads not uniquely mapping to the reference sequence or no reads originating from the genes in question. Some read mapping methodologies cause problems for splice junction mapping or multimapping (Cherbas et al., 2011). One study changed the mapping methods to reanalyse the results reported in other papers and found more novel lincRNA than the original paper (Roberts et al., 2011; Young et al., 2012). Secondly, library preparation methods can produce different transcriptome profiles (such as polyA selection or ribosome reduction method) as discussed later in this subsection. Thirdly, some tissue- or cell- specific type expression may not be observed due to their restricted expression pattern. We may need to choose the specific tissue or cell type to study (Cabili et al., 2011; Chintapalli et al., 2007; Wang et al., 2004) and also consider changing to a different approach for size selection of library preparation. Fourthly, the very low expression genes (RPKM<4) may not be detected due to the sequence coverage (see chapter 3). RNA-seq can't detect the direction; it is difficult to recognize the transcripts if the reads come from reverse and forward strands at the same position of the chromosome (Yassour et al., 2010). In this case, directional RNA-seq may help to distinguish the read direction. For this study, the greater the level of support available for transcripts from directional RNA-seq, the easier it was to confirm the results (as shown in Tables 4-13 and Table 4-14). Hence library preparation and the read mapping methodologies employed will be the main reasons for genes being under represented by RNA-seq, but other reasons still play roles in the novel genes discovery.

There are two methods to prepare RNA-seq samples, ribo-minus RNA-sequencing (rmRNA-seq), and polyA-selected RNA-sequencing (mRNA-seq or poly(A) selection). For novel transcript discovery, the poly(A) selection method, which would be enriched for coding transcripts is suitable for discovering those transcripts which have the poly(A) tail such as coding RNA and the mRNA-like noncoding RNA, but is not suitable for discovering the noncoding RNA which do not have poly(A) tail. However, the polyA+ selection did not fully exclude RNAs that are not polyadenylated. Alternatively, some of these may be polyadenylated under normal condition, or they could correspond to degradation intermediates (van Bakel et al., 2010). The other method is ribosome reduction. This process minimizes ribosomal contamination and maximizes the percentage of uniquely mapped reads covering both mRNA and a broad range of noncoding RNA species of interest including long intergenic noncoding RNA (lincRNA), small nuclear RNA (snRNA), and small nucleolar RNA (snoRNA)(Cui et al., 2010). However during the ribosome reduction, the mRNA transcription may change. This method may contribute bias to the coding gene expression.

The unpolyadenylated transcripts such as miRNA and snoRNA are not likely to be observed since the poly(A) selection method was used in this study (Daines et al., 2011). The novel coding RNA or noncoding RNA with poly(A), and some other noncoding RNAs would be found in this project. So the novel gene list we found including one coding RNA and six noncoding RNAs for which the size are longer than 200bp could be mRNA-like noncoding RNAs (mlncRNA) (Table 4-13) (Hiller et al., 2009; Soshnev et al., 2011). The one coding RNA within the list was confirmed by RT-PCR. The noncoding RNAs, either are mRNA-like noncoding RNA or contamination with genomic DNA or secondary structures of RNA that may come from the beads during the samples cleaned up. PCR confirmed the second, the third and fourth noncoding RNAs have two exons and were spliced so it could be mlncRNA. The results supported by directional RNA-seq will have more power on the data.

The novel genes that were identified and defined in this study as transcripts have single to multiple exons which appeared in multiple tissues, and the transcripts did not overlap in FlyBase gene models. If the transcripts lacked evidence of significant protein coding ability and were longer than 200bp, they were defined as lincRNA. Furthermore, the transcripts with the protein coding ability were defined as novel coding genes. During the novel gene search, we found a number of novel exons in the novel genes list (Table 4-11). However, these were not considered as novel genes themselves as previous studies revealed most of novel exons are novel exons of known genes (neighbouring, often distant, genes). These exons are located at the 5’ or 3’ ends of a gene, and some are within a gene (Cabili et al., 2011; Cherbas et al., 2011). We observed similar case to the above where we found 5’ or 3’ alternative novel exons. Thus we don’t consider the single novel exons as novel genes. In addition, some single exons may come from the genomic DNA contamination, we can’t confirm them by PCR due to genomic DNA and cDNA will have the same size if they are not the products of splicing. In this study, multiexonic transcripts are easier to confirm from different sizes of cDNA and genomic DNA when using primers that span the introns. The RNA-seq is supported by strand-specific RNA-seq; the results will be less false positive. Strand-specific RNA-seq has more power to detect the noncoding genes that are located in the reverse strand. Five out of seven novel genes (71%) supported by ssRNA-seq were and further confirmed by RT-PCR in this study, which is better than ~60% reported by another study (Daines et al., 2011). Novel transcripts Chr3R 24411661-24417150 and Chr2L9670824-9669700 are supported by ssRNA-seq but the expression level on ssRNA-seq was very low. This may be the reason for the transcripts to be undetected by RT-PCR. Novel transcripts that were detected by RNA-seq but not supported by ssRNA-seq proved more likely to fail in the RT-PCR detection (Table 4-14. Figure 4-9).

The results agree with other studies that novel genes had specific characters as follows. Novel genes would be expected to be expressed at low levels. RNA-seq has more power to detect the lowly expressed genes and allows strand-specific expression detection in contrast to the Affymetrix microarrays for example. The majority of novel transcripts contain only two exons (Daines et al., 2011). All the novel genes in this study have two exons, but they could have multiple exons in other tissues. Novel noncoding genes are often expressed in a tissue-specific manner than coding genes. One study revealed that lincRNAs are associated with specific diseases (Cabili et al., 2011). This study revealed testes have more novel genes than other tissues. Novel genes have more noncoding genes than coding genes especially in specific tissues (Daines et al., 2011). These noncoding genes may play special regulatory roles in gene transcription (Daines et al., 2011; Ponting et al., 2009), cellular function and influence alternative splicing (Tripathi et al., 2010). This study revealed more coding genes than noncoding genes for the noncanonical list of tubule-specific genes. However, it has much noncoding genes than coding genes in the canonical tubule-specific genes list (shown in Figure 4-6 A and B). These results support the evidence that novel genes are more likely to be noncoding genes.

Evidence of lincRNA functionality will be most compelling if disruption of loci frequently results in reproducible cellular or organismal phenotypes (Young et al., 2012). This may be easier to achieve in *Drosophila* than with other organisms. Novel gene discovery will have more impact with *Drosophila* as a model organism.

4.5 **Conclusions**

RNA-seq is a cutting-edge technology searching for novel genes. This technology overcomes the limitations of previous technologies, has become the most popular technology for novel gene discovery. The novel genes found by the RNA-seq poly(A) selection method are more likely to be coding genes and mRNA-like noncoding genes. Most of these genes have two exons, non abundant, belong to noncoding genes and expressed in a tissue-specific manner. Strand-specific RNA-seq has more power to verify the noncoding genes or overlapping genes. The RNA-seq ribosome reduction method may have more power for discovering all novel noncoding RNAs than the poly(A) selection method.

5. Functional studies of *Drosophila* novel gene Chr3L 23777200-23781000 using reverse genetics

Summary

This chapter describes a reverse genetics approach employed in order to assess the functional role of the *Drosophila* tubule-enriched novel gene, at Chr3L 23777200-23781000[1], which was identified by RNA-seq method from chapter 4.

Multiple RNAi knockdown constructs in the pRISE vector were generated for Chr3L 23777200-23781000, and RNA overexpression constructs in the PTW and PTWV vectors were generated for Chr3L 23777200-23781000 through Invitrogen Gateway recombination technology in which the RNAi or RNA overexpression constructs were placed under UAS control. The PTW overexpression construct was to overexpress wild-type RNA and the PTWV construct was to generate YFP (Venus) fusion. Transgenic animals bearing the above UAS constructs were generated using *Drosophila* germline transformation technology.

The RNAi and RNA overexpression fly lines were analyzed in this chapter using a combination of genetic and molecular cellular biology tools including genetic crossing, qPCR and confocal microscopy. Crossing the Chr3L 23777200-23782000-RNAi line to the c42 GAL4 driver allowed expression of RNAi construct in Malpighian tubules principal cell only; crossing RNAi construct to c724 GAL4 driver allowed expression of RNAi construct in tubules stellate cell only. Quantitative expression measurement by qPCR with the c42-chr3L 23777200-23781000-RNAi and c724-chr3L 23777200-23781000-RNAi lines together with their respective control parental lines confirmed the location of the gene in tubule principal cells. Secretion assay with these two lines compared with parental lines suggested the phenotype of this gene and potential function of this gene. Ubiquitous GAL4 driver, Actin-GAL4/CyO crossed with Chr3L 23777000-2378100 YFP fusion line also indicated the localization of the novel gene Chr3L 23777200-

[1] This gene was later named in Flybase as CG43968.

23781000. As demonstrated by this chapter, reverse genetic approach proved particularly useful for searching of novel gene functions.

5.1 Introduction

5.1.1 *GAL4/UAS system in Drosophila*

The GAL4-UAS system which is the second generation of enhancer trapping is unique in *Drosophila* to achieve cell-specific inactivation of virtually any gene (Hardy et al., 2010). The reporter gene is the yeast transcription factor GAL4 which is expressed in a cell and tissue specific manner. It is capable of driving transgenes under control of the yeast UAS_G promoter, the upstream activation sequence that is bound by GAL4 (Dow and Davies, 2003; Duffy, 2002). Cell-specific expression of the transgene is achieved in the progeny of a cross between the transgenic fly and the appropriate 'driver', a fly expressing GAL4 in the desired cell type detailed in Figure 5-1.

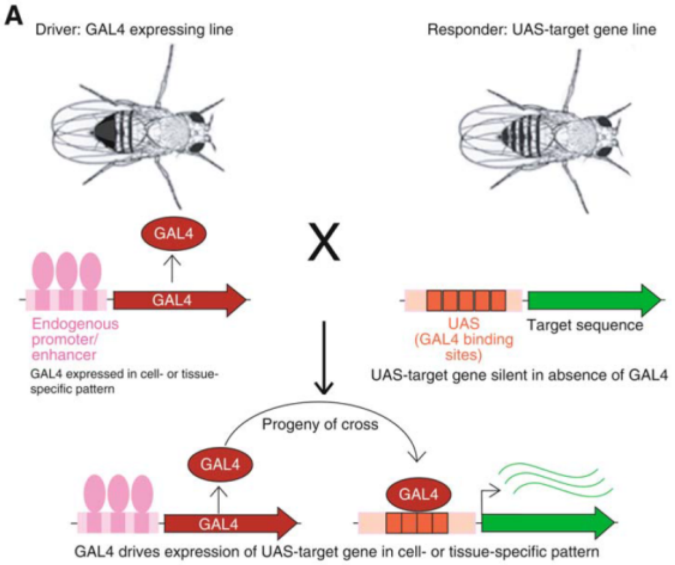

A
Driver: GAL4 expressing line
Responder: UAS-target gene line
GAL4
X
GAL4
Endogenous promoter/ enhancer
GAL4 expressed in cell- or tissue-specific pattern
UAS (GAL4 binding sites)
Target sequence
UAS-target gene silent in absence of GAL4
Progeny of cross
GAL4
GAL4
GAL4
GAL4 drives expression of UAS-target gene in cell- or tissue-specific pattern

Figure 5-1 GAL4/UAS system.

GAL4 is a transcriptional activator from yeast, which is expressed in a tissue-specific manner. UAS (Upstream Activation Sequence), an enhancer to which Gal4 specifically binds to activate gene transcription. Picture adapted from (Elliott and Brand, 2008)

So, the cell specific inactivation of any virtual gene is achieved by GAL4-UAS system in fly (Hardy et al., 2010). There are now a number of GAL4 (harbouring) fly lines, RNAi fly lines in the fly stock centres around the world (e.g. Bloomington Stock Centre), and also a number of RNAi vectors in *Drosophila*

Genomic resource centre for (https://dgrc.cgb.indiana.edu) researcher use to achieve nearly any gene knockdown in specific tissue.

5.1.2 *Drosophila as a model for novel gene discovery*

Drosophila melanogaster has been used in genetics studies for almost a century. *Drosophila* is small, easy and cheap to raise with short developmental life cycle. These factors make it an ideal model organism to study homologous gene functions. It has a small genome around 180 Mb with around approximately 14000 protein coding genes. About 75% of known human disease genes have a recognizable match in the genome of fruit flies (Reiter et al., 2001) and 50% of fly protein sequences have mammalian homologs, so *Drosophila* is increasingly used as a translation model for human development, homeostasis and disease (Graveley et al., 2011; Spradling, 2006). Genetic markers are commonly used in *Drosophila* research, for example Genetic markers in combinations with P-element inserts, easily allow one to identify transgenic animals from non-transgenic animals. Major advantages of *Drosophila* over other organisms are the relative ease of genetic manipulation, and the worldwide *Drosophila* stock centres that produce RNAi line against every gene. GAL4-UAS system is used in *Drosophila* to achieve the inactivation of genes in tissue specific manner (Dow, 2007). UAS fly lines can be made in less than 3 months for a few 100 dollars, and different vectors are available for *Drosophila* that can achieve the function of RNAi or RNA overexpression so reverse genetics can be easily applied in *Drosophila* to search the functions of novel genes including novel noncoding gene (Roberts et al., 2011).

Germline transformation of *Drosophila* with engineered P-element represents one of the most powerful methods with which study the functions of genes (Rubin and Spradling, 1982; Spradling and Rubin, 1982). P-elements can be engineered to carry the transgene of interest, as well as a marker gene, allowing the flies with insertion to be identified with genetic markers. A number of vectors engineered with P-element, marker gene, UAS sequence to make UAS transgene flies, such as pRISE, are available for applying to the *Drosophila* unique GAL4-UAS system. Microinjection with the P-element constructs will integrate into the genome and be inherited stably in the progeny of transformed individuals by the maker selection. The germline transformation scheme is outlined in Figure 5-2.

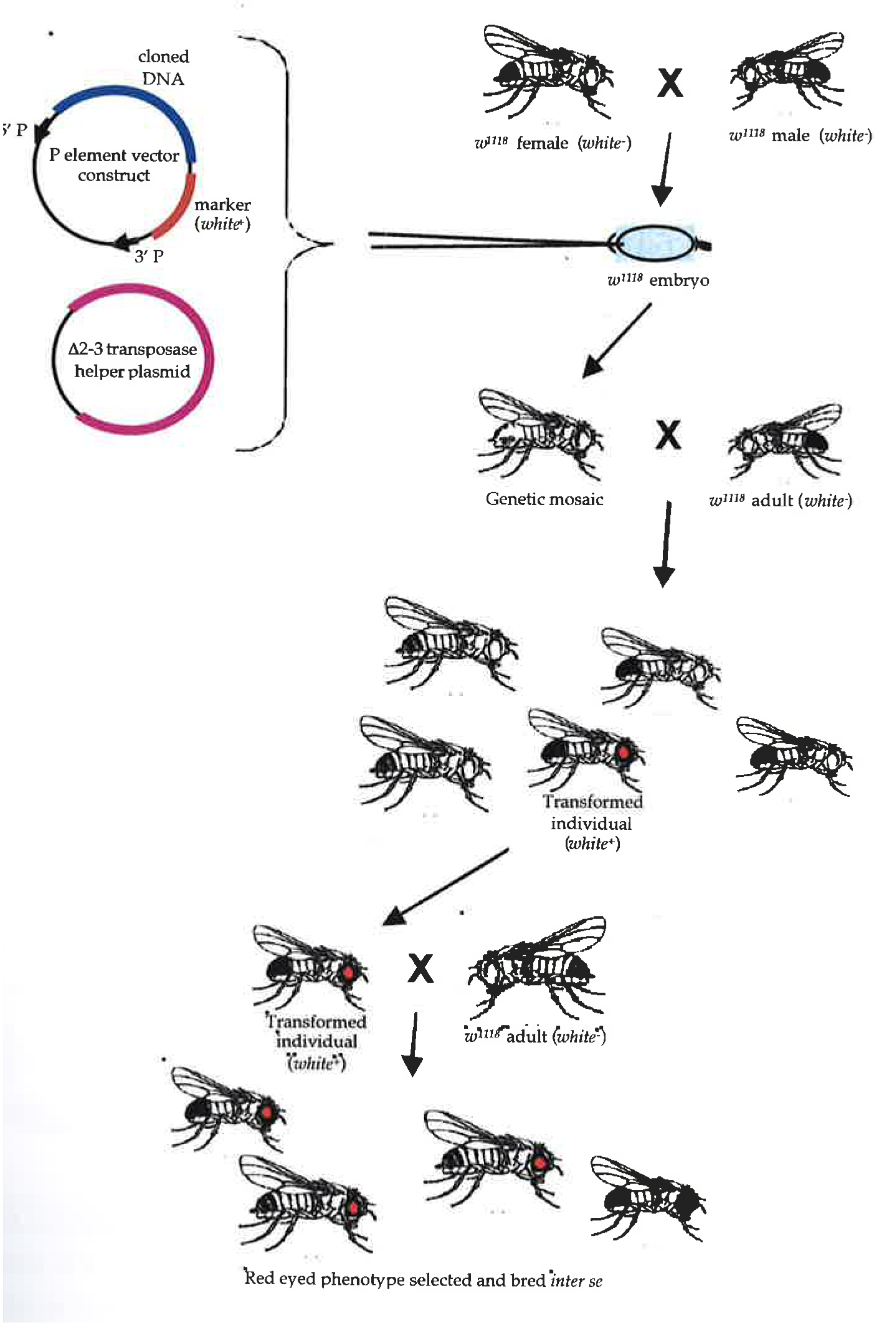

cloned DNA
5' P
P element vector construct
marker (white+)
3' P
Δ2-3 transposase helper plasmid
X
w1118 female (white-)
w1118 male (white-)
w1118 embryo
Genetic mosaic
X
w1118 adult (white-)
Transformed individual (white+)
Transformed individual (white+)
X
w1118 adult (white-)
Red eyed phenotype selected and bred inter se

Figure 5-2 Germline transformation of *Drosophila* embryos

Embryos from strain w^{1118} are co-injected with P-element with the gene of interest, marker gene and with the helper P-element plasmid that produces the functional transposase. Survival adults that potentially have the insertion will be back crossed with w^{1118}. The progeny of this cross will be screened and the flies chosen with the insertion by selecting the *white*$^{+}$ marker gene with P-element (red eyes). The progeny will be back crossed with w^{1118} and then successive generations will establish the transgene line containing the insertion, either in homogenous or heterozygous manner. Picture adapted from (Guo, 1996).

Drosophila is a common model organism for development (Graveley et al., 2011) and behaviour (Kaun et al., 2011) studies, but it is also a very useful model for physiology experiments because it is easy dissected and easy to use to make a physiology model (Dow and Davies, 2003). More recently, a number of innovative physiological techniques can be brought to bear on the transport or signalling process by using *Drosophila* tubules (Davies et al., 2012). These physiology technologies can help reverse genetics close the 'phenotype gap' by searching for the novel gene functions. The use of *Drosophila* as a model organism is also presented in chapter 1.3.1.

5.1.3 ***FlyBase annotation***

FlyBase (http://flybase.org) is an online bioinformatics database and the primary resource for molecular and genetic information on the *Drosophila* including 12 *Drosophila* species that had been sequenced.

Information in FlyBase originates from a variety of sources ranging from large-scale genome projects to the primary research literature. Data-types include sequence-level gene models, molecular classification of gene product functions, mutant phenotypes, mutant lesions and chromosome aberrations, gene expression patterns, transgene insertions, and anatomical images. FlyBase contains a complete annotation of the *Drosophila melanogaster* genome that is updated several times per year (Drysdale, 2008).

The database servers researchers of diverse backgrounds and interests, and offers several different query tools to provide efficient access to the data available and facilitate the discovery of significant relationships within the database (Wilson et al., 2008). Query tools, including the simple search tools QuickSearch and Jump to Gene are designed to help users navigate to a report page where information related to the object is presented, Other tools, such as GBrowse and the new Interactions Browser, highlight relationships between objects through a graphical interface, while QueryBuilder provides users with the ability to perform complex multi-step queries across all fields and different data sets. FlyBase also includes the recent availability of genome-wide data from the modENCODE project, next generation sequencing data (McQuilton et al., 2012).

5.1.4 *Structure and functions of Drosophila Malpighian Tubules*

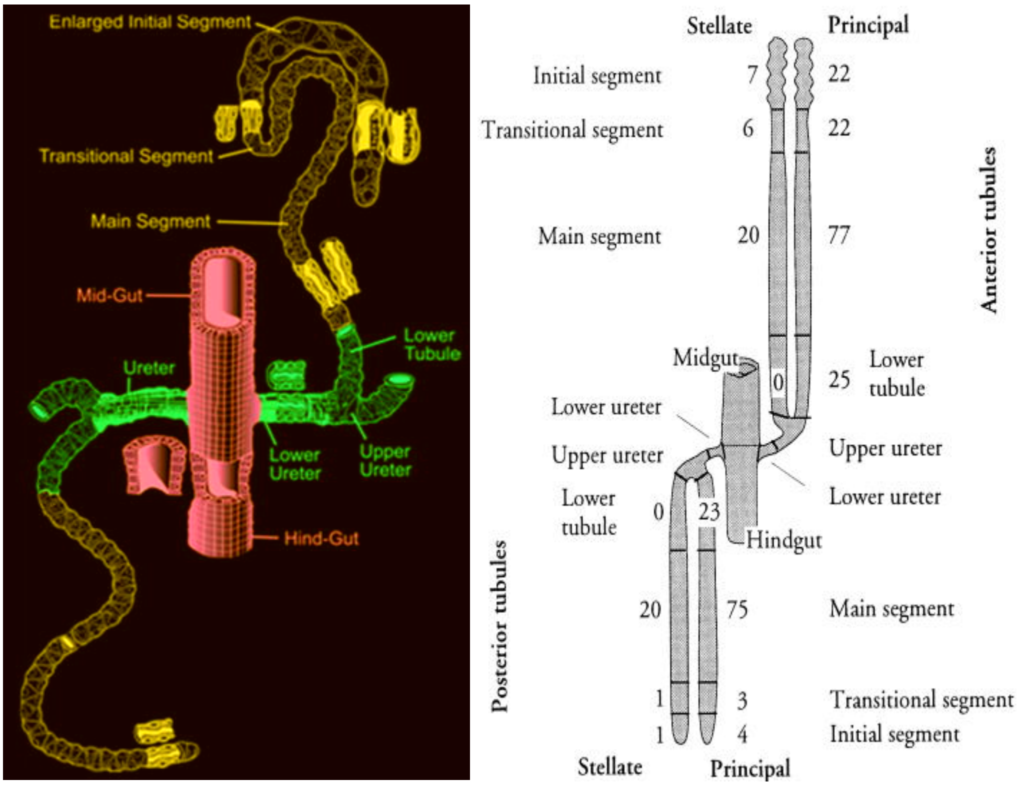


Figure 5-3 Tubules of *Drosophila* melanogaster.

Drosophila has two pairs of tubules, namely anterior and posterior tubules that ramify anteriorly and posteriorly in the body respectively. Morphologically and functionally distinct domains are labelled for anterior tubules; their posterior counterparts have the equivalent domains except that they don't have the enlarged initial segment. Left picture [Adapted from (Wessing A, 1978)].

Malpighian tubules domains as identified from enhancer trap analysis. The numbers of principal and stellate cells in each region are shown, as deduced from ethidium bromide staining. Standard errors are <1 in each case. Right picture [Adapted from (Sozen et al., 1997)].

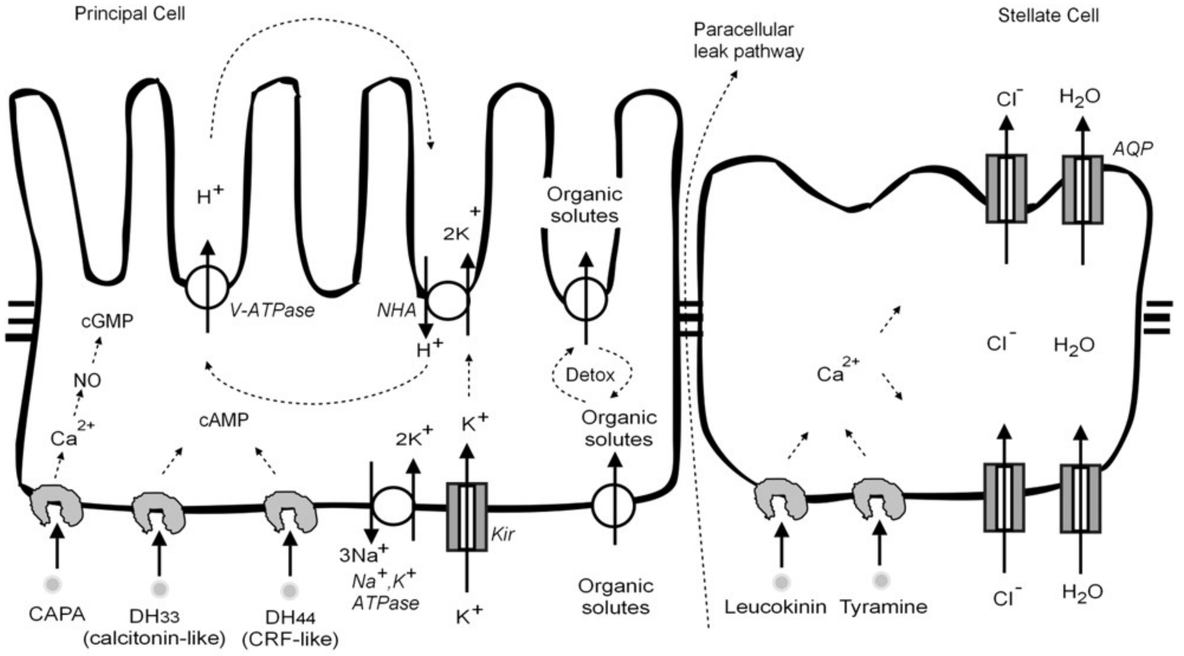


Figure 5-4 Two major cell types of tubules.

The larger, principal cells and smaller, stellate cells are joined together with septate junctions. These two cells surround the tubule lumen, thus separating the lumen from hemolymph. Details about the transport and signalling pathways are given in the texts. Picture [Adapted from (Dow and Romero, 2010)].

The two anterior Malpighian tubules are classically described as comprising a distal initial segment and a proximal main segment, joined by a narrow transitional segment; the two posterior tubules, in contrast, were thought to consist solely of a main segment. Contemporary studies, using enhancer trap lines, which place reporter genes under the control of tissue specific enhancers, confirm this viewpoint and thus the nomenclature "initial," "transitional”, "main", and “lower tubule” segments has been adopted to describe these genetically deduced domains (Sözen, 1997) (Figure 5-3).

It has been reported that the initial segment of *Drosophila* anterior tubule does not secrete detectable fluid, that the lower third of the tubule is reabsorptive, and that only the main segment is responsible for fluid production.

The initial segment is unique to anterior Malpighian tubules. Although the cells of the initial domain are thin and do not display prominent structural adaptations for ion transport, this region is excreting calcium at extremely high rates (Chintapalli et al., 2012; Dube et al., 2000). and a peroxisome-targeted isoform of SpoCk (Southall et al., 2006) on initial segment vesicles was confirmed by direct recording of peroxisomal calcium showing that the major peroxisomal calcium pool in tubules was in initial segment (Chintapalli et al., 2012). The homothorax/dorsotonals transcription factors are expressed exclusively in the initial segment of the right-hand tubules (Chintapalli et al., 2012; Wang et al., 2004).

The middle region of tubule, which is referred to as the main segment, generates the primary urine and plays a key role in excretion and osmoregulation. Two special cell types are involved, metabolically active principal (type I) and smaller intercalated secondary or stellate (type II) (Dow and Davies, 2003), which are joined together with septate junctions (Figure 5-3).

Four classes of transporters dominate the metabolically active principal cell: a basolateral Na^+- K^+-ATPase (Torrie et al., 2004) and basolateral inward-rectifier K^+ channels (Evans et al., 2005), which secrete K^+ into the lumen, and apically, a plasma membrane V-ATPase (Davies et al., 1996), which is a vital proton pump, and an alkali-metal/proton exchanger of the NHA (Day et al., 2008). This provides the first evidence that cation transport into the lumen of Malpighian tubules may be a unique property of principal, rather than type II cells.

The Basolateral Na^+-K^+- ATPase is an ouabain-sensitive, electrogentic ion pump responsible for maintaining the balance of sodium and potassium ions. It highly expressed in *Drosophila* tubule and only a single gene appears to encode the α-subunit of the Na^+-K^+- ATPase in *Drosophila* (ATPalpha) (Lebovitz et al., 1989). A further reported two genes are present in the *Drosophila* genome that are similar to the α and β subunits of Na^+/K^+-ATPase (Okamura et al., 2003). The Dow/Davies group first reported using microarray technology, that there are at least two genes to encode the α-subunit and five genes to encode the β-subunit. Na^+/K^+-ATPase may be an important part of models of tubule function.

An apical plasma membrane V-ATpase is an energizing plasma membrane proton pump. It is a large holoenzyme of at least thirteen subunits, encoded by thirty-one *Drosophila* genes. V-ATPase energizes animal plasma membrane for secretion and absorption of ion and fluid by imposing a transmembrane H^+ (proton) (Harvey and Wieczorek, 1997; Wieczorek et al., 2003). It has been recognized as the main energized pump in *Drosophila* tubules.

The basolateral inward-rectifier K+ channel is an actively pumped in tubules, and the main basolateral entry step is via barium-sensitive potassium channel, both in tubule and in other V-ATPase driven insect epithelia.

An apical plasma membrane NHA is an alkali-metal/proton exchanger, co-expressed with ATPase genes in apical plasma membrane. Their expression level affects epithelial transport of both Na^+ and K^+, one gene preferring Na^+ and the other gene preferring K^+ (Day et al., 2008).

The smaller stellate cells of tubule (type II cells) are distributed evenly throughout the initial, transitional and main segments of posterior tubules and within the main segment of anterior tubules. The main segments have hormonally-regulated chloride conductance pathways. There are three CLC-type chloride channels in the *Drosophila* tubules. A water flux pathway is also localized in tubule stellate cells (Dow and Davies, 2003). In addition to secretion of urine, *Drosophila* tubules are also involved in calcium excretion (Chintapalli et al., 2012), in immunity (Davies et al., 2012), in metabolism (Bratty et al., 2012) and detoxification of both endogenous solute and xenobiotics (Yang et al., 2007).

The tight (in insects, septate) junctions, which are between principal cell and stellate cells, may also contribute to the leak pathway for chloride movement (Dow and Romero, 2010) (Figure 5-4).

Secretion by Malpighian tubules is under hormonal control, including the insect kinins (e.g., leucokinin), corticotropin-releasing factor (CRF)-related diuretic hormones (CRF_related DH), calcitonin-like diuretic hormones (CT-like DH) and capa peptides. Capa peptide action is diuretic; via elevation of nitric oxide, cGMP and calcium in the principal cells of the Malpighian tubules.

Capa peptides were first identified as cardioacceleratory peptides (CAPs) CAP1 and 2 by Tublitz and Truman from the ventral nerve cord of *Manduca sexta* (Tublitz and Truman, 1985a, b, c). CAP2b, a cardioacceleratory peptide, is present in *Drosophila* and stimulates Malpighian tubule fluid secretion via cGMP, which in turn stimulates the nitric oxide signalling pathway (Davies et al., 2013; Terhzaz et al., 2013). Liquid chromatography analysis of adult *Drosophila* reveals the presence of a CAP2b-like peptide, which coelutes with *Manduca sexta* CAP2b, and synthetic CAP2b and that has CAP2b-like effects on the *M. sexta* heart. CAP2b stimulation elevates tubule cGMP levels but not those of cAMP. Both CAP2b and cGMP increase the transepithelial potential difference, suggesting that stimulation of vacuolar ATP action underlies the corresponding increases in fluid secretion (Davies, 1995).

Other hormones likely to be involved in Malpighian tubule function are the leucokinins, a group of widespread insect hormones. In tubules, their major action is to raise chloride permeability through stellate cells by binding to receptors on the basolateral membrane, and so ultimately to enhance fluid secretion. The action of Leucokinin is additive to both cAMP and cGMP but not to thapsigargin, suggesting that leucokinin acts by elevation of intracellular calcium (Dow, 2012; Kerr et al., 2004).

5.1.5 *Experimental plan*

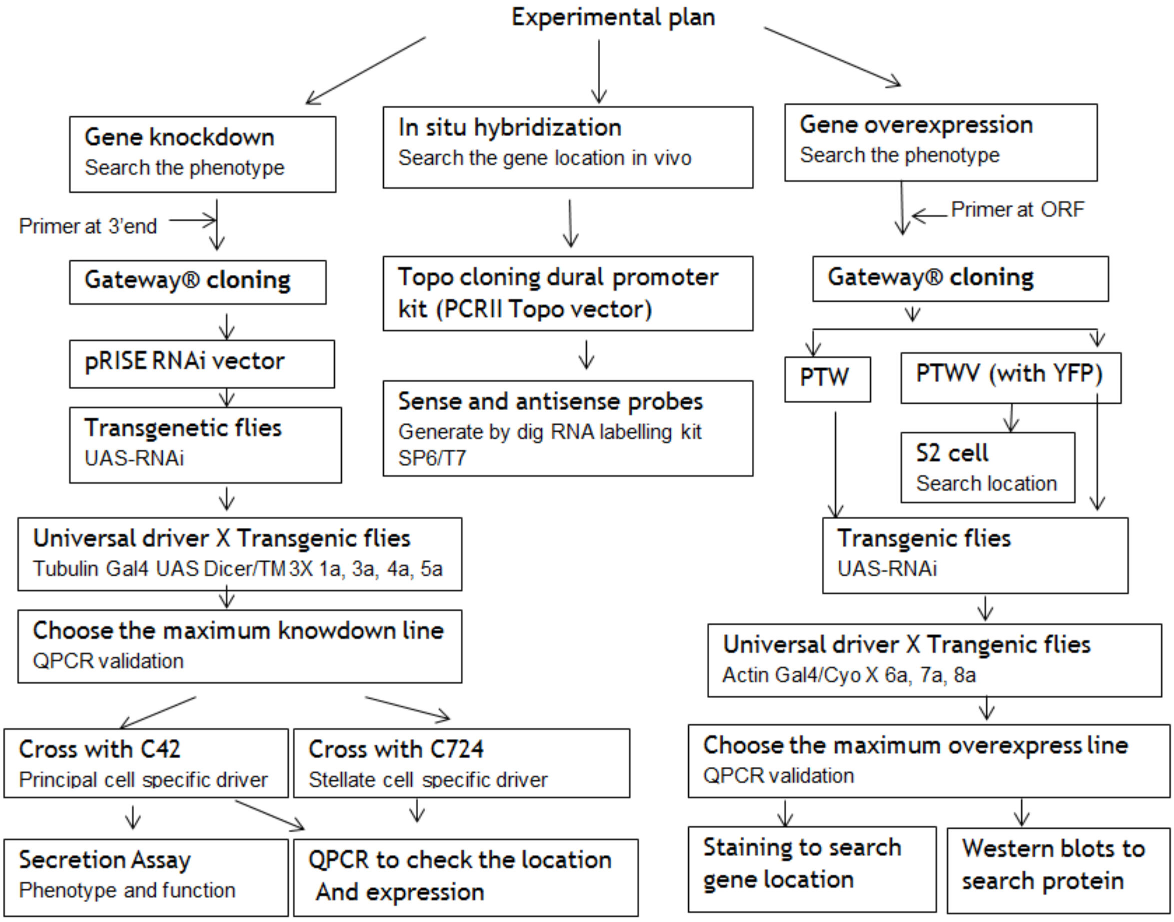

Experimental plan
Gene knockdown
Search the phenotype
In situ hybridization
Search the gene location in vivo
Gene overexpression
Search the phenotype
Primer at 3'end
Primer at ORF
Gateway® cloning
Topo cloning dural promoter kit (PCRII Topo vector)
Gateway® cloning
pRISE RNAi vector
PTW
PTWV (with YFP)
Transgenetic flies
UAS-RNAi
Sense and antisense probes
Generate by dig RNA labelling kit
SP6/T7
S2 cell
Search location
Universal driver X Transgenic flies
Tubulin Gal4 UAS Dicer/TM3X 1a, 3a, 4a, 5a
Transgenic flies
UAS-RNAi
Choose the maximum knowdown line
QPCR validation
Universal driver X Trangenic flies
Actin Gal4/Cyo X 6a, 7a, 8a
Cross with C42
Principal cell specific driver
Cross with C724
Stellate cell specific driver
Choose the maximum overexpress line
QPCR validation
Secretion Assay
Phenotype and function
QPCR to check the location
And expression
Staining to search gene location
Western blots to search protein

Figure 5-5 Experiment plan for searching for the function of novel genes by the reverse genetics method.

Figure 5.5 illustrates the reverse genetic experimental plan to verify the function of the novel gene that we found by RNA-seq analysis. It is an example of applying the reverse genetic technique to search the novel gene function. The overall plan is that knockdown and overexpression this gene to search for the phenotype, with *in situ* hybridization to search for the location. The details of this plan are:

1. Cloning of the novel *Drosophila* gene Chr3L 23777200-2378100 to RNAi vector (pRISE) and gene overexpression vector (PTW and PTWV) for *in vitro* (S2 cells) and *in vivo* (fruit flies) functional analysis.

2. Transfection of S2 cells with YFP fusion of Chr3L 23777200-23781000 constructs (PTWV) for fluorescent localisation within the cell. This would distinguish plasma membrane form endosomal (e.g., peroxisome,

mitochondria, ER, Golgi and others) localisation. This also would check if the overexpression construct has worked in order to further generate the overexpression transgenic flies.

3. Validation of overexpression and RNAi fly lines using qPCR, by crossing them with several GAL4 drivers (tubule principal cell, stellate cell specific *etc*). This shows if the mRNA levels are affected in the overexpression and RNAi flies driven using ubiquitous GAL4 lines. The 'cell-specific knockdown or overexpression' validations show in which cells this new gene is expressed.

4. Assessment of the phenotypic characters (including survival, structural, morphological defects, assay fluid secretion rates of Malpighian tubules). The percentage of survival rate would indicate if this gene is lethal and affects which developmental stage. The observation of structures and morphological characters would show if the novel gene plays a role in any developmental and morphological defects. The assay of the fluid secretion rates is to see whether knockdown or overexpression causes any impairment. This would suggest if this gene functioned in tubule secretion. If the secretion rate changes, on addition of the neuropeptides capa and leucokinin, this would suggest this change is caused by principal cell (cationic pathway) or stellate cell (anionic pathway) pathway.

5. *In vivo* localisation of YFP tagged novel gene, using different GAL4 drivers to see the cellular location of these proteins in different tissues of the fly and in specific cells of an individual tissue.

6. Western blot using the anti-GFP antibody, to see if the gene encodes a protein and the size of the protein. Because the novel gene encodes protein fused by YFP, the protein size should be the YFP plus the novel gene-encoded protein size.

7. *In situ* hybridization will show *in vivo* the localisation of this novel gene in tissue.

The results aim to use reverse genetic method to present the location and phenotype of the novel gene, then we will draw the possible function of this gene from it.

5.2 **Results**

5.2.1 ***Cufflinks result of novel gene Chr3L 23777200-23781000***

The Cufflinks report from the merged.gtf file and Cuffdiff report from isoform-_exp.diff show that the novel gene Chr3L 23777200-23781000 has two exons: Chr3L Cufflinks Exon 1 23777294-23777433, and Exon 2 23777506-23780611. The RPKM expression level of this isoform for whole flies is 0.612365, and for tubules is 86.539. The tubules/whole flies fold change is 141.319. The class _code of this gene is "u" (novel gene). Of the novel genes discovered by this project using RNA-seq with the TopHat and Cufflinks pipeline, this gene had the highest fold change. This gene had no annotation in FlyBase 5.34, and was subsequently named as CG43968. This gene was chosen for further study in this project because all the results indicated that this gene showed tubule-enriched expression with two isoforms, which were not previously annotated and are located in an intergenic region of the *Drosophila* genome. The ORF search (see 5.5.4) indicated that this gene is a protein coding gene which is easy to use reverse genetic method to search gene function. Figure 5-6 shows the merged.gtf file generated by Cufflinks as viewed in the Integrated Genome Browser (IGB), and depicts region Chr3L 23777200-23781000 indicating the novel transcript structure predicted by Cufflinks.

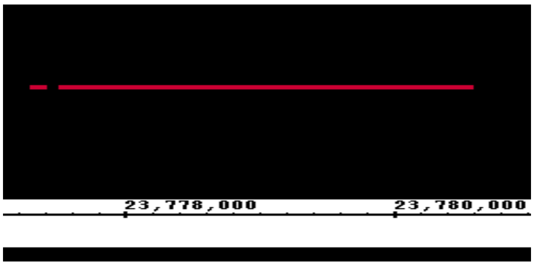


Figure 5-6 The novel transcript structure predicted by Cufflinks as viewed in IGB.

This figure shows the novel transcript in region Chr3L 23777200-23781000 has two exons. The small red bar on the left indicates a small exon, and the large red bar on the right indicates a second large exon. The gap in the middle between these two red bars indicates the intron. The scale along the bottom of the figure shows the region within chromosome 3L.

Strand-specific RNA-seq tubules data from the file transcript.gtf generated using Cufflinks were viewed using Tablet (version 1.12.03.26), and showed the predicted transcript and its direction (Figure 5-7). Figure 5-7 indicates that this transcript is on the sense strand and has two exons.

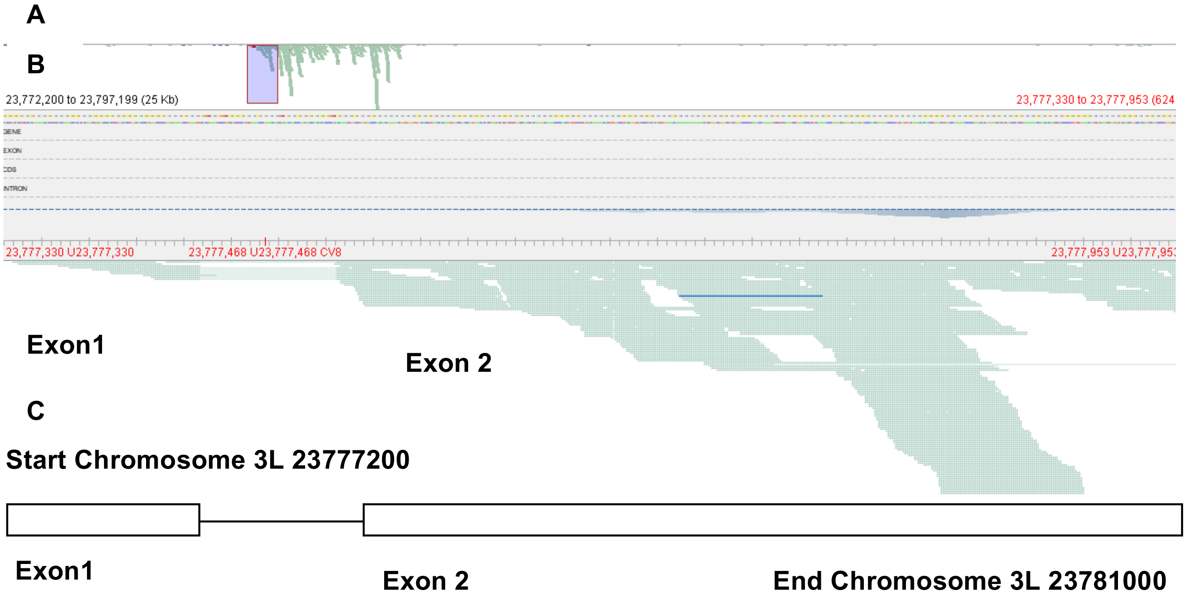

A
B
23,772,200 to 23,797,199 (25 Kb)
23,777,330 to 23,777,953 (624
GENE
EXON
CDS
INTRON
23,777,330 U23,777,330
23,777,468 U23,777,468 CV8
23,777,953 U23,777,95
Exon1
Exon 2
C
Start Chromosome 3L 23777200
Exon1
Exon 2
End Chromosome 3L 23781000

Figure 5-7 Chr3L 23777200-23781000 as viewed in Tablet.

This figure of the Tablet visualization shows the novel transcript has two exons (Exon1 and Exon 2) and is on the sense strand of chromosome 3L. Part A shows an overview of a region of chromosome 3L (23772200-23797199, 25KB). The subregion of interest (23777200-237810000) is highlighted using the red box. Part B shows a more detailed view of CG43968 between 23777200-23781000 of chromosome 3L. This detailed view shows the reads obtained for the exons in the small region at the beginning and the large region at the end of CG43968. No reads are shown in the region in between these two indicating an intron. Part C presents a schematic view of the gene structure, identifying the region of the first, small exon and the second, large exon.

5.2.2 ***FlyBase annotation of novel gene Chr3L 23777200-23781000***

FlyBase 5.34 (http://flybase.org) has no annotation for this gene (Figure 5-8).

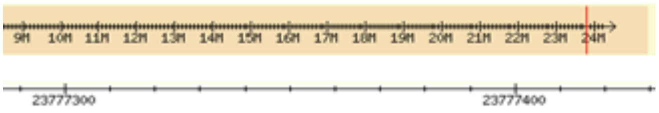


Figure 5-8 FlyBase 5.34 showing no gene model in region 23777200-23781000.

This is a visualization of Chromosome 3L (region 23777200-237810000) for FlyBase 5.34. An overview of Chromosome 3L is shown in the top part of the figure, with the subregion of interest highlighted by the red line. A more detailed view of the region is shown in the bottom part of the figure. There is no gene model is indicated under this gene region.

However recent modENCODE RNA-seq trace in FlyBase 5.36 showed the novel exon junction has been found by TopHat in *Drosophila* (Figure 5-9). This indicates a potential transcript in this region with two exons so there must be a gene here that is not annotated yet in this version.

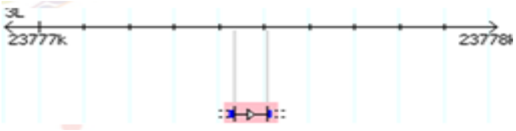


Figure 5-9 FlyBase 5.36, modECODE RNA-seq data.

The RNA-seq *Drosophila* data shows there is a novel splice junction (the lower pink bar) in region 23777200-23781000 (the upper region bar) highlighted by RNA-seq *Drosophila* data, indicating that there is a potential transcript in this area.

Very recently (16/08/2012), this gene has been annotated in FlyBase 5.48 with the symbol Dmel\CG43968 (FBgn0264699), but no further data are provided. It is a protein-coding gene from *Drosophila melanogaster*. Gene sequence location is 3L: 23777331-23780505 (http://flybase.org). The transcript is displayed in FlyBase Gbrowser as shown in Figure 5-10.

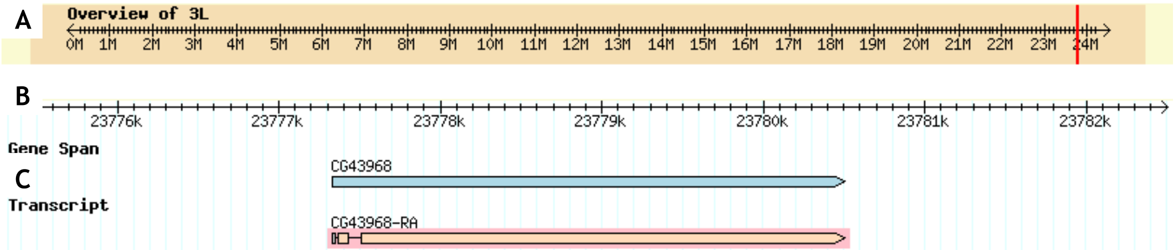


Figure 5-10 Novel gene CG43968 in FlyBase 5.48.

Part A shows an overview of chromosome 3L with the region of interest indicated by the red bar. Part B shows a more detailed view of chromosome 3L in the region of interest. Part C shows the gene model for CG43968 from FlyBase 5.48 using the blue bar, and the transcript CG43968-RA using the pink bar.

5.2.3 ***Drosophila Tiling microarrays analysis result***

Drosophila tubule Tiling 2.0R Array results were analysed by Tiling Array analysis Software (TAS). Berkeley Drosophila Genome Project version 5 (BDGP) was used as reference genome. The tubule.bed file is generated by using the parameters bandwidth of 60, default value for threshold of 4, maximum gap of 80 and minimum run of 40. The tubule.bed file is viewed in the Integrated Genome Browser (IGB) with reference genome from BDGP version 5. The position of the novel gene Chr3L 23777200-23781000 is shown in Figure 5-11. The novel transfragment of gene CG43968, which is indicated by the blue bar, was only showing as part of one exon, corresponding to the more highly expressed part of CG43968. This is because the background noise in tiling arrays due to cross hybridization affects the calculation of the gene expression. The lower expression exon of CG43968 was cut with the background.

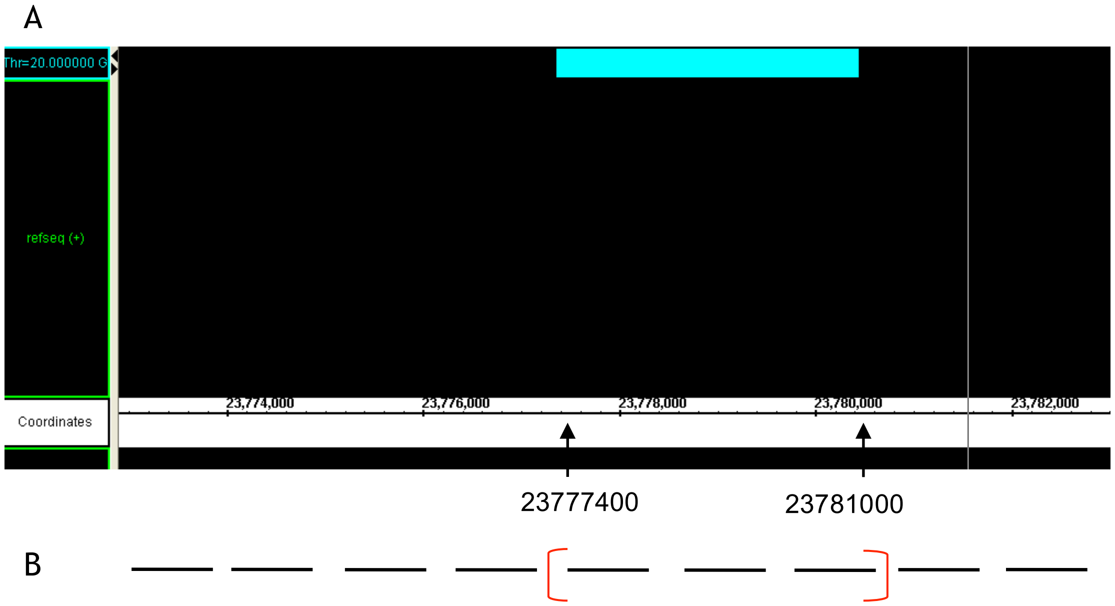


Figure 5-11 *Drosophila* tubule tiling microarray result from TAS as viewed using IGB.

Part A of this figure shows that the *Drosophila* tubule tiling microarray results as analyzed in TAS only indicate part of the transfragment highlighted by the blue bar of Chr3L 23777400-23781000 (CG43968). Only those expression levels above the background noise can be detected and so not all exons and introns for the transfragment can be distinguished. The middle of the figure shows that refseq has no gene model at this region. The scale bar at the bottom of the figure indicates the region of chromosome 3L. Part B shows the non-overlapping probe design of *Drosophila* Tiling Array 2.0 (resolution 39bp). The probes indicated within the red brackets correspond to the detected region of expression of the transfragment as indicated by the blue bar in Part A of this figure. Note, however, that this is only part of CG43968. The real transcript is much larger.

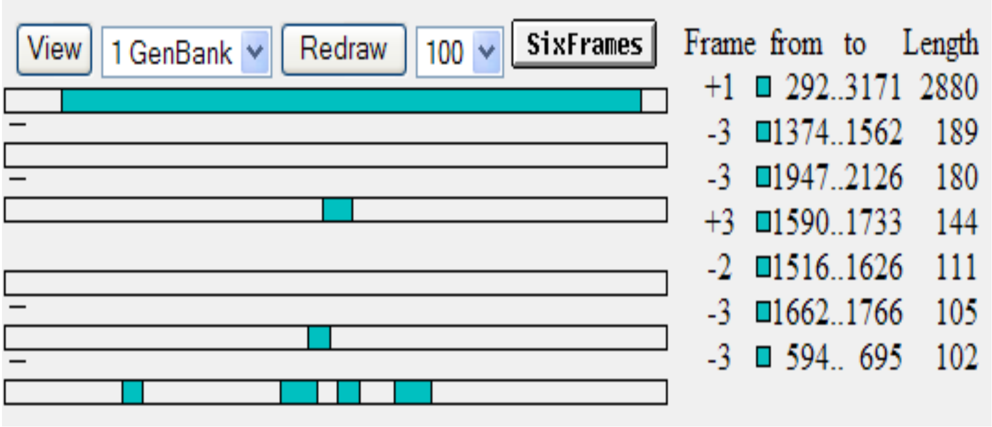
The result indicated the transfragment had been found in *Drosophila* tiling microarray but tiling microarray could not detect the splicing junction, so we cannot see the two exons in this transfragement. However RNA-seq can detect the boundary of the gene and the splicing junction, which is one of the advantages of RNA-seq over tiling microarray.

So the evidence is that *D. melanogaster* has a novel gene in this region. The rest of the chapter describes work to characterize this novel gene.

5.2.4 *Blast analysis of novel gene CG43968*

Is the gene unique to *D. melanogaster*, or is it found in other species? The BLAST analysis would increase our confidence in the gene assignment.

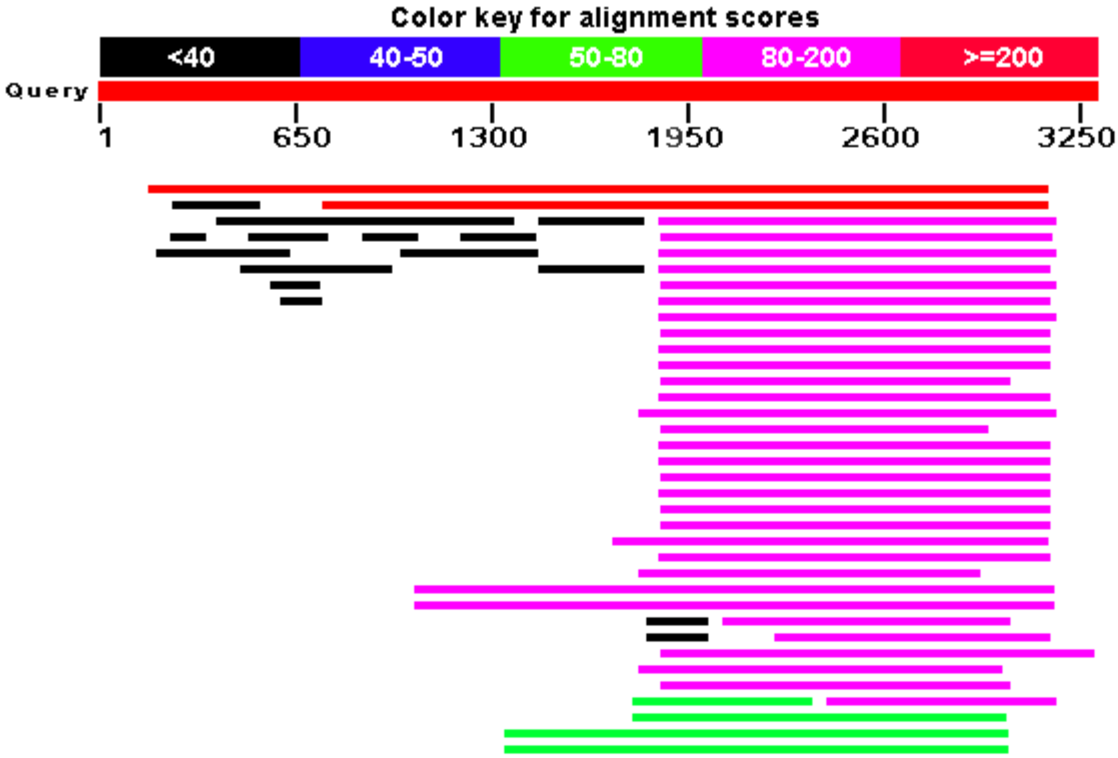
Translation BLAST (blastx), search protein databases using a translated nucleotide query. In order to find the predicted protein for novel gene Chr3L 23777200-23781000 (CG43968), blastx was performed using the nucleotide sequence of gene CG43968 to search the NCBI database to determine if the gene codes for a protein that is found in other organisms (Figure 5-13). The search took a DNA sequence and determined the sequence of six reading frames from both strand of DNA then used these sequences to search the protein database. Six frame searches found six possible open reading frames (ORF) for this gene (Figure 5-12). Because the strand-specific RNA-seq result indicated the novel gene CG43968 was on sense strand, so the top ORF possibility had more chance.

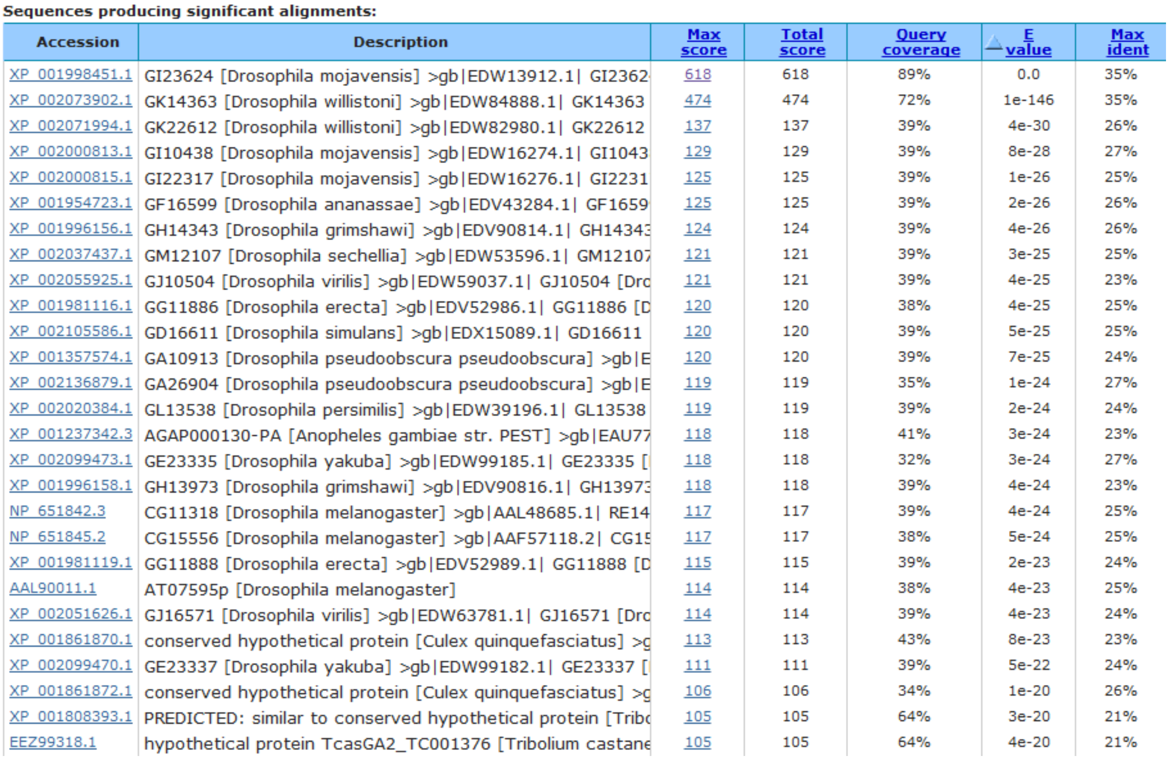


Frame	from	to	Length
+1	□ 292	3171	2880
-3	□1374	1562	189
-3	□1947	2126	180
+3	□1590	1733	144
-2	□1516	1626	111
-3	□1662	1766	105
-3	□ 594	695	102

Figure 5-12 Open reading frame search results for novel gene CG43968.

The results show the protein sequence similarity between novel genes and other species. The protein from the novel gene matched very well to ‘known’ genes in other *Drosophila* species such as *Drosophila mojavensis* and *Drosophila willistoni.* In addition, there were good similarities to other mosquito and beetle proteins. Six possible open reading frames of this gene had been found**.**





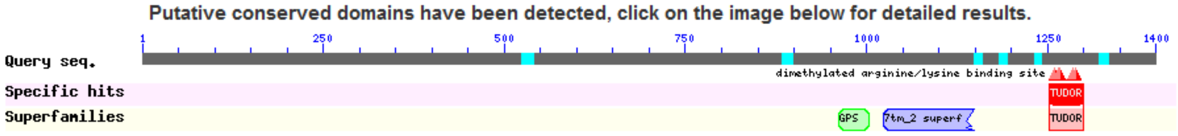
Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident
XP_001998451.1	GI23624 [Drosophila mojavensis] >gb|EDW13912.1| GI2362	618	618	89%	0.0	35%
XP_002073902.1	GK14363 [Drosophila willistoni] >gb|EDW84888.1| GK14363	474	474	72%	1e-146	35%
XP_002071994.1	GK22612 [Drosophila willistoni] >gb|EDW82980.1| GK22612	137	137	39%	4e-30	26%
XP_002000813.1	GI10438 [Drosophila mojavensis] >gb|EDW16274.1| GI1043	129	129	39%	8e-28	27%
XP_002000815.1	GI22317 [Drosophila mojavensis] >gb|EDW16276.1| GI2231	125	125	39%	1e-26	25%
XP_001954723.1	GF16599 [Drosophila ananassae] >gb|EDV43284.1| GF1659	125	125	39%	2e-26	26%
XP_001996156.1	GH14343 [Drosophila grimshawi] >gb|EDV90814.1| GH1434	124	124	39%	4e-26	26%
XP_002037437.1	GM12107 [Drosophila sechellia] >gb|EDW53596.1| GM1210	121	121	39%	3e-25	25%
XP_002055925.1	GJ10504 [Drosophila virilis] >gb|EDW59037.1| GJ10504 [Dro	121	121	39%	4e-25	23%
XP_001981116.1	GG11886 [Drosophila erecta] >gb|EDV52986.1| GG11886 [D	120	120	38%	4e-25	25%
XP_002105586.1	GD16611 [Drosophila simulans] >gb|EDX15089.1| GD16611	120	120	39%	5e-25	25%
XP_001357574.1	GA10913 [Drosophila pseudoobscura pseudoobscura] >gb|E	120	120	39%	7e-25	24%
XP_002136879.1	GA26904 [Drosophila pseudoobscura pseudoobscura] >gb|E	119	119	35%	1e-24	27%
XP_002020384.1	GL13538 [Drosophila persimilis] >gb|EDW39196.1| GL13538	119	119	39%	2e-24	24%
XP_001237342.3	AGAP000130-PA [Anopheles gambiae str. PEST] >gb|EAU77	118	118	41%	3e-24	23%
XP_002099473.1	GE23335 [Drosophila yakuba] >gb|EDW99185.1| GE23335 [	118	118	32%	3e-24	27%
XP_001996158.1	GH13973 [Drosophila grimshawi] >gb|EDV90816.1| GH1397	118	118	39%	4e-24	23%
NP_651842.3	CG11318 [Drosophila melanogaster] >gb|AAL48685.1| RE14	117	117	39%	4e-24	25%
NP_651845.2	CG15556 [Drosophila melanogaster] >gb|AAF57118.2| CG1	117	117	38%	5e-24	25%
XP_001981119.1	GG11888 [Drosophila erecta] >gb|EDV52989.1| GG11888 [D	115	115	39%	2e-23	24%
AAL90011.1	AT07595p [Drosophila melanogaster]	114	114	38%	4e-23	25%
XP_002051626.1	GJ16571 [Drosophila virilis] >gb|EDW63781.1| GJ16571 [Dro	114	114	39%	4e-23	24%
XP_001861870.1	conserved hypothetical protein [Culex quinquefasciatus] >g	113	113	43%	8e-23	23%
XP_002099470.1	GE23337 [Drosophila yakuba] >gb|EDW99182.1| GE23337 [	111	111	39%	5e-22	24%
XP_001861872.1	conserved hypothetical protein [Culex quinquefasciatus] >g	106	106	34%	1e-20	26%
XP_001808393.1	PREDICTED: similar to conserved hypothetical protein [Tribo	105	105	64%	3e-20	21%
EEZ99318.1	hypothetical protein TcasGA2_TC001376 [Tribolium castane	105	105	64%	4e-20	21%

Figure 5-13 Blastx search for the novel gene CG43968.

The blastx search found the similar protein sequence for the novel gene CG43968 in *Drosophila mojavensis* and *Drosophila willistoni*.

Although no conserved domains were found in *D. melanogaster* sequence, a reciprocal BLAST with the *D. willistoni* sequence identified possible domains. Reciprocal BLAST is a common computational method for predicting putative orthologues.



Putative conserved domains have been detected, click on the image below for detailed results.

Figure 5-14 Reciprocal Blast with the *D.willistoni* sequence identified the GPS domain.

GPS domain (G-protein-coupled receptor proteolytic site domain) had been found from the reciprocal BLAST using *D.willistoni* sequence (Figure 5-14). GPS Domain presents in latrophilin/CL-1, sea urchin REJ and polycystin. Polycystin is a protein that in humans is encoded by the *PKD1* gene Polycystic (Glücksmann-Kuis and Schneider, 1995; Hughes et al., 1995). *PKD1* is a kidney gene in humans!

Polycystin-1 is a glycoprotein, which contains a large N-terminal extracellular region, multiple transmembrane domains and a cytoplasmic C-tail (Figure 5-15).

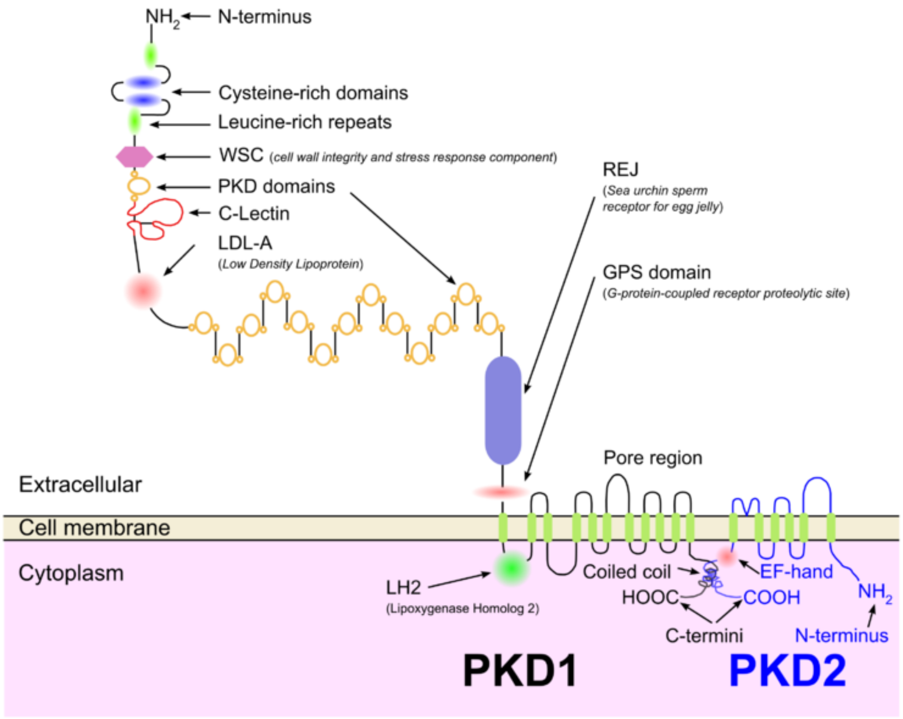


Figure 5-15 Illustration of PKD1 and PKD2 proteins.

The PKD1 (Polycystin-1) and PKD2 (Polycystin-2) proteins are at the cell membrane. The PKD1 (Polycystin-1) contains a large N-terminal extracellular region, and seven- transmembrane receptor (Secretin family) of the G-protein-coupled receptors (GCPRs). Picture derived from http://www.bio.davidson.edu/Courses/Molbio/MolStudents/spring2003/MaloneyH/Polycystins.html

From Blastx (Figure 5-13, the front half of the protein CG43968 is less well conserved, but the back half is more conserved in other species. The shape of the protein can be explained by the fact that the front half of the protein Polycystin-1 has a long extracellular N-terminus, and the back end has a conserved seven-transmembrane (7TM) receptor of the G-protein-coupled receptors (GCPRs). Polycystin-1 may modulate intracellular calcium homoeostasis and other signal-transduction pathways. It plays a role in renal tubular development as well.

5.2.5 *Protein localization prediction*

5.2.5.1 **PSORT II program for protein subcellular localization prediction**

PSORT (http://psort.hgc.jp/form2.html) is a free web-based tool used for the prediction of protein localisation sites in cells. It receives the information of an amino acid sequence and its taxon of origin (e.g. Gram-negative bacteria) as inputs. Then it analyzes the input sequence by applying the stored rules for various sequence features of known protein sorting signals. Finally, it reports the possibility for the input protein to be localized at each candidate site with additional information. PSORT was developed in 1990 and is applicable to bacterial and plant sequences. Although this technique is suited to integrating various kinds of information on protein sorting, the program requires manual adjustment of many numeric parameters. To overcome this difficulty, a new version, PSORT II was development in 1997 and is applicable to animal and yeast sequences; the *k*-nearest-neighbour method algorithm is used (Nakai and Horton, 1999). From the blastx search, the ORF sequence of novel gene CG43968 was obtained. The result of a subsequent enquiry by PSORT II is as follow:

PSORT II Prediction

30.4 %: cytoplasmic
17.4 %: vesicles of secretory system
13.0 %: nuclear
13.0 %: mitochondrial
8.7 %: Golgi
8.7 %: plasma membrane
4.3 %: endoplasmic reticulum
4.3 %: vacuolar

The PSORT II results provide support for the protein of novel CG43968 being located in cytoplasm of the cell.

5.2.5.2 **WoLF PSORT program for protein subcellular Localization prediction**

WoLF PSORT (http://wolfpsort.org) is an extension of the PSORT II program for protein subcellular location prediction and is applicable to fungi, animal, and plant sequences. WoLF PSORT converts protein amino acid sequences into numerical localization features; these features are based on sorting signals, amino acid composition and functional motifs such as DNA-binding motifs. After conversion, a simple *k*-nearest neighbour classifier is used for prediction. The evidence for each prediction is shown in two ways. Firstly, a list of proteins of known location is compared with the most similar localization feature to the query. Secondly, a table is provided of the values of each localization feature for the query and its neighbours (Horton et al., 2007). This subcellular prediction of the protein location will supply important information of the function of the protein. Unlike older programs such as PSORT and PSORTII that use one dimensional amino acid sequences of proteins, the WoLF POST uses feature selection and a flexible scoring model to increase accuracy and handle multiply localized proteins.

The ORF sequence of novel gene CG43968 is obtained from a blastx search. The result of a subsequent enquiry by WoLF POST is as follow:

Cytoplasm: 24.0%, mitochondrial: 4.0%, mito_peroxisome: 4.0%, peroxisome: 2.0%.

The result lends greater support to cytoplasmic localisation of protein for the novel gene CG43968.

5.2.6 *In situ hybridization to search for the location of novel gene CG43968*

PCR products derived from 3'UTR end of novel gene were cloned into TOPO® pCRII vectors with Sp6 and T7 dual promoter (details described in 2.19). The orientation of the PCR product was established using PCR. The T7 and Sp6 promoters of the pCR™II vector allowed *in vitro* transcription of the insert to produce sense or anti-sense products. The anti-sense probes then hybridized to mRNA within the tissue, and then the signal indicated the *in vivo* location of the gene within the tissue. The sense probes performed as a control to show the background to make sure the *in situ* hybridization signals were real (Figure 5-16).

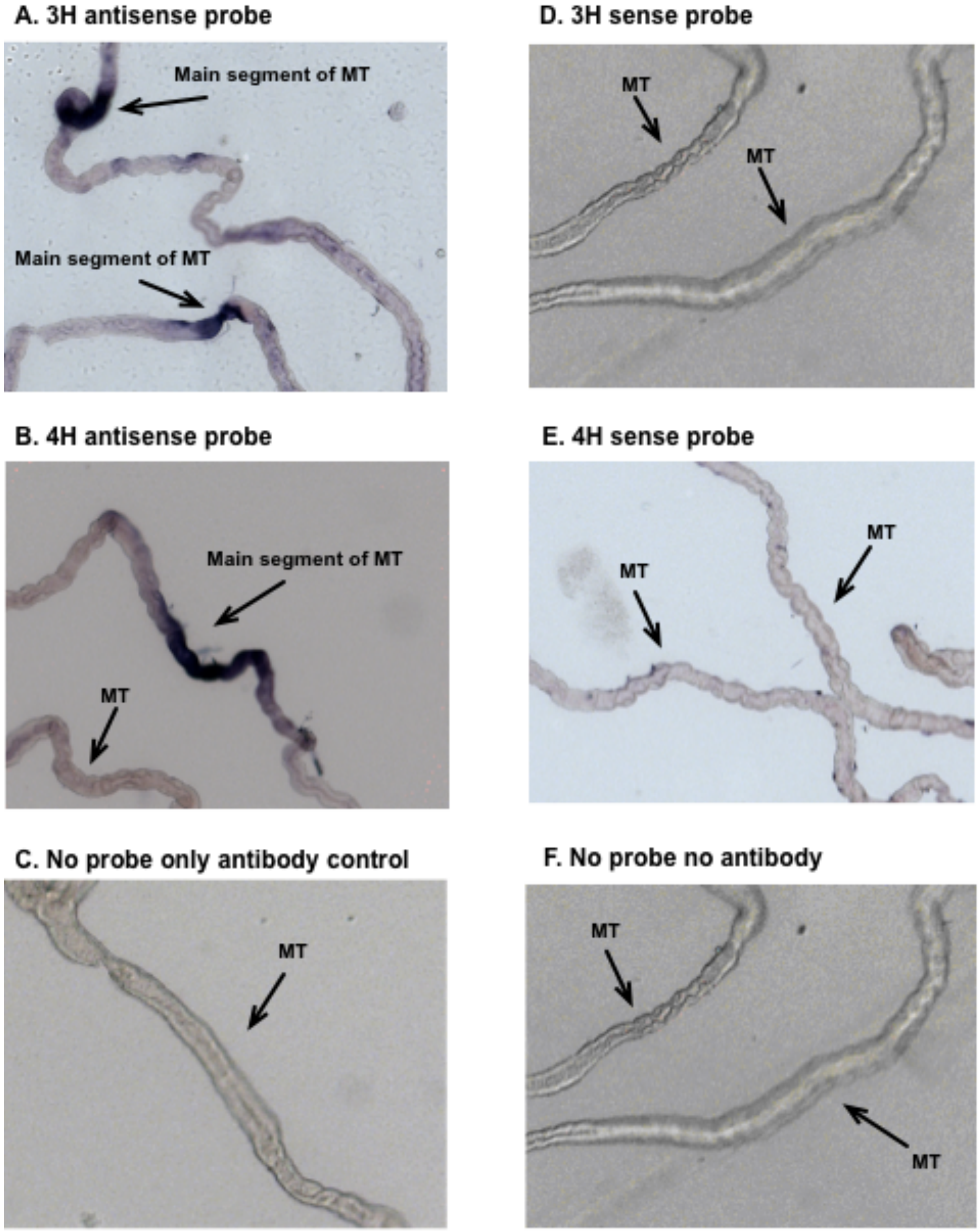


Figure 5-16 ***In situ*** hybridization images of novel gene Chr3L 23777200-23781000

This figure shows the images obtained from *in situ* hybridization analysis using 3H, 4H antisense and sense riboprobes for novel gene CG43968. A, B are different probes for the same gene only from the different PCR products at 3'UTR of novel gene CG43968, with the images showing strong expression in the main segment of the Malpighian tubule (MT). D, E were negative controls; sense probes corresponding to the antisense probes of CG43968 showed no stain in tubules. C, F all acted as controls to show the signal affected by antibody or the other factors during the *in situ* hybridization procedure, with the images showing no staining.

The main segments of tubules contain principal cells and stellate cells. The *in situ* hybridization could not distinguish the signals from the two cell types in the main segments. Principal cells are responsible for cation and organic metabolites secretion, and stellate cells are responsible for water and chloride secretion. Because the *in situ* hybridization could not clearly identify the gene expression in specific cells types, cell specific gene knockdown and qPCR was used to identify the location of novel gene expression in the next experiment.

5.2.7 ***Loss-of-function analysis (dsRNA knockdown analysis using pRISE vector for CG43968)***

5.2.7.1 **RT-PCR, pENTR/D-TOPO® vectors and sequencing**

RT-PCR primer design was performed using Invitrogen Primer design tool-perfect primer ™ designer (http://tools.lifetechnologies.com) and Snap Dragon dsRNA design (http://www.flyrnai.org). Invitrogen Primers designed two primers on different exons and spanning an intron. SnapDragon designed primers on the 3' end of the gene. The novel gene CG43968 sequence was obtained from Download Sequence Region (*Drosophila melanogaster* release 5.30, http://www.ncbi.nlm.nih.gov/projects/mapview/seq_reg). Four amplified fragments in this region were cloned into the pENTRY/D-TOPO (Invitrogen) vector using the Gateway technology. Details of the method are presented in section 2.15.3.

Construct 1: Primer design uses Invitrogen software. See Table 5-1 for details.

RT-PCR used primers designed on the two exons, so the transcript spanned an intron. This method also confirmed that this novel transcript was a real transcript not a genomic contamination. Result is shown in Figure 5-17. The PCR program used 94^{o}C 2mins then 94^{o}C 30s, 55^{o}C 30s, 72^{o}C 30s for 30 cycles, 72^{o}C 5 mins. The construct-1 was sequenced (Figure 5-18), and the splicing junction was confirmed.

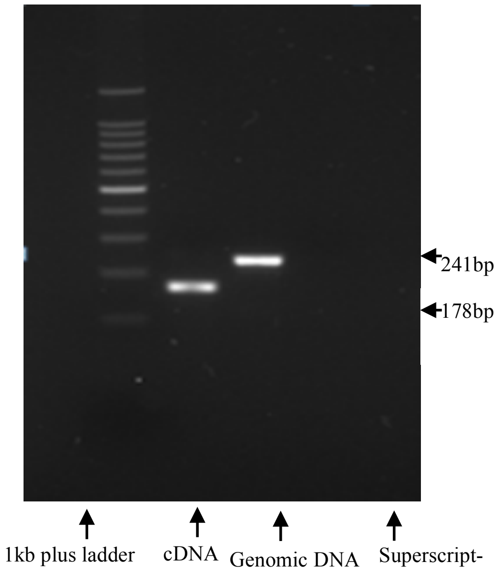
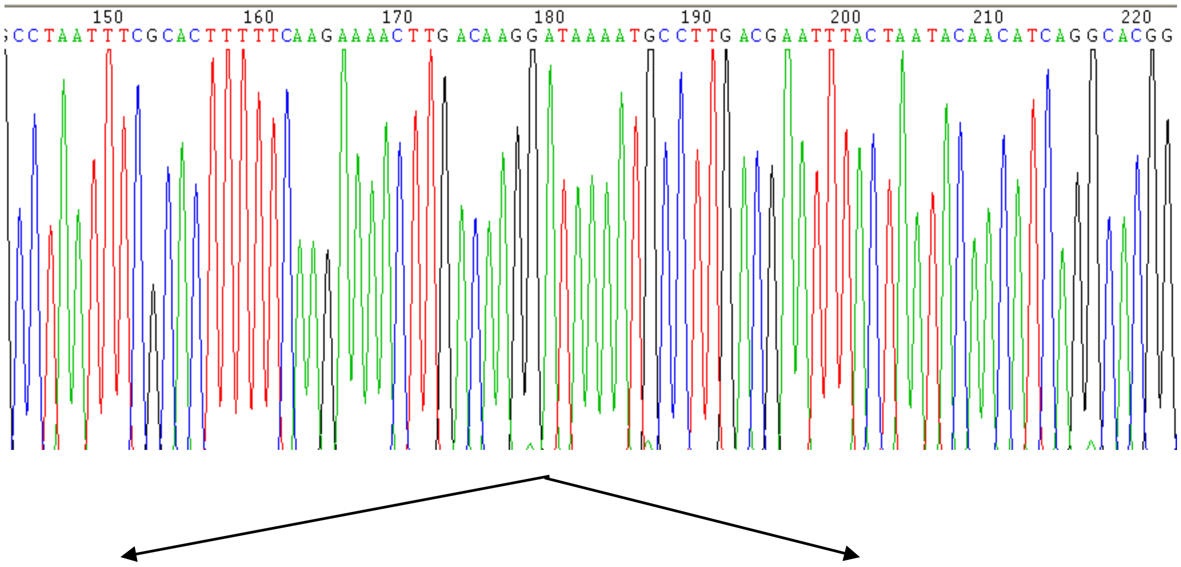


Figure 5-17 RT-PCR of novel gene CG43968.

RT-PCR used primers which spanned an intron showing two bands, cDNA (178bp) and Genomic DNA (241bp). It also confirmed that the intron was spliced out in the transcript. The superscript minus control indicated the cDNA samples did not have genomic contamination.



GTATGAAAAGTTTGAGTAGTTTTAGGTCAAATTTCAAA
TAAGTAGAATAATTTACGAATTAATATATTTTA

Figure 5-18 Sequence of CG43968 of pENTR/D-TOPO vector.

The intron (73bp) was confirmed of RNA-seq prediction after the CG43968 construct of pENTR/D-TOPO® vector was sequenced. The two black arrows show the sequence of the intron that was spliced out during reverse transcription.

The other three constructs (2,3,4) of pENTR/D-TOPO® vectors were made for the same gene CG43968 primers designed at the 3’ end of the gene by Snap Dragon, as detailed in Table 5-1.

Table 5-1 Primers of CG43968 for pRISE constructs.

pRISE construct	Primer	Primer sequence	Primer location	cDNA product (bp)	Genomic DNA product (bp)
Construct 1	Forward Reverse	GTAGCATTTATGTGCCTATTGC GATGGTTAGACTTAAGGCACA	Exon1 Exon2	173	241
Construct 2	Forward Reverse	CCTTGCAAGCTTCAACCAAT AAACGCCTTAAACGGCATAG	Exon2 Exon2	310	310
Construct 3	Forward Reverse	TTTGTTCATGGCGCATATTG TGTTGCGTTTAGCTCAGCAG	Exon2 Exon2	369	369
Construct 4	Forward Reverse	AGTTTCAAGCTATCGCACCG AATCCAAAACACAACGCACA	Exon2 Exon2	304	304

Summarizing the primers for *Drosophila* tubule-enriched novel gene CG43968 that were designed in SnapDragon, and used to generate pRISE constructs.

Sequences have been confirmed for these four constructs.

5.2.7.2 **Destination vector pRISE to generate UAS-RNAi line**

The four trigger sequences for the same gene from the pENTRY/D-TOPO vector using Gateway technology could be transferred easily to pRISE by an *in vitro* reaction mediated by LR Clonase. The RNAi constructs generated were based on the protocol by Kondo T *et al.* (Kondo et al., 2006) with the generation of construct containing inverted repeats the attR1-ccdB-attR2 cassette. This was achieved by cloning two identical fragments of the gene of interest into the vector of the Gateway cassette in the opposite orientation separated by a hairpin loop (intron), under UAS control that acted as dsRNAi into the tissue. Since pRISE carries a pentamer of UAS^{GAL4} in the promoter region, RNA silencing can be controlled by selecting appropriate ‘driver’ GAL4 transgenes (Duffy, 2002). Further functional analysis was performed when the RNAi transgenes flies were generated. Details of this method are presented are in section 2.15.3

5.2.7.3 **Making the transgenic fly lines and validation**

The pRISE plasmid was purified by columns. Insert size and direction of pRISE vectors were checked. 50 µg plasmid with concentration at least 0.5 µg/µl was sent to BestGene Inc (thebestgene.com) for microinjection to make transgene flies. The process is illustrated in Figure 1-6. The Gateway™ destination vector pRISE consists of a P-element, UAS sequence and mini-white marker for genomic integration, and coinjected with a helper P-element plasmid that produces a functional transposase in *Drosophila* germline, allowing insertion of the genetic payload. Using the *mini-white*$^{+}$ marker, the transformants were selected. The transgenic lines were made either homozygous or heterozygous (with balancer). Then, the essential process of validating the fly lines using GAL4/UAS bipartite system was performed by driving the transgene CG43968-RNAi expression using a variety of cell- or tissue-specific GAL4s or using a ubiquitous GAL4 such as Tubulin-GAL4 UAS Dicer-2/TM3, Sb; or Actin-GAL4/CyO. A schematic diagram of the cross scheme is shown in Figure 5-19.

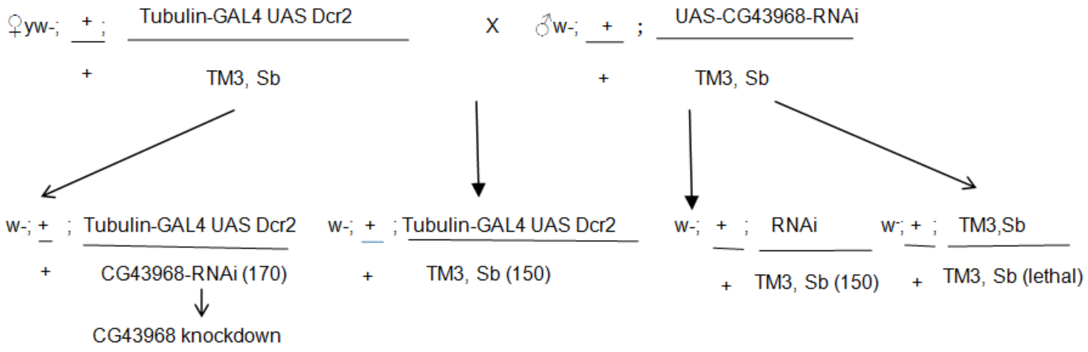


Figure 5-19 A schematic diagram of the RNAi knockdown cross scheme.

CG43968-RNAi crossed with ubiquitous GAL4 driver Tubulin-GAL4 UAS Dicer-2/TM3, Sb obtained CG43968-RNAi knockdown in adult whole flies. Four genotypes were produced (Tubulin-Gal4 UAS Dcr-2/CG43968-RNAi; Tubulin-Gal4 UAS Dcr2/TM3, Sb; CG43968-RNAi/TM3 Sb; TM3Sb/TM3, Sb). The number in the bracket indicated the survival numbers of the F1 progeny. The survival rate of each genotype is 1:1:1:0. The novel gene CG43968 is not lethal.

CG43968-RNAi was knocked down by using ubiquitous Tubulin-Gal4 UAS Dcr2/TM3, Sb driver in adult whole flies. The F1 progeny were also counted for the four genotypes to check the survival rate that confirmed the novel CG43968 was not lethal when after knockdown.

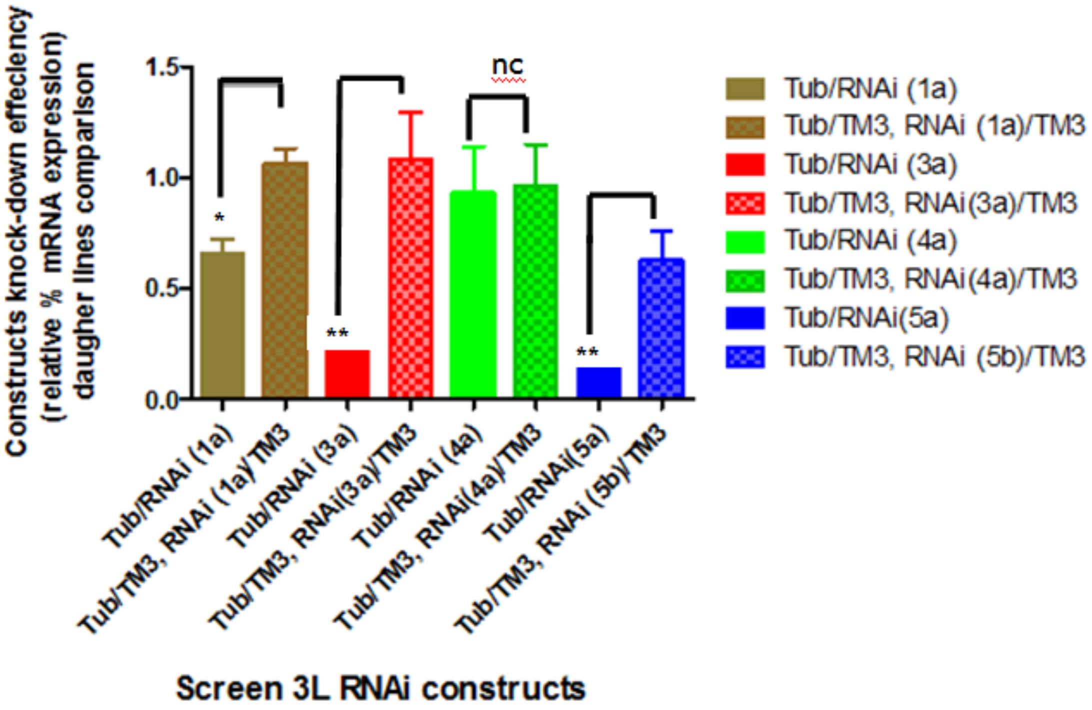


Figure 5-20 Screen of the UAS-CG43968 constructs.

The UAS-CG43968-RNAi (3a) achieved over 77% knockdown over control. The comparison of the knockdown efficiency of four fly lines was obtained from three UAS-RNAi constructs of novel gene CG43968 by using the ubiquitous driver GAL4. The percentage knockdown was 38% (1a), 77% (3a), 4% (4a), and 70% (5a) respectively. UAS-CG43968- RNAi (3a) achieved the most efficient knockdown.

qPCR performed gene expression comparison between the siblings of progeny to determine the efficiency of the knockdown as shown in Figure 5-20. RNAi (1a) was from CG43968 construct 1; RNAi (3a), RNAi (4a) were from construct 2, RNAi (5a) was from construct 4. The absolute percentage of knockdown for UAS-RNAi (1a), driven by Tubulin-GAL4, was 38% (*t*-test, $P < 0.01$) to its heterozygous siblings control. The absolute percentage of knockdown for UAS-RNAi (3a), driven by Tubulin-GAL4 UAS Dcr2, was 77% (*t*-test, $P < 0.05$) to its heterozygous siblings control. The absolute percentage of knockdown for UAS-RNAi (4a), driven by Tubulin-GAL4, was 4 % (*t*-test, $P >0.5$) to its heterozygous siblings control, 4a was not knocked down. The absolute percentage of knockdown for UAS-RNAi (3a), driven by Tubulin-GAL4, was 70% (*t*-test, $P < 0.05$) to its heterozygous siblings control. The four transgene fly lines showed only three lines contained the knockdown. UAS-RNAi (3a) and (5a) were more efficiently knocked down than UAS-RNAi (1a). UAS-RNAi (4a) was not knocked down. Consequently, at least one RNA line [UAS-CG43968-RNAi (3a)] produced a good knockdown (>75%).

5.2.7.4 **Effect of principal cell-specific knockdown of CG43968**

The following procedure was used to create the c42-GAL4/UAS-CG43968-RNAi (3a) line in fly tubule principal cells. The c42-GAL4 is a specific tubule principal cell driver. c42 was crossed with UAS-CG43968-RNAi/TM3, Sb. The progeny was c42/UAS-CG43968-RNAi, c42/TM3, Sb. The survival rate of F1 progeny is 1:1 for the two genotypes, so this gene was not lethal in principal cell.

In order to produce reliable comparison to see where this gene was expressed and to control the genetic background effect, we needed to create a single copy of the heterozygous parental line for the control. To achieve this, virgin females c42 were crossed with males UAS-CG43968-RNAi/TM3, Sb; virgin females c42 were crossed with male Canton S; virgin females Canton S were crossed with UAS-CG43968-RNAi/TM3, Sb. The F1 progeny were selected as c42-GAL4/UAS-CG43968-RNAi (normal hair), c42/Canton S, Canton S/UAS-CG43968-RNAi (normal hair).

The entire cross was set up at 26 ºC and 70-80% relative humidity (Duffy, 2002; Haley et al., 2003), three females and six males in one vial, allowed to mate for 48 hours and laid eggs. After 48 hours, the flies were transferred to new vials, and this was repeated another two times. The F1 progeny were selected according to the marker, for example, normal hair or short hair. The F1 progeny were transferred to fresh vial in every two days over seven days. The seven days flies were selected according to the marker, and equal numbers of females and males (3 males and 3 females) were flash frozen in liquid nitrogen then stored at -80 ºC.

5.2.7.5 **Effect of stellate cell-specific knockdown of CG43968**

The following procedure was used to create the c724-GAL4/UAS-CG43968-RNAi (3a) line in fly tubule stellate cells. The c724-GAL4 is a specific tubule stellate cell driver. c724 was crossed with UAS-CG43968-RNAi/TM3, Sb. The F1 progeny was c724-Gal4/UAS-CG43968-RNAi, c724-Gal4/TM3, Sb. The survival rate of the progeny was 1:1 for the two genotypes, so this gene was not lethal in stellate cell.

In order to produce reliable comparison to see where this gene existed and to control the genetic background effect, we needed to create a single copy of the heterozygous parental line for the control. So virgin females c724 were crossed with males UAS-CG43968-RNAi/TM3, Sb; virgin females c724 were crossed with males Canton S; virgin females Canton S were crossed with males UAS-CG43968-RNAi/TM3, Sb. The F1 progeny were selected as c724-GAL4/UAS-CG43968-RNAi (normal hair), c724/Canton S, Canton S/UAS-CG43968-RNAi (normal hair). The F1 progeny were transferred to fresh vials every two days over seven days. The seven days F1 progeny flies were selected according to the marker, and equal numbers of females and males (3 males and 3 females) were flash frozen in liquid nitrogen then stored at -80 °C.

5.2.7.6 **Where is CG43968 expressed?**

RNA extraction was performed using Qiagen Micro kit (see methods section 2.5.1) for c42-Gal4 crossed samples and c724 crossed samples. The quantity was checked by Nanodrop, and quality was checked by Agilent bioanalyzer (2.5.2). Relative standard qPCR were performed not only between c42-Gal4/UAS-CG43968-RNAi line with parental lines but also between c724-Gal4/UAS-CG43968-RNAi line with parental line and between c42-Gal4 crossed fly and c724-Gal4 crossed fly in order to find the location of the novel gene CG43968 (Figure 5-20).

In the comparison between c42-GAL4/UAS-CG43968-RNAi, c42-GAL4/Canton S, Canton S/UAS-CG43968-RNAi, significantly less expression of c42-GAL4/UAS-CG43968-RNAi than parental line c42-GAL4/Canton S and Canton S/UAS-CG43968-RNAi line was shown. This result confirmed that CG43968 was knocked down in c42-Gal4/UAS-CG43968-RNAi line. The comparison between c42-GAL4/UAS-CG43968-RNAi with sibling line c42-GAL4/TM3, Sb also confirmed CG43968 was knocked down in c42-*GAL4*/UAS-CG43968-RNAi line (Figure 5-20).

In the comparison between c724-GAL4/UAS-CG43968-RNAi, c724-GAL4/Canton S, Canton S/UAS-CG43968-RNAi, the same expression level of c724-GAL4/UAS-CG43968-RNAi with parental line c724-GAL4/Canton S and Canton S/UAS-CG43968-RNAi line was shown confirming that there was no knockdown in the c724-Gal4 crossed fly. The comparison between c724-GAL4/UAS-CG43968-RNAi with sibling line c724-GAL4/TM3, Sb also confirmed the novel gene was not knocked down in c724-GAL4/UAS-CG43968-RNAi line (Figure 5-21).

Further comparisons were performed between c724-GAL4/UAS-CG43968-RNAi, c724-GAL4/Canton S, Canton S/UAS-CG43968-RNAi along with c42-GAL4/UAS-CG43968-RNAi, c42-GAL4/Canton S, and Canton S/UAS-CG43968-RNAi. The c42/TM3 Sb and c724/TM3, Sb also in the same plate showed the comparison between the knockdown line and siblings in order to determine the efficiency of the knockdown. The results are shown in Figure 5-20. The absolute percentage of knockdown for UAS-CG43968-RNAi, driven by c42-GAL4, was 80% (*t*-test, $P < 0.05$) to its heterozygous sibling c42/TM3, Sb control. The absolute percentage of knockdown for UAS-CG43968-RNAi, driven by c42-GAL4, was 70% (*t*-test, $P < 0.01$) to its heterozygous c42/Canton S parental control, indicating that CG43968 was knocked down in principal cells. The comparisons between c724 lines showed no change, indicating this gene was not knocked down in stellate cells. The absolute percentage of knockdown for UAS-CG43968-RNAi, driven by c42-GAL4, was 70% (*t*-test, $P < 0.01$) compared to c724/Canton S control indicating CG43968 is in principal cells but not in stellate cells of tubules. c42-GAL4/UAS-CG43968-RNAi compared to parental line Canton S/UAS-CG43968-RNAi showed no significant change possibly due to the RNAi leakage when the crossed happened.

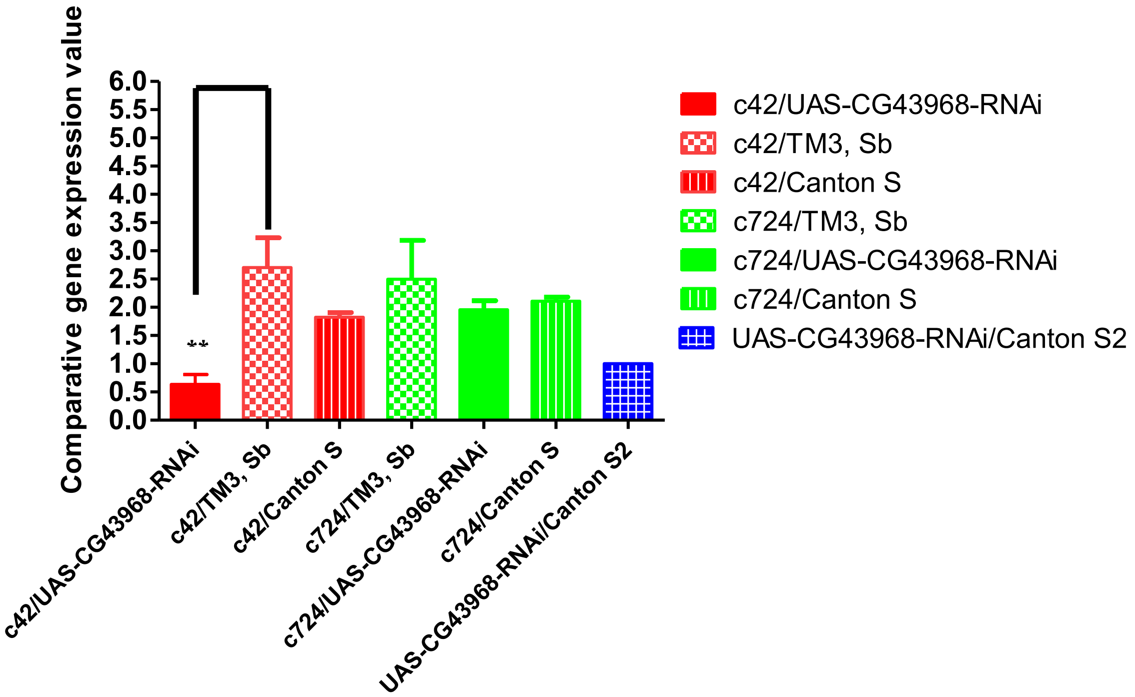
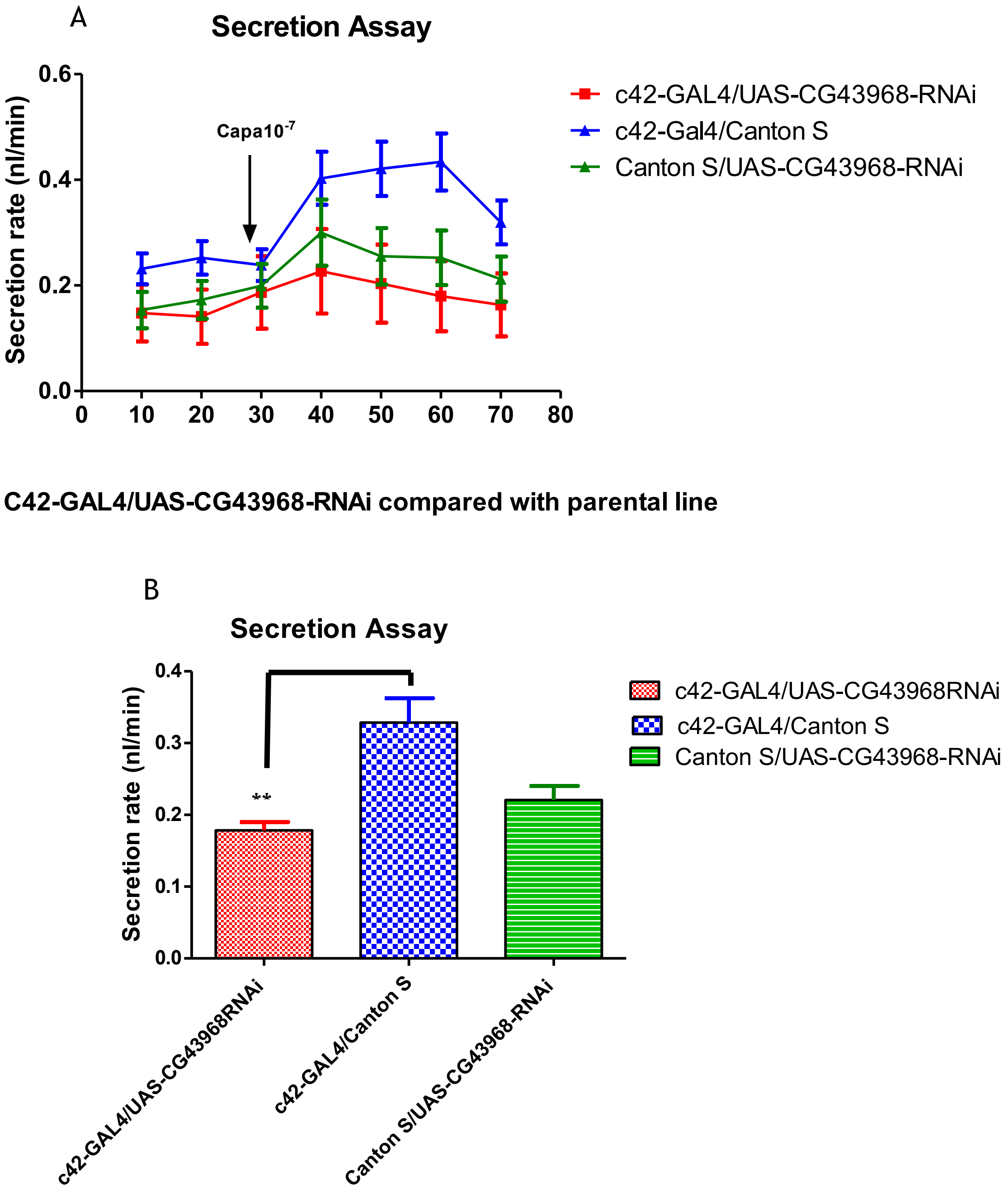


Figure 5-21 Comparative gene expression showing CG43968 is expressed mainly in principal cell.

Relative standard qPCR was performed between c42-GAL4 /UAS-CG43968-RNAi, c42/Canton S, Canton S/UAS-CG43968-RNAi, c42/TM3, Sb; between c724-GAL4/UAS-CG43968-RNAi, c724/Canton S, Canton S/UAS-CG43968-RNAi, c724/TM3, Sb; between c42 crossed lines and c724 crossed lines. The novel gene CG43968 was significantly less expressed in c42-GAL4/UAS-CG43968-RNAi line than in all the other lines as shown. This indicated that CG43968 is mainly expressed in tubules principal cells.

5.2.7.7 Secretion Assay of c42-GAL4/ UAS-CG43968-RNAi compared with parental line

Secretion assay was performed between F1 progeny c42-GAL4/UAS-CG43968-RNAi, c42/Canton S and Canton S/UAS-CG43968-RNAi from the cross scheme (section 5.2.7.4). This comparison was used to find out if CG43968 knockdown would affect the tubule secretion phenotype. If it did affect the secretion, the relation to principal cell or stellate cell would be checked by adding the neuropeptide capa or leucokinin (LK or *Drosophila* kinin). Capa increased the secretion through principal cell, whilst Leucokinins increased secretion of tubules through stellate cell. The mechanism of CG43968 function involved in secretion is summarized in Figure 5-22.



C42-GAL4/UAS-CG43968-RNAi compared with parental line

c42-GAL4/UAS-CG43968-RNAi compared with parental line

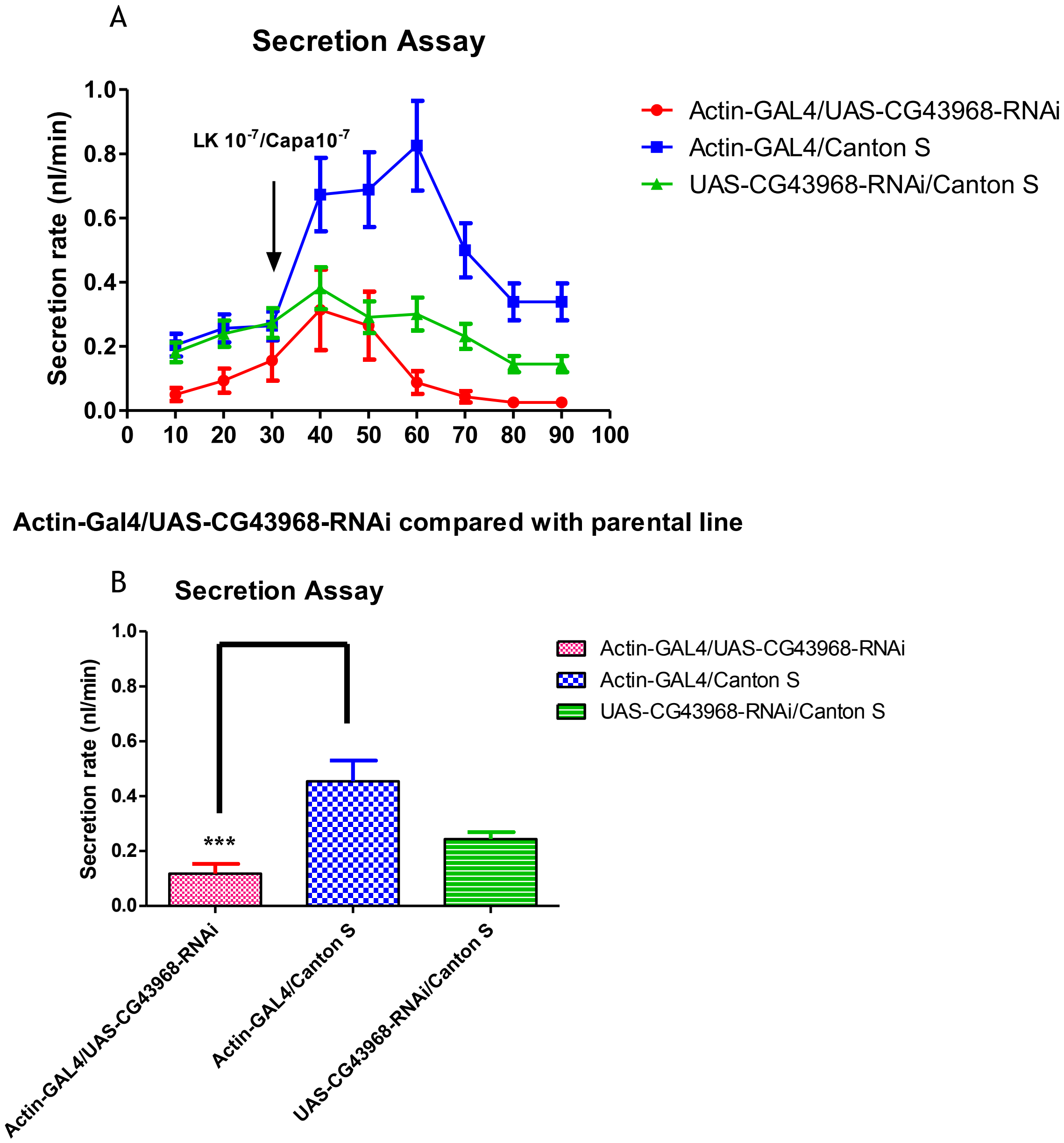
Figure 5-22 Secretion assay for c42-Gal4/UAS-CG43968-RNAi compared with parental line.

The secretion was significantly decreased in the c42-GAL4/UAS-CG43968-RNAi line compared to the parental line indicating CG43968 is involved in the tubule secretion in principal cell. (A) Tubule secretion assays were performed using modified Ramsay assay (section 2.21). The Ca^{2+} agonist capa was added after 30 min of basal readings every 10 min. An additional 40 min of secretion reading were taken every 10 min. (B) The secretion rates were averaged over the 70 mins as three lines and presented as a graph for statistical significance using a *t*-test, *P*-value. In c42-GAL4/UAS-CG43968-RNAi (3a) (red), the secretion significantly reduced with a mean difference of 0.1784 ± 0.01158 (*t*-test, $P<0.001$) (red line), it was not changed significantly in the Canton S/UAS-CG43968-RNAi line (green line). The secretion went significantly up in the c42/Canton S after adding Capa 10^{-7}M with a mean difference 0.3285 ± 0.03398 (*t*-test, $P<0.05$) (blue line).

The results confirmed that secretion was decreased in the c42-GAL4/UAS-CG43968-RNAi line compared to parental lines c42/Canton S and UAS-CG43968-RNAi/Canton S. Furthermore, the secretion of c42-GAL4 line was slightly increased when capa-1 (10^{-7} M) was added after 30 mins compared to the parental lines c42/Canton S where secretion was significantly increased and UAS-CG43968-RNAi (3a)/Canton S where secretion was moderately increased. The secretion of Canton S/UAS-CG43968-RNAi was not significantly changed due to the UAS-CG43968-RNAi leakage affecting the result. Capa-1 peptide action is diuretic via elevation of nitric oxide, cGMP and calcium in the principal cells of the Malpighian tubules. The c42-GAL4/UAS-CG43968-RNAi did not respond to Capa as well as the parental lines, indicating the Capa pathway had been blocked in some part after CG43968 knockdown. So this novel gene is in tubule principal cell and is involved in tubule secretion.

5.2.7.8 **Secretion Assay of Actin-GAL4/ UAS-CG43968-RNAi compared with parental line**

The flies were crossed using the universal driver Actin-GAL4/CyO. The crosses were created as females Actin-GAL4/CyO x males UAS-CG43968-RNAi; females Actin-GAL4/CyO x males Canton S and females Canton S x males UAS-CG43968-RNAi. The F1 progeny were chosen as Actin-GAL4/UAS-CG43968-RNAi, Actin-GAL4/Canton S, and Canton S/UAS-CG43968-RNAi identified by the marker CyO (curly wing). The secretion assay was performed between F1 progeny Actin-GAL4/UAS-CG43968-RNAi, Actin-GAL4/Canton S, and Canton S/UAS-CG43968-RNAi. Capa 10^{-7}M and *Drosophila* leucokinin 10^{-7}M were added after 30 mins (Figure 5-23).



Actin-Gal4/UAS-CG43968-RNAi compared with parental line

Actin-Gal4/UAS-RNAi (3a) compared with parental line

Figure 5-23 Secretion assay for Actin-Gal4/UAS-RNAi compared with parental line.

The secretion decreased in the Actin-GAL4/UAS-CG43968-RNAi line compared to the parental line indicating CG43968 is involved in the tubule secretion in tubule principal cells but not stellate cells. (A) Tubule secretion assays were performed using modified Ramsay assay (section 2.21). The Ca^{2+} agonist capa-1 and *Drosophila* kini were added after 30 min of basal readings every 10 min. An additional 60 min of secretion reading were taken every 10 min. (B) The secretion rates were averaged over the 90 mins as three lines and presented as a graph for statistical significance using a *t*-test, *P*-value. In Actin-GAL4/UAS-CG43968-RNAi (red), the secretion significantly reduced with a mean difference of 0.1177 ± 0.03539 (*t*-test, $P<0.001$) (red line), it did not change significantly in the Canton S/UAS-CG43968-RNAi line (green line). The secretion went significantly up in Actin-Gal4/Canton S after adding Capa 10^{-7}M /Lk 10^{-7}M with a mean difference 0.4542 ± 0.07520 (*t*-test, $P<0.05$) (blue line). However, Actin-Gal4/UAS-CG43968-RNAi (red) also increased more than c42-Gal4/UAS-CG43968-RNAi after added Capa 10^{-7}M only indicated that CG43968 was not in stellate cell.

From section 5.2.7.7, the c42-GAL4/UAS-CG43968-RNAi did not respond well to Capa-1. Secretion was assayed after adding Capa and *Drosophila* kinin to Actin-GAL4/UAS-CG43968-RNAi, Actin-GAL4/Canton S, and Canton S/UAS-CG43968-RNAi.The secretion rate for Actin-GAL4/UAS-CG43968-RNAi had a better response to Capa-1 and *Drosophila* kinin than just adding Capa 10^{-7}M itself (Figure 5-23) but still a much less response compared to the parental line Actin-GAL4/Canton S.

As we know, Capa peptide acts in the principal cells of the Malpighian tubules. In tubules, *Drosophila* kinin's major action is to raise chloride permeability through stellate cells by binding to receptors on the basolateral membrane, and so ultimately to enhance fluid secretion. The action of *Drosophila* kinin is additive to both cAMP and cGMP by elevation of intracellular calcium (Dow, 2012; Kerr et al., 2004). The slightly increased the secretion of Actin-GAL4/UAS-CG43968-RNAi line compared to the c42-GAL4/UAS-CG43968-RNAi line after adding both peptides indicated the CG43968 knockdown was not affecting stellate cell but only affecting principal cells. So the difference in secretion rate was greater between Actin-GAL4/Canton S and c42-GAL4/UAS-CG43968-RNAi. The Canton S/UAS-CG43968-RNAi showed no significant change due to the UAS-CG43968-RNAi leakage affecting the result. These results supported the qPCR results that CG43968 gene only existed in tubule principal cell.

5.2.8 *Overexpression analysis for CG43968*

5.2.8.1 **Generation of an overexpression construct by using Gateway destination vectors PTW and PTWV**

Two primers had been designed by using the entire ORF of Chr3L 23777200-2378100. Primers were designed with the sequence CACC on the 5' end of the 5' primer for using Gateway entry clone (pENTR/D-TOPO Cloning Kit) followed by the gene specific sequence in the first reading frame. One primer, ORF1, included the stop codon taa; the other primer, ORF2, excluded the stop codon taa for adding tag-YFP (Table 5-2). ORF1 and ORF2 were cloned into the pENTRY/D-TOPO (invitrogen) vector using the Gateway technology. This trigger sequence could be transferred easily to Gateway destination vector PTW and PTWV by an *in vitro* reaction mediated by LR Clonase.

PTWV contains a C-terminal tag [Venus, improved YFP (515/528 nm)]. Venus YFP shows fast and efficient maturation, allowing detection of reliable fluorescent signals that was not previously possible (Nagai et al., 2002).

Table 5-2 Primers for PTW and PTWV overexpression

Constructs	Primer sequence	Products size
PTW	**Forward:** CACCATGCGGGTGTGCGATACA **Reverse:** TTACAAACGTCTAAATATGCACTTGC	2880bp
PTWV	**Forward:** ATGCGGGTGTGCGATACA **Reverse:** CAAACGTCTAAATATGCACTTGC	2877bp

5.2.8.2 **Verification of PTWV in *Drosophila* S2 cells**

In order to verify the PTWV construct worked efficiently before sending the construct to generate transgenic flies, a PTWV plasmid was transfected into the S2 cell by using Insect GeneJuice Transfection method (section 2.16.2). The overexpression protein in S2 cell cytoplasm by confocal microscope can be determined (section 2.20.2) (Figure 5-24)

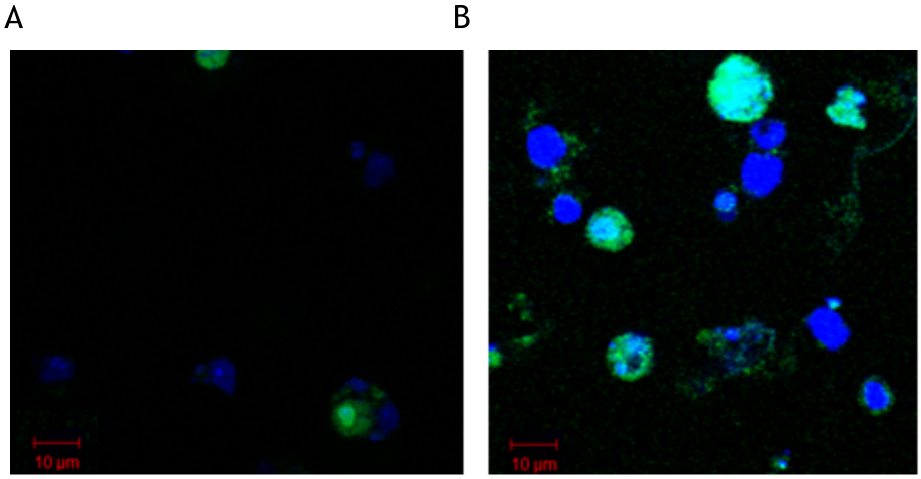


Figure 5-24 UAS-CG43968-PTWV25 plasmid overexpression in *Drosophila* S2 cell.

UAS-CG43968-PTWV25 plasmid (containing Venus) successfully showed the fluorescent signals. The UAS-CG43968-PTW25 plasmid was transfected into *Drosophila* S2 cell using Insect GeneJuice Transfection method, and used to show the YFP signals in S2 cell membrane or cytoplasm or both. (A) The control cells were not transfected by PTWV. No fluorescent signal was detected. (B) Fluorescent signal (Venus) was detected in cytoplasm. The blue fluorescent signals were DAPI which were stained nuclei, the green fluorescnt signals were YFP of the novel gene.

The fluorescent signal (Venus, YFP) was detected in cytoplasm of *Drosophila* S2 cell indicating that the novel gene CG43968 was successfully transferred into PTWV vector by using Gateway system and transfected into S2 cell. So the UAS-

CG43968-PTWV25 overexpression construct was confident to send to generate the transgenic fluorescent flies. PTW was sent to generate transgenic flies without tag in order to test the gene function without the effect of the fluorescent tag.

5.2.8.3 Making and validating the overexpression transgenic flies lines

PTW and PTWV plasmid with ORF1 and ORF2 was purified and 50 μg plasmid with concentration at least 0.5 μg/μl was sent to BestGene Inc (thebestgene.com) for microinjection to make transgene flies (Figure 5-1).

In order to validate the overexpression constructs, UAS-24/CyO, UAS-25/Cyo lines were chosen to cross with Actin-GAL4/CyO universal driver in adult whole flies.

The cross scheme (Figure 5-25):

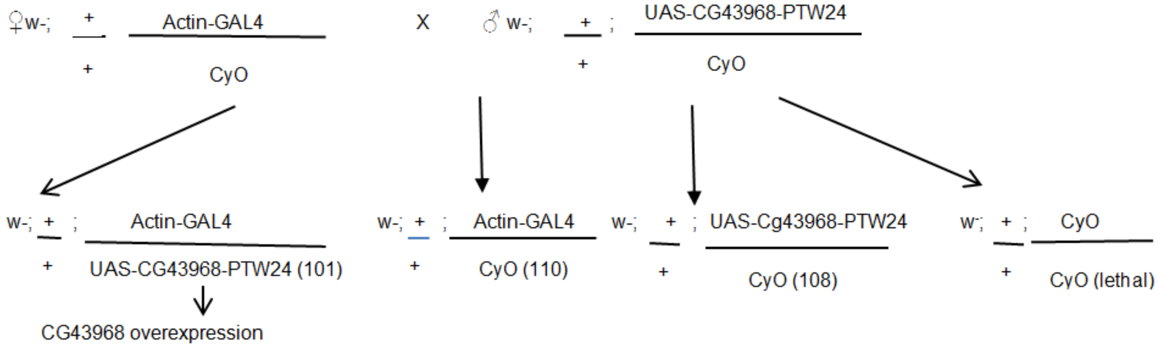


Figure 5-25 A schematic diagram of the RNAi overexpression cross scheme.

Schematic diagram of using ubiquitous driver (Actin-GAL4/CyO) crossed with overexpression fly line UAS-CG43968-PTW24 and UAS-CG43968-PTWV25 to obtain the CG43968 overexpression in adult whole fly lines (Actin-GAL4/UAS-CG43968-PTW24 and Actin-GAL4/UAS-CG43968-PTWV25) through the GAL4-UAS system. The survival numbers for each genotype are shown in brackets. The survival ratio for each genotype was the same 1:1:1 indicating overexpression gene CG43968 was not lethal.

The F1 progeny were also counted (with the numbers shown in the bracket) for the four genotypes to check the survival rate in order to confirm the novel CG43968 was not lethal when overexpression was happened.

The qPCR validation of novel gene overexpression of F1 progeny from Actin-GAL4/CyO x UAS-CG43968-PTW24/CyO was performed relative to Actin-GAL4/UAS-CG43968-PTW24 vs. Actin-Gal4/CyO; Actin-GAL4/CyO x UAS-CG43968-

PTWV25 was validated relative to Actin-GAL4/UAS-CG43968-PTWV25 vs. Actin-Gal4/CyO (Figure 5-25). The result showed no change for Actin-GAL4/UAS-CG43968-PTW24 vs. Actin-Gal4/CyO (First two red bars); Actin-GAL4/UAS-CG43968-PTWV25 vs. Actin-Gal4/CyO showed the expression increased by 58% (Figure 5-26) (the last two green bars). The UAS-CG43968-PTWV25 was chosen for further analysis of function.

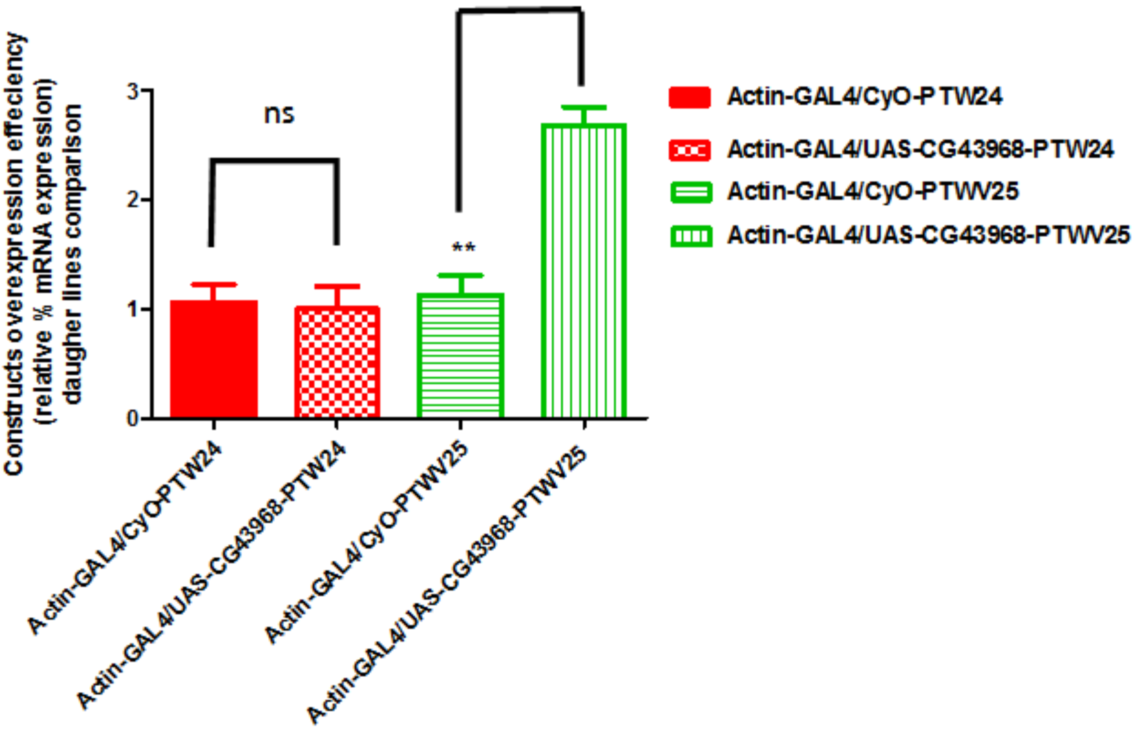


Figure 5-26 Screen of CG43968 overexpression constructs.

The Actin-GAL4/UAS-CG43968-PTWV25 fly line achieved overexpression of 58% over control. The comparison of the overexpression efficiency of two fly lines was performed for UAS-CG43968-PTW24, UAS-CG43968-PTWV25 constructs by using the ubiquitous driver GAL4. The percentage of overexpression was 58% (Actin-GAL4/UAS-CG43968-PTW25) compared to their siblings, Actin-GAL4/UAS-CG43968-PTWV25 achieved the more efficient overexpression. Actin-GAL4/UAS-CG43968-PTW24 showed the expression had no change compared to their siblings.

5.2.8.4 **Immunocytochemistry (ICC) for overexpression line**

Overexpression over control for the F1 progeny of Actin-GAL4/UAS-CG43968-PTW25 from Actin-GAL4/CyO xUAS-CG43968-PTW25/CyO was confirmed by qPCR (Figure 5-25). UAS-CG43968-PTWV25 was analysed in fusion with YFP. The Actin-*GAL4*/UAS-25 line was chosen to run ICC in order to determine the location of the novel gene CG43968. Actin-GAL4/UAS-CG43968-PTWV25 flies were checked using a fluorescent microscope with YFP channel. The flies were lit up in two third of the body, giving confidence that this line had the fluorescent signals. The control line is UAS-CG43968-PTWV25 flies or Canton S flies.

Firstly, tubules were dissected from Actin-GAL4/UAS-CG43968-PTWV25 line and UAS-CG43968-PTWV25 line, and viewed by fluorescent microscope. The YFP signal was viewed by choosing the green light channel (Figure 5-27). The fluorescent signals of Actin-GAL4/UAS-CG43968-PTWV25 line were mainly seen in the tubule ureter, the main segments of tubules and the basolateral of tubules (principal cells?) indicating CG43968 may be located in these places.

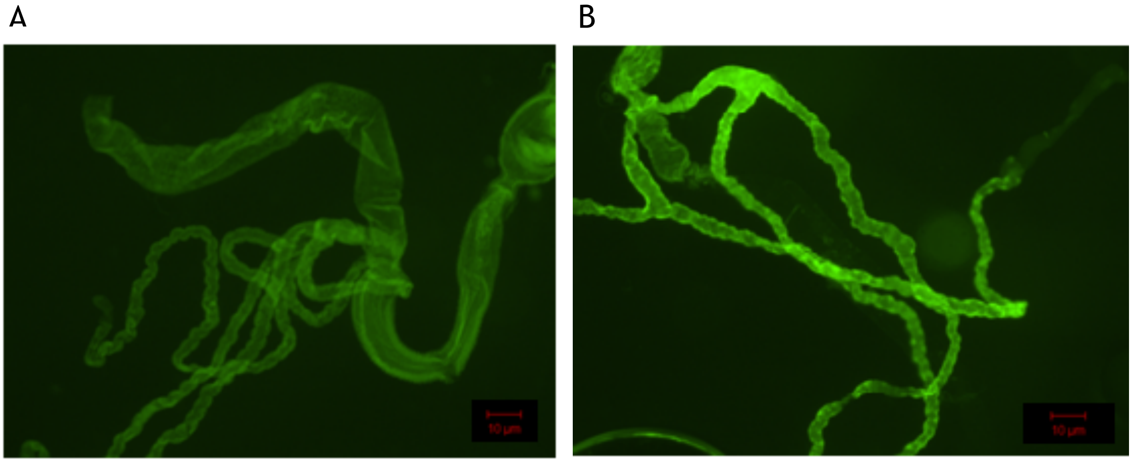


Figure 5-27 Overexpression of Actin-GAL4/ UAS-CG43968-PTWV25 line.

The tubules from Actin-GAL4/ UAS-CG43968-PTWV25 line that were viewed by fluorescent microscope showed the signals being detected in tubules ureter, main segments and basolateral principal cells. (A) Tubules from UAS-CG43968-PTWV25 line as control. The leaky YFP gave rise to a weak fluorescent background but no real fluorescent signals. (B) Tubules from Actin-GAL4/ UAS-CG43968-PTWV25 overexpression line. The ureter, the main segments of tubules and the basolateral of tubules were lit up by Venus (principal cells?) which showed the green fluorescent signals.

Second, live GFP imaging does not always produce clear images, so the tubules were also fixed and stained with anti-GFP antibody. Tubules from Actin-GAL4/ UAS-CG43968-PTWV25 line and control Canton S fly line were dissected and stained by antibodies, using the primary antibody Anti-GFP antibody (1:000) and the secondary antibody Fluorescent Goat anti-mouse- IgG-FITC (1:500) (section 2.18). The images were viewed by confocal microscope (section 2.20.2) as depicted in Figure 5-28.

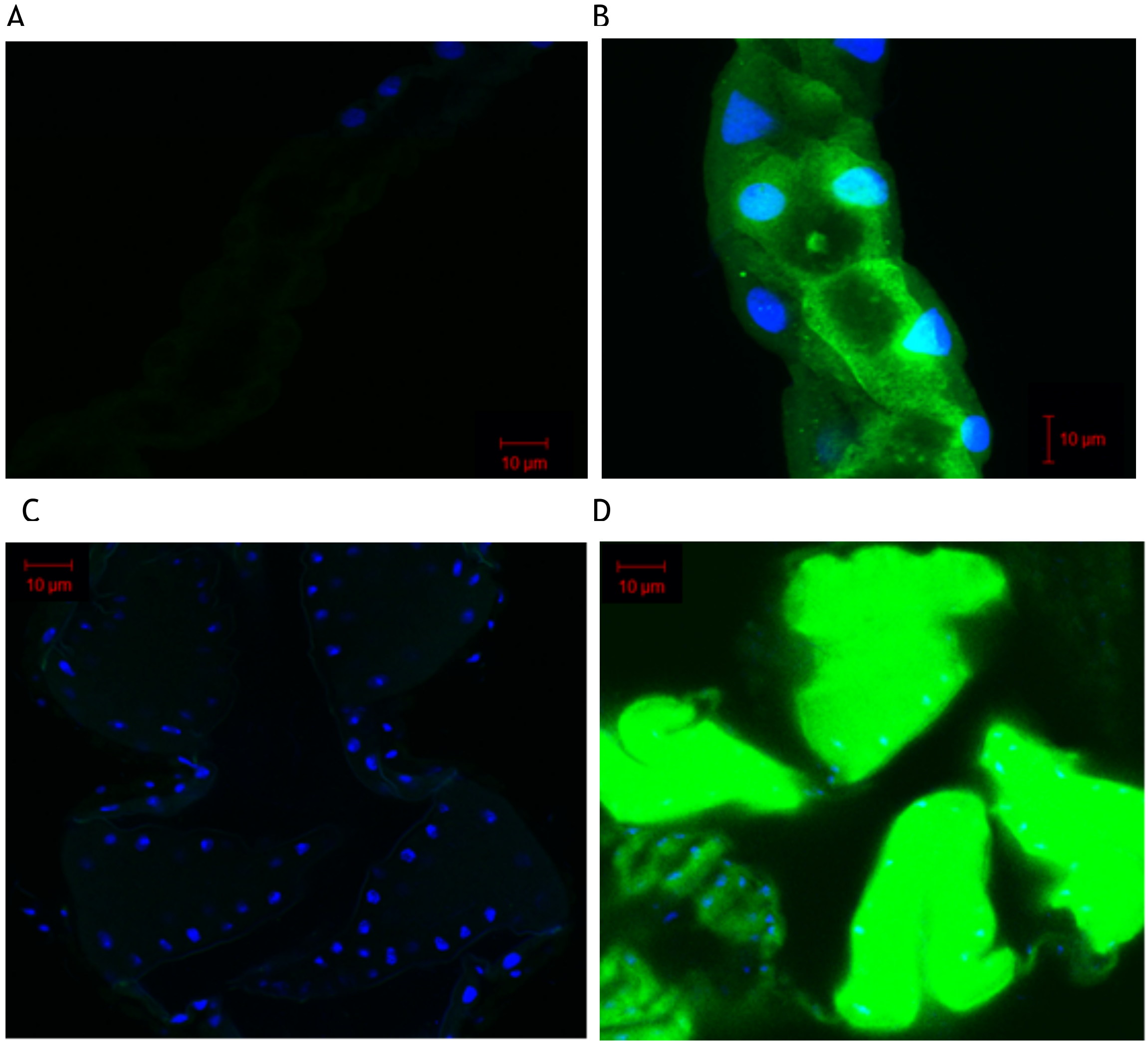


Figure 5-28 ICC of tubules, hindguts of Actin-GAL4/UAS-CG43968-PTWV25 fly line.

ICC of tubules, hindguts of Actin-GAL4/UAS-CG43968-PTWV25 line showed the YFP signals might locate in membrane or cytoplasm or both. (A) ICC of Canton S tubules, viewed by confocal microscope, showing no signals were detected. (B) ICC of Actin-GAL4/ UAS-CG43968-PTWV25 tubules, viewed by confocal microscope showing YFP signals (green fluorescent) were seen in principal apical membrane, probably in microvilli and cytoplasm. The blue fluorescent signals were from DAPI which were stained nuclei. (C) ICC of Canton S rectum, viewed by confocal microscope, showing no YFP signals were detected. The blue fluorescent signals were from DAPI which were stained nuclei. (D) ICC of Actin-GAL4/ UAS-CG43968-PTWV25 rectum, showing YFP signals (green fluorescent) were detected in cytoplasm and membrane.

The ICC signals from the confocal microscope suggested that in tubule, the fusion protein from tubule-enriched gene CG43968 was located either in cell apical membrane probably concentrated in the microvilli, or possibly in cytoplasm as well.

5.2.8.5 **Western blotting to search the protein of novel gene CG43968**

Actin-GAL4/CyO females were crossed with UAS-CG43968-PTWV25/CyO male flies to overexpress the CG43968 in whole flies. Six flies of F1 progeny Actin-

GAL4/UAS-CG43968-PTWV25 were selected to be frozen in protein lysis buffer RIPA. Six Canton S flies were frozen in protein lysis buffer. Protein were extracted and quantified by using Bradford assay (section 2.17). 20 µg protein was used to run the gel. The primary antibody was anti-GFP antibody (1:1000), the secondary antibody was goat anti mouse IgG-HRP (1:5000).

The result is shown in Figure 5-29. The size of protein for CG43968 was 118.76 KDA. The YFP size is 28kDa. We expected to see the detected size of the protein attached to YFP at size 146.76, but we can only see one band detected at size 28 kDa. This indicated that the YFP was not attached to the protein of the novel gene, but smear bands from 28 kDa to 148 kDa were detected.

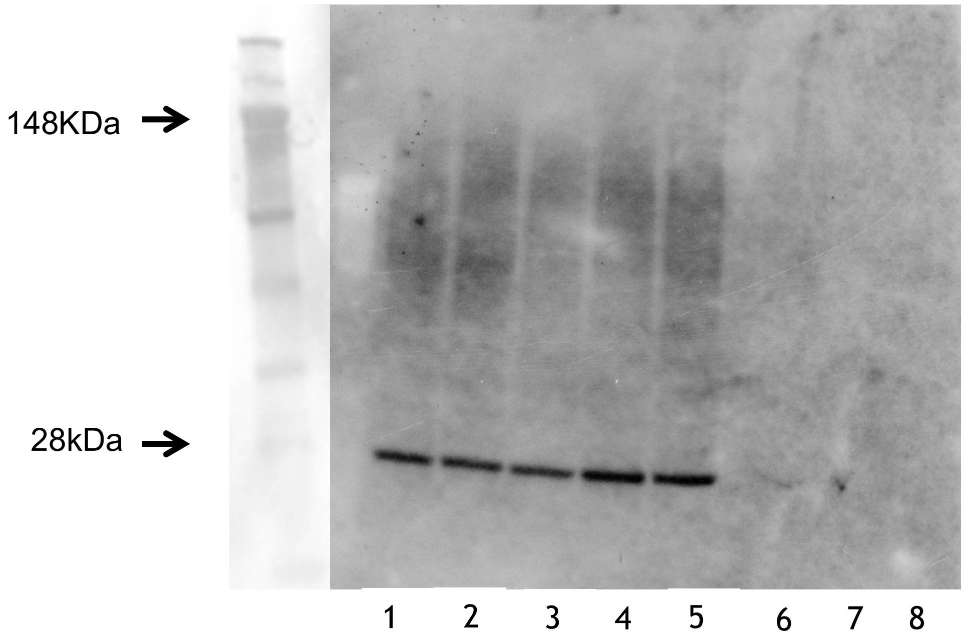


Figure 5-29 Western blotting of CG43968 overexpression line.

Western blotting showed the signals from YFP but not from the novel protein CG43968. Western blotting of five replicate samples of Actin-GAL4/UAS-CG43968-PTWV25, lane1-5 showed the same size bands 28 kDa, which was the same size as the YFP only. Three triplicates of Canton S control flies proteins were shown from lane 6-8. No signals were detected.

Different methods to treat the proteins were tested including denaturing the protein by heating at 95°C instead of 100°C to protect the protein-YFP bond: the results were the same. Pre-denaturing the protein by adding 0.05M EDTA was also tested: the results were the same. Given this, there are multiple reasons why the protein-YFP products cannot be detected. Firstly, the novel protein is soluble; GFP is normally distributed throughout the cytoplasm. So perhaps the protein is partially degraded in the cell, leading to a smear band at the expected size and a prominent degradation product at 28kDa. Secondly, the protein-YFP bond was broken during the sample preparation or during the west blotting process. Thirdly, the construct failed, and the UAS-CG43968-PTWV25 line only contained YFP but not attached on the novel protein from novel gene CG43968.

If this protein was a soluble protein, it might have more chance to exist in cytoplasm rather than membrane. Membrane protein has helix interactions so they are more stable and tight. However soluble proteins are alpha-bundle proteins, which are easily dissolved in water (Eilers et al., 2002).

CG43968 is only confirmed as a coding gene if the coded protein is detected. So I plan to design a CG43968 specific antibody in order to prove the novel protein will be necessary in the future.

5.3 **Discussion**

In this chapter, a tubule-enriched novel coding gene, named recently as CG43968 by FlyBase recently, was found by the RNA-seq poly(A) selection method. The reverse genetic approach was used to elucidate the function of CG43968.

5.3.1 ***CG43968 is a real protein coding gene***

The following evidence supports CG43968 as a tubule-enriched coding gene.

Firstly, the splicing junction was found by tubule RNA-seq poly(A) selection method in our project. The splicing junction was also confirmed by modENCODE RNA-seq trace in FlyBase. The novel gene was named as CG43968 by FlyBase on 16/08/2012, FlyBase 5.48. The novel gene CG43968 was also supported by strand-specific RNA-seq data of tubule in this project (Figure 5-7).

Secondly, the splicing event was confirmed by RT-PCR (Figure 5-17) with the cDNA and genomic DNA having two different sizes, and also confirmed by sequencing the CG43968 construct of pENTR/D-TOPO® vector (Figure 5-18).

Thirdly, open reading frame (ORF) search found a long ORF coded by this gene (Figure 5-12), and the blastx search found the novel protein conserved in *Drosophila mojavensis* and *Drosophila willistoni*, also conserved in mosquito and beetle (Figure 5-13).The coding genes are more conserved in other species than noncoding genes (Eddy, 2001).

Fourthly, Coding Potential Calculator (CPC calculator) program had predicted that this gene was coding gene (Kong et al., 2007; Roberts et al., 2011).

Lastly, a CG43968 transfragment was also found by *Drosophila* Tiling 2.0R Array of tubules in this project (Figure 5-10). Tiling microarrays are another technology for novel gene discovery with more limitations when compared to RNA-seq (Manak et al., 2006; Wang et al., 2009). However, it can be used to confirm the RNA-seq results.

5.3.2 *The possible location of CG43968*

Different techniques have been employed to search the *in situ* localization of *CG43968.*

Firstly, the expression of *CG43968* is found in tubule main segment. *In situ* hybridization confirmed the Dig detection signals of anti-sense probes were in the tubule main segment (Figure 5-16).

Secondly, the protein prediction programs, both PSORT II (Protein Subcellular Location Prediction) and WoLF PSORT predicted CG43968 to be more cytoplasmic but still may locate in plasma membrane.

Thirdly, CG43968 presents in principal cells. Taking advantage of the *Drosophila* Gal4/UAS system, transgenic RNAi flies were generated by Invitrogen Gateway system using pRISE RNAi vector (Kondo et al., 2006) (Materials and Methods Chapter 2.15). c42-Gal4 (principal cell specific driver) crossed with UAS-CG43968-RNAi produced a gene knockdown specific in principal cells was confirmed by qPCR, however c724-Gal4 (stellate cell specific driver) crossed with UAS-CG43968-RNAi failed to ablate the expression also confirming CG43968 is only expressed in principal cells but not in stellate cells.

Tubule secretion assay using cell-specific RNAi knockdown line c42-Gal4/UAS-CG43968-RNAi compared with parental line c42-Gal4/Canton S confirmed the knockdown line having less secreted fluid than the parental line. After stimulating the tubules with cell-specific diuretic neuropeptides (Capa-1 acts on principal cells and *Drosokinin* acts on stellate cells), c42-Gal4/UAS-CG43968-RNAi tubules showed no response to CAPA in contrast to *drosokinin* stimulation compared to parental line c42-Gal4/Canton S. This result further supported that CG43968 is in principal cells. Lastly, CG43968 showed apical plasmamembrane localisation, concentrated in microvilli of the principal cells. Transgenic overexpression flies were generated by Invitrogen Gateway system using PTWV vector with the fuses CG43968 with fluorescent YFP to help identify subcellular localization of the protein (Figure 5-24). Actin-Gal4 a ubiquitous *GAL4 driver* crossed with UAS-CG43968-PTWV s suggested CG43968 localization in principal cell plasma apical membrane microvilli (Figure 5-28), but also possibly located in cytoplasm.

5.3.3 *Possible function of CG43968*

Secretion assay performed between c42-Gal4/UAS-CG43968-RNAi and parental line showed less secretion in c42-Gal4/UAS-CG43968-RNAi line compared with the parental line suggested CG43968 may play an essential role in apical plasma membrane secretion upon stimulation.

Secondly, CG43968 has a Latrophilin/CL-1-like GPS domain that may modulate intracellular calcium signaling events. The same domain is implicated in renal tubular development in humans.

The blastx search found out CG43968 is conserved in other *Drosophila* species and there are good similarities to other species including mosquito and beetle proteins. Further reciprocal BLAST found CG43968 has a Latrophilin/CL-1-like GPS domain.

The GPS domain presents in Latrophilin/CL-1, sea urchin REJ and polycystin. Polycystin-1 is a protein in humans is encoded by the *PKD1* gene related to human kidney disease. Polycystin-1 is a glycoprotein which contains a large N-terminal extracellular region, multiple transmembrane domains and a cytoplasmic C-tail. The CG43968 protein only back half is conserved may be related to the structure of the protein. PKD1, CL-1 and REJ possess functional similarities that are likely to be due to their common GPS domains and transmembrane regions. Each of these molecules is suggested to mediate transmembrane influx of Ca^{2+} (Ponting et al., 1999). Polycystin-1 also may function as an integral membrane protein involved in cell-cell/matrix interactions, and may modulate intracellular calcium homoeostasis and other signal-transduction pathways. It plays a role in renal tubular development, and mutations in this gene have been associated with autosomal dominant polycystic kidney disease (Van Adelsberg and Frank, 1995).

In *Drosophila*, Calcium (Ca^{2+}) is a ubiquitous second messenger molecule in all cell types and tissues. Calcium signalling and calcium homeostasis are essential for life. *Drosophila* tubule function requires extracellular calcium influx into principal cells *via* plasma membrane channels (Davies and Terhzaz, 2009). Calcium entry is essential for CAP2b to induce a physiological response in the whole organ (Rosay et al., 1997). CAP2b, a cardioactive neuropeptide that stimulates fluid secretion by a mechanism involving nitric oxide, causes a rapid, dose dependent rise in cytosolic calcium in 77 principal cells in the main (secretory) segment of the tubule. So calcium signalling plays important role in the modulation of the nitric oxide signalling pathway in tubules.

The impact of extracellular calcium on $[Ca^{2+}]_{cyt}$ and fluid transport by the Malpighian tubule is dramatic. In the absence of external calcium, the CAP2b-induced calcium response is abolished (Davies and Terhzaz, 2009).

Rise in tubule cAMP, cGMP or calcium in the principal cells stimulates fluid secretion by tubule. However raising calcium in stellate cells also stimulated fluid secretion, so did both cAMP and cGMP (Dow, 2007).

If CG43968 localised in plasma membrane, it may be a membrane protein of tubule which contains a GPS domain, and is responsible for taking extracellular calcium into the principal cell. So CAP2b will be induced and cytosolic calcium being raised, nitric oxide production will be activated and cGMP will be stimulated to increase tubule secretion. If CG43968 is in cytoplasm, it may be involved in stimulating the nitric oxide pathway to accelerate the secretion of tubule. That is the reason when CG43968 is knocked-down, the secretion of tubules less than control (Figure 5-22).

5.3.4 *Future work*

More assays needed to be done to further confirm the function of CG43968.

Firstly, calcium assays. There are many methods for calcium assay, including protein-based recombinant Aequorin probes to assay the intracellular calcium level after CG43968 knockdown to compare with control. This will confirm if GPS domain of CG43968 impacts on calcium signalling or transport events.

Calmodulin-based fluorescent reporters as calcium probes have also been of immense value in calcium signalling/transport research (Meldolesi, 2004).

In *Drosophila* protein-based aequorin calcium probes allow expression in specific cells and tissues *in vivo via* targeted expression of recombinant aequorin. The GAL4-UAS system has been developed in order to investigate calcium signalling using aequorin in *Drosophila*, UAS–apoaequorin flies were generated (Rosay et al., 1997). These flies, the first transgenic animals for a calcium reporter, were used to monitor calcium signals in live intact tissue and provided the first *in vivo* measurements of cytosolic calcium concentration ([Ca2+]cyt) in *Drosophila* Malpighian tubule (Rosay et al., 1997) and brain. This is more popular method to measure the calcium in the lab.

Secondly, design antibody for CG43968 protein to further confirm the coding ability and the location of the protein are important for the function search.

Thirdly, *In situ* hybridization of the CC43968 knockdown line may further confirm the gene location.

5.4 **Conclusions**

RNA-seq is most advantageous technology for novel gene discovery. Reverse genetics is an important technique for discovering the novel gene function. *Drosophila* is an ideal model organism for reverse genetics. So RNA-seq combines with reverse genetics and *Drosophila* model will be a powerful method to discover the novel genes and elucidate their functions. Integrative physiology and functional genomic will supply all the possible techniques to discover the phenotypes and close the ‘phenotype-gap’ in *Drosophila* tubules and other organisms.

6. Species Array

Summary

Microarray technology is the first technology to measure gene expression at the whole genome level. This technology has been dominant in the transcriptomic field for more than a decade, represented by the success of the Affymetrix GeneChip. However the GeneChips are only available for model species. Using genomic DNA as a mask will increase the sensitivity of measuring gene expression when apply heterologous microarrays for non-model species (Hammond et al., 2006; Hammond et al., 2005).

Next-generation sequencing is a recently developed technology for transcriptome analysis at the whole genome level. This technology doesn’t require *a priori* knowledge of the genome sequence, and can be used to sequence the genomes whilst measuring gene expression. As a result, this technology can be applied to all species, and so it may render the microarray genomic mask method as obsolete in the future. However, the analysis of next generation sequence data is a challenge in that it requires computing resources and requires specialist bioinformatics knowledge. In contrast, applying the microarray genomic masks method is simple, quick and so may remain a useful approach in the future.

This chapter will examine the method for applying a genomic mask to measure the *Drosophila pseudoobscura* and *Drosophila simulans* transcriptome on the *Drosophila melanogaster* Genechip.

6.1 Introduction

6.1.1 *Cross-species and cross-strain hybridization*

Two methods are applied for using microarray technology for species that have no representative microarray platforms. One is cross-species hybridization (CSH), in which the RNA from one species (the target) is hybridized to a microarray representing another (reference) species. The other method is cross-strain (within species) hybridization (CTH), which is used to study variation between strains of the same species. (Bar-Or et al., 2007).

The advent of whole-genome transcriptome profiling using high-density microarrays has had a substantial impact on the understanding of biological systems, including development processes and disease responses and how these are regulated at the transcriptome level (Hammond et al., 2005). Several microarray platforms are used for these studies. The Affymetrix GeneChip is one of the most successful platforms.

After ten years of development, GeneChips are available for a large numbers of species. However the GeneChips are still limited to model species. For example, *Drosophila* has only one type of Genechip available for *Drosophila melanogaster* but there are more than 1700 species of *Drosophila* known in the world. Only twelve of the *Drosophila* species have been sequenced. Little attention has been paid to the transcriptome analysis of these species.

Cross-species approaches have been used widely to study the transcriptome of species that don't have complete genome arrays available (Bar-Or et al., 2007; Davey et al., 2009; Hammond et al., 2006; Hammond et al., 2005; Nuzhdin SV, 2004). However, microarrays are designed for species-specific hybridization or cross-strain hybridization. Cross-species use of the method is controversial and considered a non-standard application for microarrays, but mostly the results are still meaningful. To increase the ability to reflect biological processes, care should be taken when choosing microarray platforms, including issues encompassing experimental design to data analysis.

6.1.2 *Data analysis of cross-species hybridization*

In terms of data analysis, the data need to be filtered to obtain valid biological results. Two approaches have been used to filter the data. One approach is to filter sets of microarray probes using available genomic data. Matching is determined based on the sequence similarity between the probes; filtration excludes the probes that have a low level of similarity to the target species (Bar-Or et al., 2006). However, this method can only be applied to species whose genome has been sequenced. The second approach uses genomic DNA (gDNA) hybridizations as a mask to exclude the weak hybridization probes and enable the selection of probes with higher level of matching to the transcripts. This method can also be applied to species whose sequence is not available (Hammond et al., 2005).

Graham *et al.* put equine brain and liver samples on the human U133 plus 2 chips. They used two methods to filter the data to generate the probe masks. One method involves using BLAST search to compare the equine genome and human genome to select the perfect match; the other method is gDNA selection to mask off the unmatched probes. Comparing these two methods found that fewer probe-pairs and probe-sets are retained using the BLAST search compared to the gDNA probe selection method (Graham et al., 2010). Differences due to non-perfect 25bp matches between the probes and the equine gDNA were still hybridized and thereby still retained in the probe-mask files. BLAST search and gDNA selection combined method may be more accurate in the analysis.

6.1.3 *Platforms for cross-species hybridization*

What is the best type of probes for CSH: Affymetrix short oligomer (~25 mer) probe-sets, longer oligomer (~30-60-mer) probes, or even cDNA microarray probes? Research by Walker *et al* suggested that the longer the probes, the better the CSH performance (Walker et al., 2006). Although long-probe microarray seems to be the preferred platform for CSH, data filtration can increase the validity of CSH results obtained from either long- or short-probe microarrays. The design of Affymetrix GeneChips makes them ideal for cross-species hybridization (Bar-Or et al., 2007) because Affymetrix GeneChips are designed to use 11-20 probes pairs to represent one transcript. Each probe-pair consists of a perfect-match (PM) and mismatch (MM) probe. The PM probe is a 25-base sequence complementary to the target gene, whilst the MM probe is identical to the PM probe but with a single mismatch at thirteenth base. Using a genomic mask can exclude probes with weak hybridization signals but retain the high level signal probes, so that the one or two probes retained after selection can still represent the expression of the genes.

6.1.4 *Examples for cross-species hybridization*

The cross-species method using genomic DNA as a mask has been successfully used in the analysis of a number of different non-model species for which no commercial GeneChip is available. Hammond *et al.* examined *Brassica oleracea* under phosphate stress and a control group on the *Arabidopsis thaliana* chip, successfully detecting ninety-nine genes that were significantly regulated in the shoot under phosphate stress (Hammond et al., 2005). Graham *et al.* applied sheep tissues on Human U133 plus 2 arrays. The results of the RNA analysis comparing skeletal muscle and liver transcriptomes demonstrated that the gDNA probe selection method is suitable for studying gene expression profiles in sheep tissue and produces biologically relevant data (Graham et al., 2011).

6.1.5 *Cross-species hybridization in Drosophila*

There are more than 1700 species of *Drosophila* in the world. However, only one type of GeneChip for *Drosophila melanogaster* (*Drosophila* Genome 2.0 Array) is available at the moment.

Nuzhdin *et al.* had compared 10 heterozygous *D. simulans* genotype and a pool sample of 10 *D.melanogaster* lines using Affymetrix GeneChip (*Drosophila* Genome 1) provided a genome approach for cross species hybridization to identify candidate genes potentially responsible for adaption and specification in *D.simulans* and *D.melanogaster* on the basis of rapid divergence in expression. This was demonstrated that a large fraction of the genome may be involved in adaptation via expression (Nuzhdin SV, 2004). Their research revealed the common pattern of evolution of gene expression level and protein sequence in *Drosophila*. Ranz *et al* also used within-species microarrays to demonstrate the sex-dependent gene expression and evolution of the *Drosophila* transcriptome (Ranz et al., 2003).

6.1.6 *Advantages and disadvantages of cross-species hybridization*

Cross-species microarrays are used in comparative, evolutionary and ecological studies of closely related species. This method is particularly useful for those species for which the genome has not been sequenced, or no specific microarray is available, or the annotation is poor or at the early stage. Designing a new array for those species is either not possible or would require spend a lot of money or effort. Research in this approach demonstrates that reasonable and biologically meaningful results can be obtained after data filtration either by BLAST or by gDNA hybridization. This method makes possible research using the already available GeneChips. However, this method is not a standard microarray approach; the results are to be interpreted with caution. Choosing suitable microarray platforms, careful experimental design, and more effort in data filtration are necessary to obtain good data. The performance of CSH depends on the degree of probe-transcript sequence-similarity matching. If during the hybridization, there were low match probes, probes not matching to any target species or more than one matching to target species, the data generated will bias the biological results (Bar-Or et al., 2007).

6.1.7 *Aim of the experiment*

The aim of this experiment was to investigate the feasibility of using the genomic mask method in conjunction with *Drosophila* microarrays. The experiment was designed to use cross-species hybridization to analyse tubule samples of *Drosophila* similans and *Drosophila* pseudoodscura using standard *Drosophila* melanogaster expression microarrays. A genomic DNA mask (one sample of *D.*similans whole fly, one sample of *D.* pseudoobscura whole fly) was applied to the standard *Drosophila* melanogaster microarrays to eliminate the different sequence probes and increase the sensitivity of the useful probes for expression measurement.

Gene expression was compared for *D.* simulans (four replicate tubule samples on *D.* melanogaster microarrays, *D.*pseudoobscura (four replicate tubule samples on *D.* melanogaster microarray) and *D.* melanogaster (four replicate tubule samples of Oregon R) before and after genomic masks were applied. The expression results were used to investigate the sensitivity of the cross-species hybridization and the function of the genomic masks.

6.2 **Results**

6.2.1 ***Quality control analysis***

Figure 6-1 presents the results obtained by applying the replicates of the three tissues OR (*D.melanogater*, Oregon-R), PS (*D. psedoobscura*) and SM (*D.simulans*) to *D. melanogaster* GeneChip. Histograms present a 2-D view of the distribution of data, where the values of each variable are split into equal-size bins and the number of counts in each bin is represented by the height of the bar. The histogram showed that OR and SM grouped together but PS showed the different distribution. However, the replicate samples within the species of OR and SM were grouped together but PS showed the variation within the replicate samples. The results indicated that the variation of the hybridization within tissues is less than the variation between tissues. Second, the OR and SM are very close but PS are different from OR and SM. It indicated the main variation was caused by between tissues hybridization.

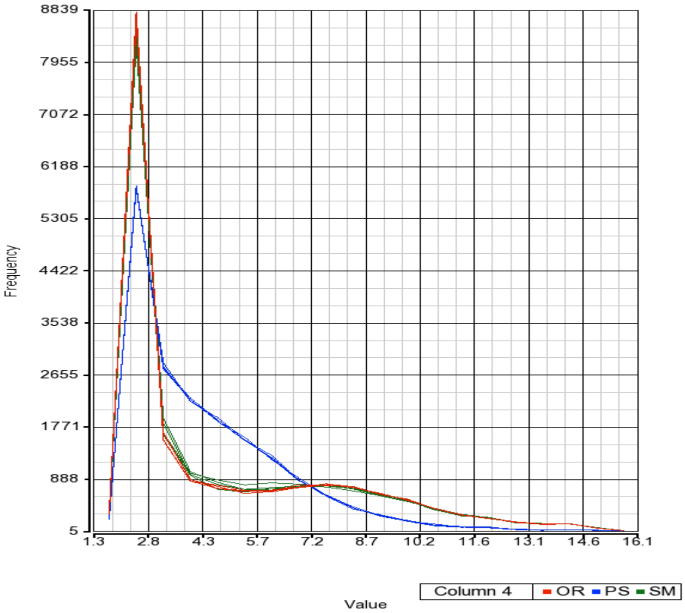


Figure 6-1 Histogram view of three types of tissue samples.

This is the histogram view of three types of tissue samples of OR (*D. melanogaster* Oregon-R) red, PS (*D. psedoobscura*) blue, SM (*D.simulans*) green to show the distribution of the three types of tissue samples. One line for each of the samples with the $\log_2$ intensity of the probes was graphed on the X-axis and the frequency of the probe intensity on the Y-axis. This allows viewing the distribution of the intensities to identify any outliers. OR and SM showed the similar distribution, However, PS showed the different distribution with OR and SM indicated that PS expression may affect by the *D. melanogaster* chip due the difference of sequences.

Principal Component Analysis (PCA) is a statistical procedure concerned with elucidating the covariance structure of a set of variables. In particular it allows us to identify the principal directions in which the data varies. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible.

PCA in Partek Genomics Suite (Patek Inc v6.6) was used to assess the behaviour of all of the genes in each individual tissue and then to build the relationships among the tissues (Figure 6-2). PCA analysis included firstly calculating PC scores by computing the standard correlation between each gene's expression profile vector and each principal component vector (eigenvector) and secondly, calculating the standard correlation between each condition vector and each eigenvector. A correlation matrix in the form of PCA scores was then presented to indicate the relation between two conditions (tissues).

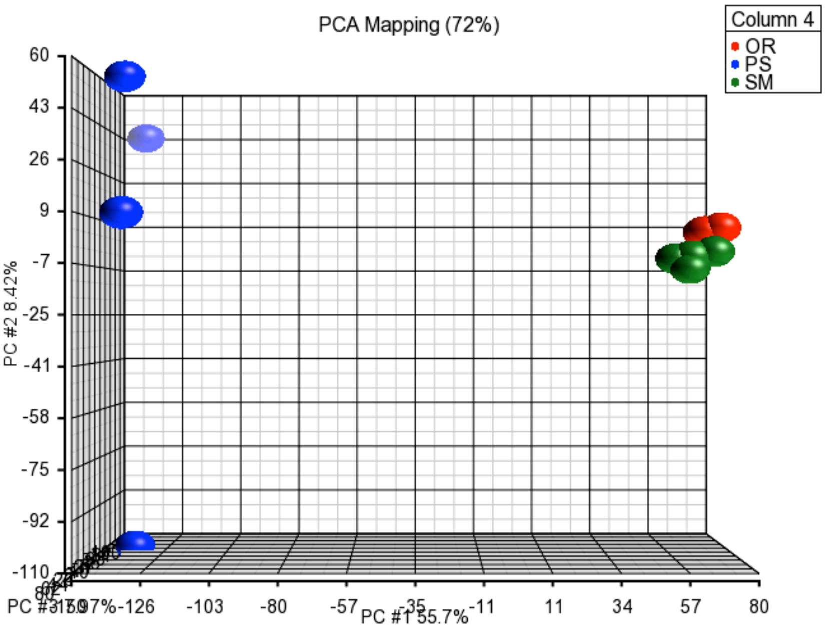


Figure 6-2 Principal component analysis (PCA) of OR, PS and SM transcriptomes.

Principal component analysis of OR (D. melanogaster Oregon-R) red, PS (D. psedoobscura) blue, SM (D.simulans) green showed the variation between OR, PS and SM. PC1 was the first major factor affected the data that was different tissues. OR and SM were much close than PS indicated the sequence distance between the PS with OR and SM, and also the PCA showed the second variation PC2 was between the replicate samples. PS showed large variation within the replicate samples but not between OR replicate samples and SM replicate samples indicating that PS hybridized on *D. melanogaster* GeneChip could not obtain the consistent results.

From Figure 6-2, we can see the first principal component accounting for 55.7% of the variation is between the different tissues. The component values for *D. melanogaster* and *D. simulans* are closely grouped but those for D. *pseudoobscura* are far apart, indicating the significant difference between *D.pseudoobscura* and *D. melanogaster* and *D.simulans*. The second component is the variation within the same species. We can see the large variation between the replicate samples of *D.pseudoobscura*; however the replicate samples of *D. simulans* and *D. melanogaster* are grouped very closely.

6.2.2 ***Genomic-DNA hybridization and probe-selection***

We used *D. melanogaster Drosophila* Genome 2.0 Array to study the transcriptome of *D. simulans* and *D. pseudoobscura*. The sequence polymorphism between the species will likely result in an underestimate of the transcripts abundance if all probe sets are used within individual sets (Ji et al., 2004). As a consequence, the probe pairs were selected by the hybridization efficiency of the genomic DNA of *D. similans* and *D. pseudoobscura* on the *D. melanogaster* GeneChip within the probesets in order to increase the sensitivity of detection.

Four replicates of *D. simulans* tubules, four replicates of *D. psedoobscura* tubules and four replicates of *D.melanogaster* Oregon-R tubules were biotin-labelled and hybridized on the *D. melanogaster* GeneChips. One sample genomic DNA of *D. simulans*, and one sample genomic DNA of *D. pseudoobscura* were biotin-labelled and put on the *D. melanogaster* GeneChip.

Figure 6-3 and Figure 6-4 show the different gDNA hybridization threshold and probe selection and probe-sets obtained for *D. simulans* and *D. pseudoobscura*. Generating the different threshold of gDNA hybridization requires generating chip definition files (CDF). This was performed by converting the gDNA hybridization .CEL files for *D. simulans* and *D. pseudoobscura* from binary to text format using AGCC power tools suite apt-cell-convert tool. Next, the CDF-_masking_2.1 Perl script available at http://affymetrix.arabidopsis.info/xspecies/ was used to convert these text files to chip definition files, generates different gDNA intensity threshold mask file for *D. simulans* and *D. pseudoobscora* in the range as 0, 50, 100, 150, 200, 300, 400, 500, 600,700, 800, 900, and 1000. The different CDF files were used for a one-way ANOVA using the Partek software to compare against with *D. melanogaster*. Different gDNA hybridization intensity threshold and probe selection for *D. simulans* see Figure 6-3. For *D. pseudoobscura* see Figure 6-4.

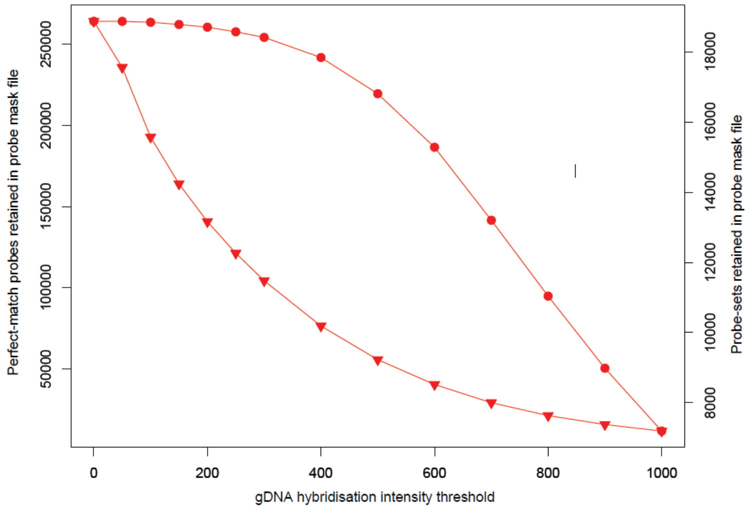


Figure 6-3 *D.simulans* gDNA intensity threshold and the probe-set and perfect-match probes retained on *D.melanogaster* GeneChip

Number of *D. melanogaster* probe-pairs and probe-sets retained from the GeneChip *Drosophila* Genome 2.0 array by the number of gDNA hybridization intensity thresholds of *D. simulans* used to generate the probe mask files. Filled circles are scaled to the right-hand y-axis and filled triangles are scaled to the left-hand y-axis.

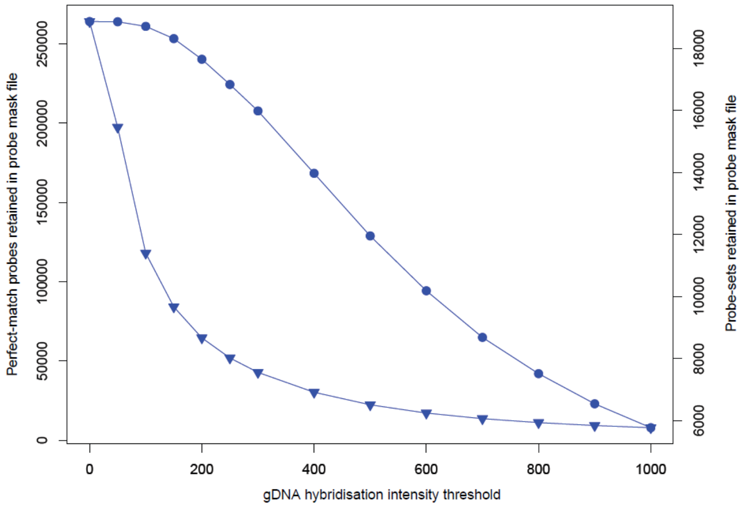


Figure 6-4 *D.pseudoobscura* gDNA intensity threshold and the probe-set and perfect-match probes retained on *D.melanogaster* GeneChip

Number of *D. melanogaster* probe-pairs and probe-sets retained from the GeneChip Drosophila Genome 2.0 array by the number of gDNA hybridization intensity thresholds of *D. pseudoobscura* used to generate the probe mask files. Filled circles are scaled to the right-hand y-axis and filled triangles are scaled to the left-hand y-axis

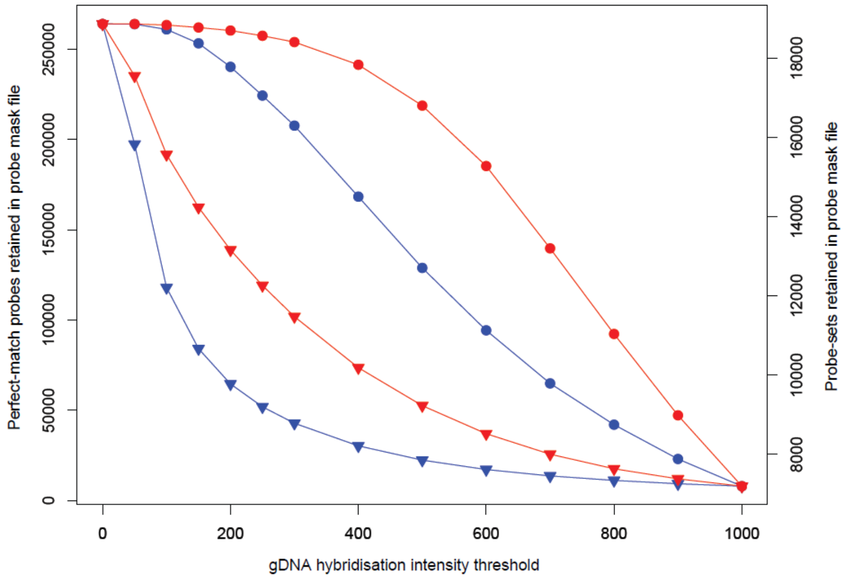


Figure 6-5 *D.pseudoobscura* and *D.simulans* gDNA intensity threshold and the probe-set and perfect-match probes retained on *D.melanogaster* GeneChip

Number of *D. melanogaster* probe-pairs and probe-sets retained from the GeneChip Drosophila Genome 2.0 array by the number of gDNA hybridization intensity thresholds of *D. simulans* (red), *D. pseudoobscura* (blue) used to generate the probe mask files. Filled circles are scaled to the right-hand y-axis and filled triangles are scaled to the left-hand y-axis.

Figure 6-3 and Figure 6-4 were optimized empirically by generating 13 probe mask files with gDNA hybridization thresholds ranging from 0 to 1000. The 13 masks were evaluated in turn to investigate which threshold is the best to represent the transcriptome of *D. simulans* and *D. pseudoobscura*. Figure 6-3 that for *D. simulans* shows the number of probe-sets dropped slowly as the gDNA intensity threshold increased. Figure 6-4 that for *D. pseudoobscura* shows the number of probe-sets dropped sharply only a gDNA intensity threshold of 200 was selected. Figure 6-5 shows the same gDNA intensity threshold but with different probe-sets and probe pairs of *D. simulans* and *D. pseudoobscur*a retained in the *D. melanogaster* GeneChip. Figure 6-5 clearly shows the difference between the behaviour of the two species on the same chip indicating that in term of sequence polymorphism, *D. simulans* and *D. melanogaster* are quite similar but *D. pseudoobscura* and *D. melanogaster* are much different.

6.2.3 *Comparing the differential gene expression of D. simulans and D. melanogaster by applying gDNA masks*

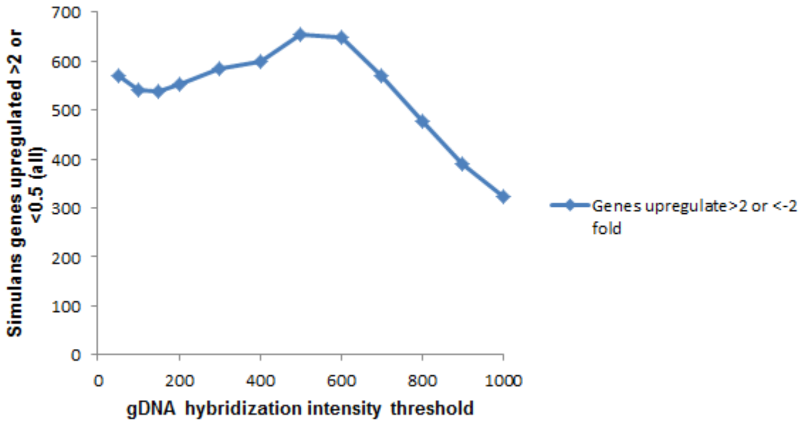


Figure 6-6 *D. simulans* and *D. melanogaster* comparison by one-way ANOVA with different gDNA masks

Different *D.simulans* gDNA hybridization thresholds used to generate probe mask files for the transcriptome analysis. Different gDNA hybridization thresholds yield different amount of genes of *D.simulans* tubules which compared with *D. melanogaster* tubules by one-way ANOVA in Partek Genomic Suit 6.6 (p<0.05). Figure 6-6 indicated at *D.simulans* gDNA threshold 500-600 yield maximum amounts of different genes.

Different *D. simulans* gDNA intensity thresholds generating different CDF files were applied to the one-way ANOVA analysis between *D. simulans* and *D. melanogaster*. The number of two folds genes with p value<0.05 and gDNA intensity thresholds were plotted in Figure 6-6. Figure 6-6 showed that at the *D. simulans* gDNA threshold 500-600 produced the maximum amount of differential expression genes between *D. simulans* and *D. melanogaster*. Figure 6-7 further using volcano plot showed the differential gene expression between *D. simulans* and *D. melanogaster*.

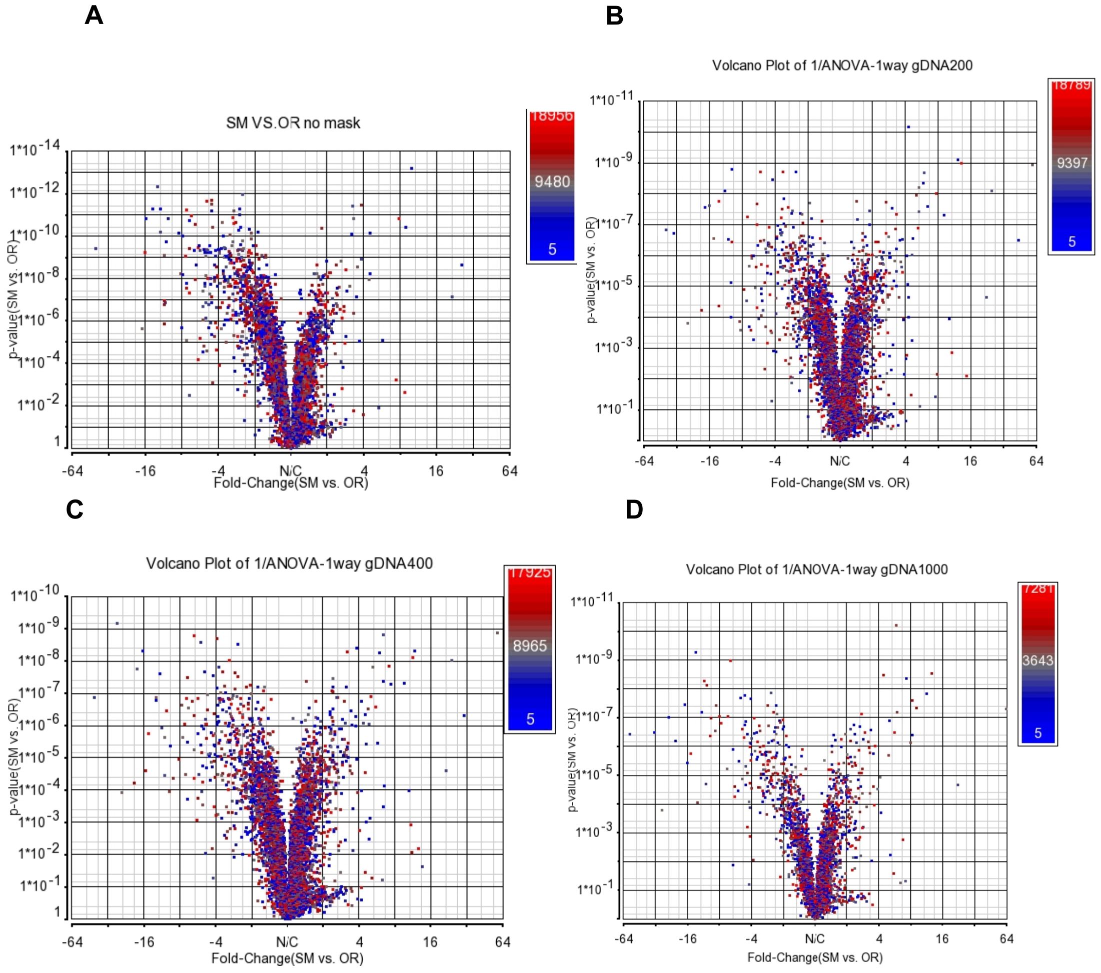


Figure 6-7 'Volcano' plot of fold change of *D. simulans* versus *D. melanogaster* by different gDNA masks

'Volcano' plots illustrating the fold change of *D. simulans* vs *D. melanogaster* (X-axis) and the relations of p-value (Y-axis) derived from one-way ANOVA from Partek Genomic Suite v6.6 (P<0.05). (A) no probe selection, during transcriptome analysis,(B), (C), (D) using probe mask file during transcriptome analysis, generating at gDNA hybridization intensity thresholds of 200, 400 and 1000 respectively.

Figure 6-7 showed the different gDNA intensity thresholds and the fold change between *D. simulans* and *D. melanogaster*. The different gDNA intensity thresholds produced the number of differential expression genes between the *D. simulans* and *D. melanogaster* was not much difference. However, gDNA intensity threshold at 400 was produced more differential gene expression than gDNA threshold at 200 and 1000 indicated the two species have very similar gene expression at the transcriptomic level.

6.2.4 *Comparing the differential gene expression of D. pseudoobscura and D. melanogaster by applying gDNA masks*

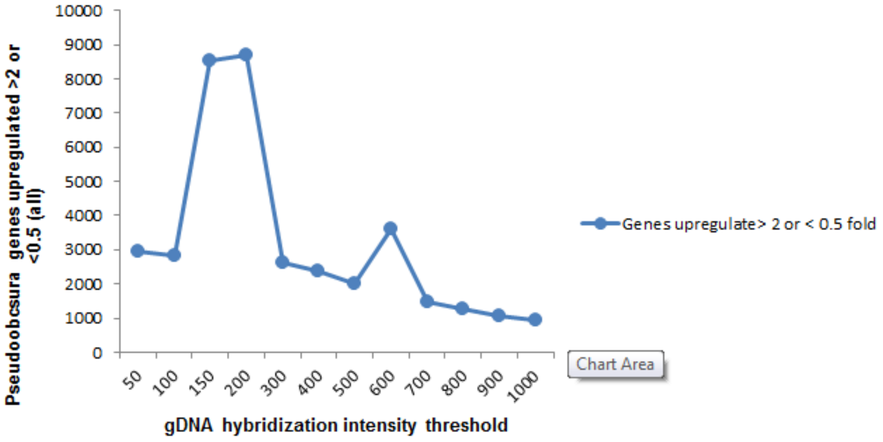


Figure 6-8 *D. pseudoobscura* and *D. melanogaster* comparison by one-way ANOVA with different gDNA masks

Different gDNA hybridization thresholds yield different amount of genes which compared *D.pseudoobscura* with *D. melanogaster* by one-way ANOVA in Partek Genomic Suit 6.6 (p<0.05). Figure 6-8 indicated at gDNA threshold 150-200 yield maximum amounts of different genes.

Different *D. pseudoobscura* gDNA intensity threshold generating different CDF files were applied to the one-way ANOVA analysis between *D. pseudoobscura* and *D. melanogaster*. The number of two folds genes with p value<0.05 and gDNA intensity thresholds were plotted in Figure 6-8. Figure 6-8 showed that at the *D. pseudoobscurs* gDNA threshold at 150-200 produced the maximum amount of differential expression genes between *D. pseudoobscura* and *D. melanogaster*. Figure 6-9 further using volcano plot showed the differential gene expression between *D. pseudoobscura* and *D. melanogaster*.

Figure 6-9 showed the different gDNA intensity threshold and the fold change between *D. pseudoobscura* and *D. melanogaster*. The different gDNA intensity thresholds produced the number of differential expression genes between the *D. pseudoobscura* and *D. melanogaster* was large difference. However, gDNA intensity threshold at 150 was produced much more differential gene expression than gDNA threshold at 400 and 1000 indicated the power of the cDNA selection affected gene expression at the transcriptomic level.

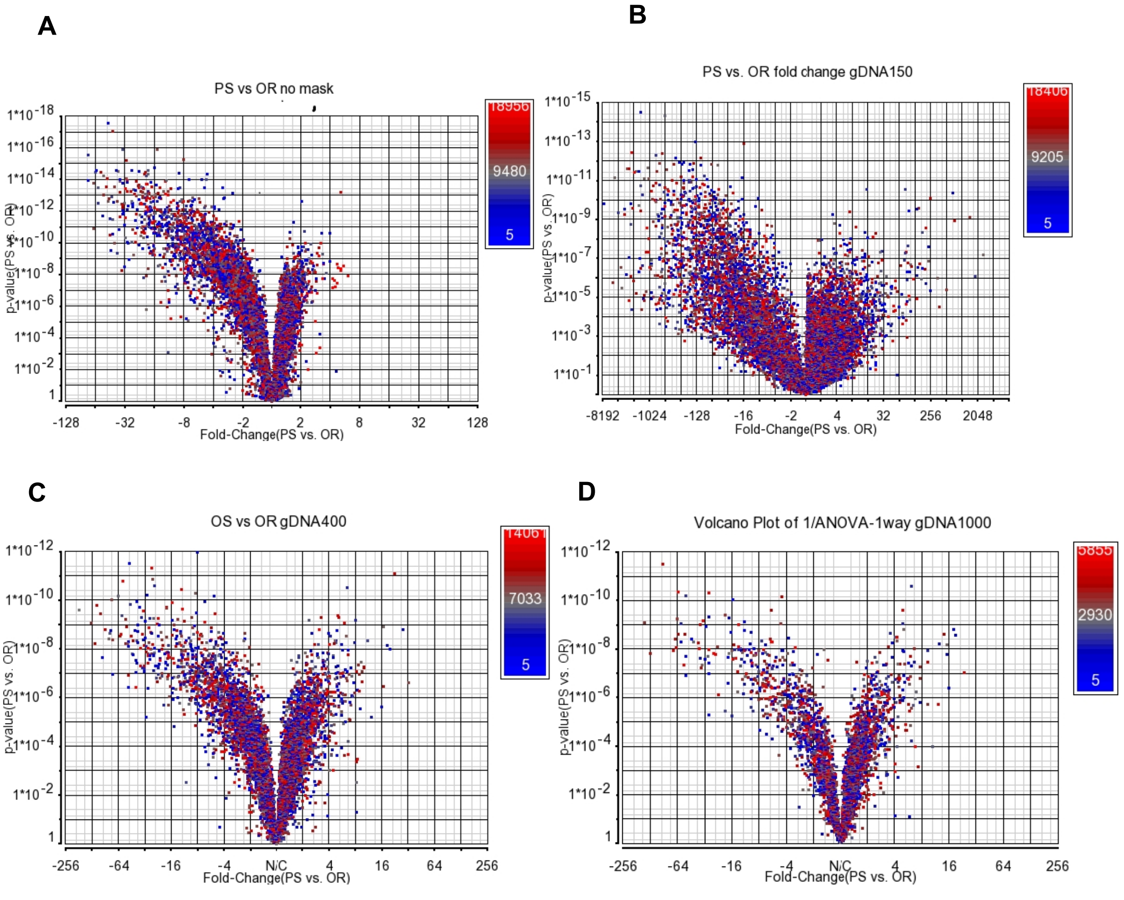


Figure 6-9 ‘Volcano’ plot of fold change of *D. simulans* versus *D. melanogaster* by different gDNA masks

‘Volcano’ plots illustrating the fold change of *D. psedoobscura* vs. *D. melanogaster* (X-axis) and the relations of p-value (Y-axis) derived from one-way ANOVA from Partek Genomic Suit v6.6. (A) no probe selection, during transcriptome analysis,(B), (C), (D) using probe mask file during transcriptome analysis, generating at gDNA hybridization intensity thresholds of 150, 400 and 1000 respectively.

6.3 **Discussion**

Following the sequences of 12 *Drosophila* species in 2007, the sequence information has provided an extensive resource for the study of the relationship between sequence and phenotypic diversity. As illustrated in Figure 6-10, the genomes of these species provide an excellent model for studying how conserved functions are maintained in the face of sequence divergence (Clark et al., 2007). Here we used gDNA masked cross-species hybridization method to investigate the possibility of measuring the transcriptomes of a species closely related to *D. melanogaster* namely *D. simulans* and are related at a medium distance namely *D. pseudoobscura* by using the *Drosophila* Genome 2.0 Array platform developed for *D. melanogaster*.

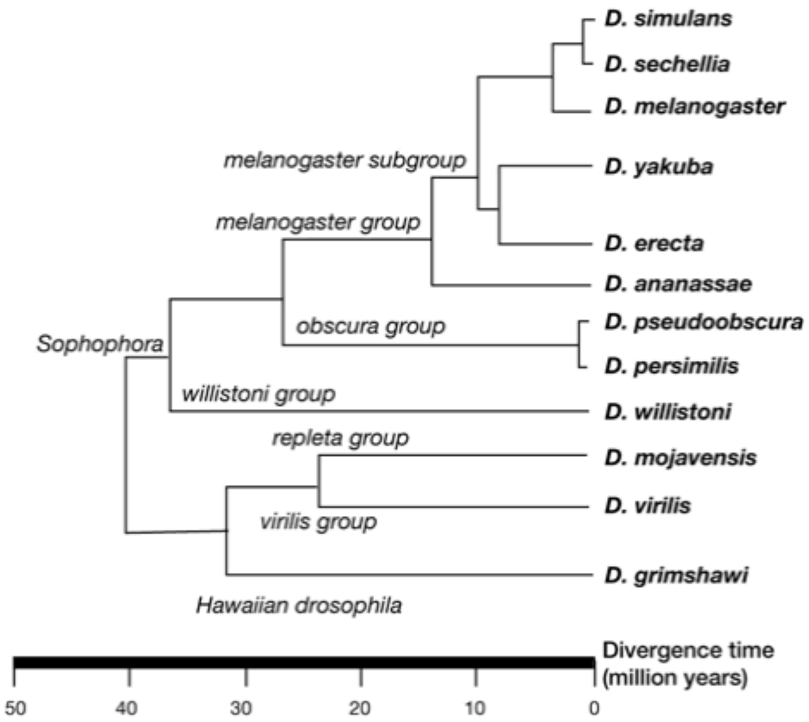


Figure 6-10 Phylogenetic tree of the *Drosophila* genus.

This figure shows the phylogenetic relationships between the *Drosophila* species of which the genomes have been sequenced (adapted from http://rana.lbl.gov/drosophila website). Closely *related - D. simulans and* medium distance – *D. pseudoobscura.*

From Figure 6-10, we can see that *D. simulans* diverged from *D. melanogaster* approximately 5.4 million years ago. So it is a close relative of *D. melanogaster* in Africa. The whole-genome alignment between *D. simulans* and *D. melanogaster* is around 82.7% (Garrigan et al., 2012). *D. pseudoobscura* diverged from *D. melanogaster* approximately 25 to 55 million years ago. So it is a medium distance relative of *D. melanogaster*. For the whole-genome alignment of *D. melanogaster* and *D. pseudoobscura*, just 48% of bases can be reliably aligned (Macdonald and Long, 2006; Richards et al., 2005). However, the level of homology with *D. melanogaster* in terms of protein-coding genes is 80% for *D. simulans* and 78.2% for *D. pseudoobscura*. So in terms of evolution, the coding sequences are more conserved than the genomic sequences. The GeneChips are designed using the 3-prime UTR genomic sequences so the sequences are much less conserved between the *Drosophila* species. The details of the sequence alignment between *D. pseudoobscura* and *D. melanogaster* can be seen in Figure 6-11. There is greater conservation in the coding region (CDS) (20-75%), but less conservation in the 3'UTR region (20-40%).

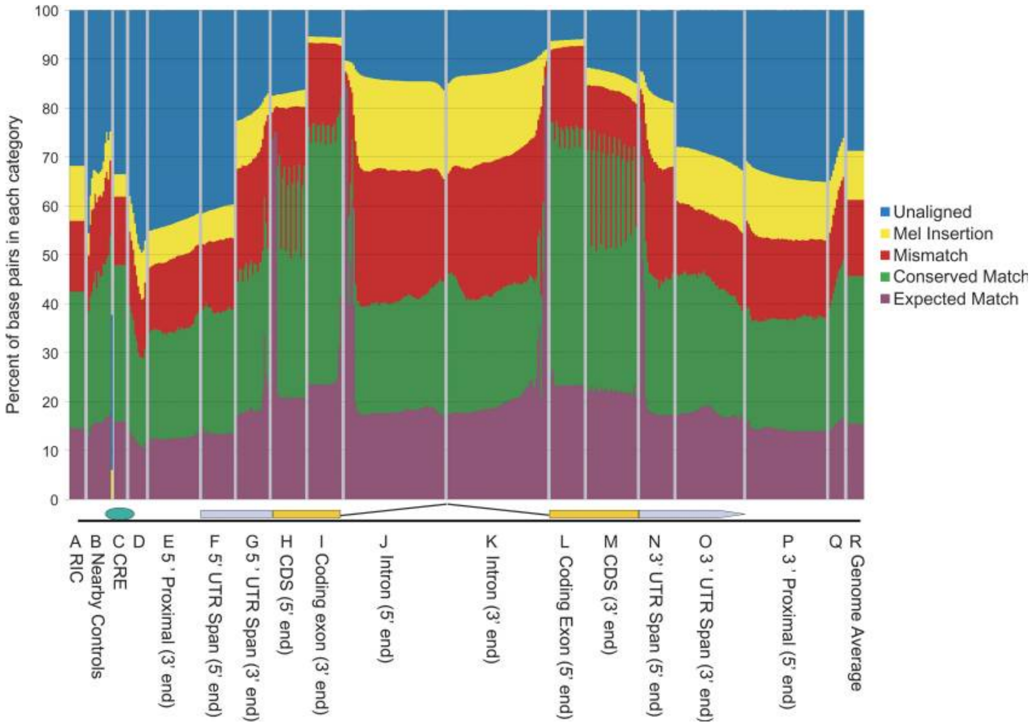


Figure 6-11 The conservation between *D. psedoobscura and D. melanogaster*

Figure 6-11 showed the average conservation of different segments of a "prototypical gene" between *D. pseudoobscura* and *D. melanogaster*. Conserved (green), Mismatch (red), Expected match (purple), Mel insertion (yellow) and unaligned (blue). At the CDS (3' end) and CDS (5' end) showed the highest conservation (green), lowest mismatch (red) and at the intron (3' end) and intron (5' end) showed the lowest conservation (green), highest mismatch (red) between *D.pseudoobscura* and *D. melanogaster*. Picture adapted from (Richards et al., 2005)

Both the histogram of $\log_2$ intensity of the probe values in Figure 6-1 and the PCA results shown in Figure 6-2 indicate similarity between D. *simulans* and D. *melanogaster*, and the difference between *D. pseudoobscura* from both *D.*

simulans and *D. melanogaster*. In particularly, the PCA results show the first component difference corresponds to most of the variability was caused by the difference between species, especially the big difference between *D. pseudoobscura* in comparison to both *D. simulans* and *D. melanogaster*. The second component difference is with the replicate samples within the species where shows the large variation within the replicate samples of *D. pseudoobscura*.

The main reason of the variation is the sequence difference between *D. melanogaster* and *D. pseudoobscura,* which may cause variation during the hybridization. As a consequence, the hybridization results may not be reproducible. For cross-species hybridization (CSH), a certain number of microarray probes are imperfectly hybridized to transcripts of the target species. Therefore, CSH performance depends on the degree of probe-transcript sequence-similarity matching (Bar-Or et al., 2007). During hybridization, a variable number of probes to target species may exhibit a perfect match, a low match, more than one match (cross-hybridization) or no match. These variations may bias the biological results (Bar-Or et al., 2007). For *Drosophila*, the further the distance between species is, the greater the variation that will occur in hybridization. The variation within the replicate samples of *D. pseudoobscura* but not between the replicate samples of *D. simulans* and *D. melanogaster* also support the less conserved sequences at the 3' end of *D. pseudoobscura* and *D. melanogaster* (Figure 6-11) caused the irreproducible results of hybridization within the species.

The percentage of present calls is reported by the Affymetrix GeneChip Operating Software (GCOS), where present calls indicate that the targeted transcript was present. The average percentage present calls for *D. pseudoobscura* tubules hybridized on *Drosophila* Genome 2.0 Array was reported by GCOS as 5.8%, for *D.simulans* tubules as 34.8%, *and for D. melanogaster* tubules as 41.3%. The sequences difference between *D. pseudoobscura* and the *D. melanogaster* is evident.

Figure 6-3, 6-4, 6-5 demonstrate that as the gDNA threshold increased, the number of perfect-match probes dropped sharply but the number of probe-sets which represent the target genes dropped slowly. At a gDNA threshold of 400 for

D. simulans, the number of perfect-match probes dropped by 30% but the probe-set represents the target genes remained at 95%. At a gDNA threshold of 150, for *D. pseudoobscura*, the number of perfect-match probes dropped by nearly 70% but the probe-set representing the genes of interest remained at 97%. This indicates the gDNA selection only reduces the mismatch probes of the probe-sets but not affect the number of probe-sets which represent the target genes. It also means that the gDNA selection will increase the sensitivity of the GeneChip measurement.

Figure 6-6, 6-8 show that as the gDNA intensity threshold increases, the level of detection of differential expression genes changes. The maximum number of differential genes at greater than 2 or less than 0.5 fold changes compared *to D. melanogaster* was detected for *D. simulans* at gDNA threshold of 500-600. Relative good numbers of differential genes were also detected for *D. simulans* at gDNA intensity threshold 400 and lower. In contrast, for *D. pseudoobscura*, the maximum number of differential genes at greater than 2 or less than 0.5 fold changes relative to *D. melanogaster* was detected at a gDNA threshold of 150-200. Based on a combination of the gene expression change and the number of genes retained, gDNA intensity threshold at 400 for *D. simulans* was chosen and a gDNA intensity threshold at 150 for *D. pseudoobscura* as the best optimized thresholds. The 'volcano' plot in Figure 6-7, 6-9 also show the large number of genes change at a gDNA intensity threshold of 400 for *D. simulans*, gDNA intensity of 150 for *D. pseudoobscura* suggest the gDNA selection method increases the sensitivity of GeneChip used by the heterologous species.

The gDNA mask results are reasonable based on a consistent level of hybridization quality. However, the PCA results lead to question the reproducibility of the gene expression result in the case of *D. pseudoobscura*. Although after the gDNA mask, the number of differential genes increase in the comparison between *D. pseudoobscura* and *D. melanogaster*. The results will be invalid if it was a poor hybridization (low match or no match). This suggests that the gDNA mask principle is only suitable for application to the closely related species. Greater caution must be taken for medium and distanced species, depending on the sequence similarity and hybridization reproducibility.

With the ongoing development of next-generation sequencing, it may prove advantageous to use RNA-seq technology in preference to cross-species hybridization array technology for transcriptome profiling of medium or distantly related *Drosophila* species or other species without a genome array available. However array technology may prove easier, faster and cheaper to use for transcriptome profiling for closely related species.

7. Conclusions and Future work

Summary

The main conclusions identified during this project are summarised in this chapter, including 3-way comparison of Drosophila tiling arrays; Drosophila expression arrays and Drosophila RNA-seq; RNA-seq is a best technology for novel gene discovery so far and reverse genetics method is a best tool to investigate the novel gene functions. Genomic DNA can be used as a mask to increase the sensitivity of arrays when applied to heterologous species.

7.1 Conclusions

Tiling arrays, three-prime expression arrays and RNA-seq are popular technologies for quantifying the gene expression levels of the entire genome. This thesis demonstrates that RNA-seq is currently the best technology, which overcomes the disadvantages of microarrays for novel gene discovery. Reverse genetics is a valuable tool for searching for novel gene functions. Microarrays remain useful, particularly in terms of using genomic DNA as a mask can increase the sensitivity to apply arrays to heterologous species.

7.1.1 *3-way comparison of Drosophila tiling arrays, expression arrays and RNA-seq*

Drosophila expression arrays have been applied in Drosophila research for over a decade as a tool for expression detection. Drosophila tiling arrays have been used as an analytical tool in recent years. RNA-seq is a cutting technology that has been used as a discovery tool as well as a gene expression tool in recent years, and has come to dominate the field of genomic research. The 3-way comparison of these three technologies can give us a view of their relative merits and drawbacks, and help us to choose the most suitable technology for our research.

Comparing Drosophila RNA-seq with Drosophila expression arrays has demonstrated that these two technologies are correlated well for detecting gene expression; both technologies have problems with genes at low expression levels. Microarrays have problems with cross-hybridization, hybridization noise, miss target affects, lower dynamic range and three-prime bias, all of which will reduce its capability. However, it still offer a means to measure three-prime end processing (Cui and Loraine, 2009). In contrast, RNA-seq supports a wider dynamic range than microarrays through its increased read coverage, and so is able to detect genes at low expression levels. RNA-seq does not depend on previous knowledge of the genome, can be used for any organism, and also offers increased power for novel gene and alternative splicing discovery. RNA-seq had taken genomic research into a new level. Microarrays remain useful and complementary to RNA-seq for transcriptome profiling.

Comparing Drosophila tiling arrays with Drosophila expression arrays revealed that the two technologies have a reasonable correlation for gene expression levels. However, the two technologies suffered from cross-hybridization, background noise and low dynamic range, and so it can prove difficult to obtain agreement about the lower and higher expression genes.

Comparing RNA-seq with tiling arrays showed that both technologies had the ability to discover the transcription “dark matter” within the genome. However, tiling arrays have a high false-positive discovery rate, may over estimate the transcriptional activities in the genome, and are ill-suited to accurately detect transcripts at low levels (van Bakel et al., 2010). As a consequence, the results of tiling arrays must be interpreted with caution or must be confirmed by other molecular methods.

In summary, RNA-seq, tiling arrays and expression arrays complement each other in terms of their performance for transcriptome profiling as well as novel gene discovery.

7.1.2 *RNA-seq is the best gene discovery technology so far*

Poly (A) selection RNA-seq has been used as a tool for discovering novel genes in *Drosophila* tubules. By applying the TopHat and Cufflinks pipeline, we have been able to find a number of novel genes that belonged to coding and noncoding RNAs and were confirmed by RT-PCR. The results suggested that the novel genes were lowly expressed, most of them had two exons, belonged to noncoding RNAs and existed in a tissue-specific manner.

7.1.3 *Reverse genetics is the best technique so far for functional study of novel genes*

Drosophila as an experimental organism for functional genomics is suitable for applying reverse genetics. The Gal4/UAS system made Drosophila uniquely suitable to knock-down any genes within the genome. In the work presented in this thesis, RNA-seq was used to find the novel gene CG43968. Reverse genetics had been applied to CG43968 for investigation of gene function.

Using Gal4/UAS-RNAi, the novel gene CG43968 was knocked down to confirm the functional involvement in tubule secretion located in tubule principle cells. The possible function of CG43968 may play an essential role in apical plasma membrane secretion upon stimulation. CG43968 has a Latrophilin/CL-1-like GPS domain, which may play a role in renal tubular development. CG43968 may be a membrane protein of tubule which contains a GPS domain, and is responsible for taking extracellular calcium into the principal cell.

Using Gal4/UAS-PTWV to over express CG43968, the novel gene has been shown to be present in cytoplasm and cell membrane. In situ hybridization also revealed the novel gene CG43968 located in the main segments of tubules.

7.1.4 *Species array*

Applying genomic DNA-based probe-selection method increased the sensitivity of the arrays in order to measure gene expression from the *Drosophila melanogaster* closely-related species *Drosophila simulans,* and the medium-distance-related species *Drosophila pseudoobscura* by using Drosophila melanogaster expression arrays. The results revealed that genomic DNA-based probe-selection method would indeed increase the sensitivity of the arrays, making them suitable to apply to closely-related species of the model organism but not suitable to apply to medium- or far-distance-related species.

7.2 Future Work

Many aspects of the work presented in this thesis could be extended to provide further insight into the function of *Drosophila Malpighian* tubules. Future work could be carried out in the following areas:

1. Further investigation of the function of novel gene CG43968.

    a) **By using a calcium assay**, since CG43968 has been confirmed as having function in tubule secretion, possibly related to calcium transport in principal cell of tubules. Measuring the calcium levels in the normal flies and the CG43968 knockdown flies will help us understand if CG43968 has function related to calcium transport.

    b) **By designing an antibody,** since CG43968 is a novel gene for which no antibody is available at the moment. Designing an antibody for CG43968 will further help to identify the location of the protein product of this gene in the cell.

    c) By signaling pathway investigation: CG43968 has a Latrophilin/CL-1-like GPS domain that may modulate intracellular calcium, and is responsible for taking extracellular calcium into the principal cells to induce CAP2b to activate nitric oxide production and then stimulate cGMP for tubule secretion. Further investigating the pathways will help in understanding the function of CG43868.

    d) **By fluid secretion metabolomics.** Comparing the different metabolomics of the secreted fluid of the tubules of CG43968 knockdown flies and that of normal flies may help us to further understand the function of CG43968.

2. The 3-way comparison results from this thesis suggested that tiling arrays overestimated the genome transcription activities. Investigating previous *Drosophila* tiling arrays research by using matched RNA-seq data will find out if the novel transcriptional fragments are real or affected by false-positive signals (Manak et al., 2006; van Bakel et al., 2010).

3. Search for novel alternative splicing units in Drosophila tubules by using the existing data set, especially those alternative splicings that change the coding sequence. Confirm any results by molecular genetics method.

4. Search for more novel genes (including noncoding genes) and their functions in tubules.

a) **By using ribosome reduction strand-specific RNA-seq.** Poly (A) selected RNA-seq mainly finds the poly (A) related transcripts. However, more novel transcripts are noncoding genes. In recent years, more efforts have been put into investigating the functions of noncoding genes. Noncoding genes have been considered to play important gene regulatory roles in the genome such as in directing post-transcriptional regulation of gene expression or in guiding RNA modifications (Carpenter et al., 2013; Eddy, 2001).

b) **Looking for regions transcribed on both strands**. Using existing data or ribosome reduction RNA-seq data to look for the regions transcribed on both strands. This will help us understand the role of antisense RNA.

c) **Perform reverse genetics to non-coding RNAs.** There is not (yet) a huge catalogue of mutations in ncRNAs that have been shown to affect phenotype, compared to those in protein-coding sequences. However, on the assumption that most ncRNAs have regulatory roles and that most regulatory regions have not yet being identified, next-generation sequencing technology may help to identify the ncRNAs in *Drosophila* and reverse genetics methods may help to verify the function of those ncRNAs.

d) **Assigning genetic signatures**: it is no surprise that this may be the case (Mattick, 2009). *Drosophila* has been armed with the GAL4/UAS system and wealthy physiology data that make it an ideal model to perform reverse genetics technique (siRNA-mediated gene knockdown and overexpression). These techniques are emerging major tools to investigate noncoding gene function.

e) **Metabolomics studies for function of noncoding RNAs**: Noncoding RNAs has been reported to play important roles in metabolic pathways. For example, small RNAs such as Spot42, Glmz and others are post-transcription regulators of bacteria sugar and control sugar metabolism (Gorke and Vogel, 2008). Gene expression data combined with metabolomics data may help us understand the noncoding RNA function at system biology levels.

Appendix I: Fly food recipe

Fly food recipe (Mix the contents in 1 litre of H_2O in the below order of preference.
10 g Tayo agar
1 tbsp Soya fluor
15 g Sucrose
33 g Glucose
15 g Maize meal
10 g Wheat germ
30 g Treacle
35 g Yeast
Bring to boil, stirring constantly; simmer 10 min; allow to cool slightly to about 70 ºC; leave for 20 min and then add:
10 ml Nipagin (of below formulation)
5 ml Propionic acid
[Nipagin = 25 g. Nipagin M (Tegosept M, p-hydroxybenzoic acid methyl ester) in 250 ml Ethanol]
Dispense:
Fly Vials = 8 ml
Fly Bottles = 70 ml

Appendix II: Phosphate Buffer Saline (PBS)

Phosphate Buffer Saline (PBS) (in H_2O)
137 mM NaCl
2.7 mM KCl
10 mM Na_3PO_4
2 mM KH_2PO_4, pH 7.4
Other solutions using PBS
For PBST: 0.25% TritonX-100 was added.
For PBSTw: 1% Tween20 was added.
For blocking buffer for westerns: 10% non-fat milk power was added to PBSTw.
For blocking buffer for ICCs: 10% goat serum was added to PBST.

Appendix III: *Drosophila* Schneider's media

www.invitrogen.com, (accessed on 26th August 2011)

COMPONENTS	Molecular Weight	Concentration (mg/L)	mM
Amino Acids			
Glycine	75	250	3.33
L-Arginine	174	400	2.3
L-Aspartic acid	133	400	3.01
L-Cysteine	121	60	0.496
L-Cystine	240	100	0.417
L-Glutamic Acid	147	800	5.44
L-Glutamine	146	1800	12.33
L-Histidine	155	400	2.58
L-Isoleucine	131	150	1.15
L-Leucine	131	150	1.15
L-Lysine hydrochloride	183	1650	9.02
L-Methionine	149	800	5.37
L-Phenylalanine	165	150	0.909
L-Proline	115	1700	14.78
L-Serine	105	250	2.38
L-Threonine	119	350	2.94
L-Tryptophan	204	100	0.49
L-Tyrosine	181	500	2.76
L-Valine	117	300	2.56
beta-Alanine	89	500	5.62
Inorganic Salts			
Calcium Chloride ($CaCl_2$-$2H_2O$)	147	794	5.4
Magnesium Sulfate ($MgSO_4$-$7H_2O$)	246	3700	15.04
Potassium Chloride (KCl)	75	1600	21.33

Potassium Phosphate monobasic (KH_2PO4)	136	450	3.31
Sodium Bicarbonate ($NaHCO_3$)	84	400	4.76
Sodium Chloride (NaCl)	58	2100	36.21
Sodium Phosphate monobasic (NaH_2PO_4-$2H_2O$)	156	1321	8.47
Other Components			
Alpha-Ketoglutaric acid	146	200	1.37
D-Glucose (Dextrose)	180	2000	11.11
Fumaric acid	116	100	0.862
Malic acid	134	100	0.746
Succinic acid	118	100	0.847
Trehalose	342	2000	5.85
Yeastolate		2000	-

Appendix IV: *E. coli* growth media

COMPONENTS	grams/litre
LB-broth	
Bacto-tryptone	10
Dried Yeast	5
NaCl	10
LB-agar	
Bacto-tryptone	10
Dried yeast	5
NaCl	10
Bacto-agar	15
SOC broth	
Bacto-tryptone	2 % (w/v)
Dried yeast	0.5 % (w/v)
NaCl	10 mM
KCl	2.5 mM
$MgCl_2$	10 mM
$MgSO_4$	10 mM
Glucose	20 mM

Appendix V: Buffers for SDS-PAGE and Westerns

From Sambrook and Russell, 2001
6 x SDS-PAGE Loading buffer
0.35 M Tris HCl, pH6.8
10.28 % (w/v) SDS
36 % v/v glycerol
5 % v/v b-mercaptoethanol
0.012 % w/v bromophenol blue
in 0.5 ml aliquots stored at -20°C
Tris-Glycine Running Buffer (in 500 ml of H_2O)
7.2 g Glycine
1.5 g Tris Base
6 ml 10% (w/v) SDS
Staining Solution
465 ml Brilliant blue R concentrate (Sigma)
535 ml H_2O
Destaining Solution (in H_2O)
10 % (v/v) Acetic Acid
45% (v/v) Methanol
Ponceau S Staining Solution (in 500 ml H_2O)
1.5 g TCA
0.5 g Ponceau S stain
Transfer Buffer (in 1 litre of H_2O)
20 % (v/v) Methanol
14.4 g Glycine
3 g Tris Base

Resolving and Stacking gels for SDS-PAGE (from Sambrook and Russell, 2001)	
COMPONENTS	Vol. (ml)
Resolving gel 10%, volume for 2x 5 ml gels	
H_2O	4
30 % acrylamide mix	3.3
1.5 M Tris (pH 8.8)	2.5
10 % (v/v) SDS	0.1
10 % (v/v) APS	0.1
TEMED	0.004
Stacking gel 5%, volume for 2x 1.5 ml	
H_2O	2.1
30 % acrylamide mix	0.5
1.0 M Tris (pH 6.8)	0.38
10 % (v/v) SDS	0.03
10 % (v/v) APS	0.03
TEMED	0.003

Appendix VI RT-PCR products and primers for novel genes

RT-PCR products and primers for novel genes from section 3.2.10

>Chr3L 23777293-23780611 cDNA RT-PCR product

GTAGCATTTATGTGCCTATTGCCTAATTTCGCACTTTTTCAAGAAAACTTGACAAGGGATAAAA
TGCCTTGACGAATTTACTAATACAACATCAGGCACGGCATATTGGAAGTATAAAGGCATTCCTT
ACGCAATATCTGAAGATCTGCTGTGCCTTAAGTCTAACCATC

Primer Forward GTAGCATT TATGTGCCTATTGC

Primer Reverse GATGGTTAGACTTAAGGCACAG

> Chr2R 16189454-16190517 cDNA RT-PCR products (red sequence are introns)

Product1.
ACTCGGAGGCTTCTTCTGGTTGCGGCCAAGTGTCAGTACATTAATCATGAGGTGATTTATGGC
GCCCACGCCATCGACGCATGTGCTCACGTTTTTCTCGCTGCCAGCAGCAACATCAGCGCGTG
TGGGCAGCCCACCGCCACCCACCCACTCCCACTGCCACCACCACCTCCTCCGATTCCCCCGA
AAACACCTGGCCCCAACGCATGCATTGTTATGCTCCGTTAGCCGTTTTAGAAGCG

Product2.
ACTCGGAGGCTTCTTCTGGTTGCGGCCAAGTGTCAGTACATTAATCATGAAACATCAGCGCGT
GTGGGCAGCCCACCGCCACCCACCCACTCCCACTGCCACCACCACCTCCTCCGATTCCCCC
GAAAACACCTGGCCCCAACGCATGCATTGTTATGCTCCGTTAGCCGTTTTAGAAGCG

Primer Forward ACTCGGAGGCTTCTTCTGGT

Primer Reverse CGCTTCTAAAACGGCTAACG

> Chr2R 184996-184503 cDNA RT-PCR products (red sequence are introns)

Product1

GCAAGAACTTGGCTTCGTAAGGGTGAGAGGAGTCAGTGGTCGGTACAGGTGGCCCCAGGAC
GAGCGTTGCCTCGCGGACGATATACCCTGCCCCATAATAATCCTAAACCCATACCGACCGGC
AGGTGGTCTTCCAGAGAGAC

Product 2

GCAAGAACTTGGCTTCGTAAGGGTGAGAGGCCCCATAATAATCCTAAACCCATACCGACCGG
CAGGTGGTCTTCCAGAGAGAC

Primer Forward GTCTCCTCTGGAAGACCACC

Primer Reverse GCAAGAACTTGGCGTGTTCG

>Chr3R 1096703-095876 cDNA and gDNA RT-PCR Product (346bp)

TAATTCGCACAATTCGCGGCAGATATTCGGCCAGGTATGCTTCAGATATGCATATAATATACAC
ATACATATGTACCCCTTCTTAGAGATAGATTTGCGATTGTTAGGTGCTGAAGACGACCTCCGCT
TTTTCAGTTCGACCCTGTAGAATGCTGATTGTAGAACCGCGCGATTGTATAAACTCCACGTAG
AAGGGAGCACCACTCTATCTATCCAGGCCACAACTTAATGTCCATGCCACATGCCACACATGT
ATGTTAAGTGGGTGACTGGACGAGAGGAAGGATTTTACAAAGGATACAGATAAATCGATCGGA
GATTGAGGCAGTTGGATGTGGATGCAGCA

Primer Forward TGCTGCATCCACATCCAACT
Primer Reverse TAATTCGCACAATTCGCGGC

>Chr3R 5951869-5965492 cDNA RT-PCR product
CCTAGAGAATGGCGGAACGGACTGTTCGGCTGACAAACACAGAAGGCAATATTTACTGTTCTG
TCATAGGTGTCACTGATGTTTAAAAATACACTACCCGAACACTAGAGATGCAAAAATAAACAAA
CGAAATGAAAGGTCTATTAAATGTGTGTGGCATGTGAATGGCTG

Primer Forward CCTAGAGAATGGCGGAACGG
Primer Reverse CAGCCATTCACACTGCCACAC

>Chr2L 9669700-9670824 gDNA RT-PCR product (835bp)

ACAATGGCCGGGTAATAACTGAAAGGTGAGCACAGAACACAACTGTCAGTTGGATCTAAAAAT
ATTTTAAAATTTCCGATAAGCTGTCACTTCTAGTATATCCCTTACTTTTAGTCACACGGCTTCCG
TCTTTCACAGATATCTCTAATCGAGCCACCACCTTTCGATTAGCTGCACTCCAATCAAGGCAGC
TTATCTGTTCGTGATTGGCATGATTCCTGCCAGCGGGTGCCCACTGTAATAAATAGTGCCCGA
AATGGCACTCAAGTGTCGGCCACTTAATAACGAATTTCTGGCTGCCCGAACAAGTCGTAAAGA
TGCATCGCAGCTCGGATTGTGGTCCCATCGGAACTGCACTTTAAGAGATGTTTGCAAAAAGAA
AGTGAAAGAGCGCAAAGGTCAGCGGTGGGGGGAAATTCAATGTGAAAAGCGGCGATGTCGG
CTGGGTTACAGCGTTTTCAGAAGGGCTTCCCATCTGCATTTTCCTCCTCCATTCACTGACTATT
TATTTGTTCTTATTGTTTTTTTTTTTTGCTGCTCTTCATAACTGTTGCATATACATATAAATGCGAA
CCATTGCGTATACTCAATTAATTGAGACAAATTACCCATACGCCGGGTGGGTGAATGACTGTC
GAGAGTTCAATTCAGTTTGAATTGAAGTGTTTTACTTTGGTTGTTATTTTTTATTGCATTGTCTG
GTTGCACCAACCGAATAAAAATGAAAATTTCGCCATAATAACTTTAATTTGCCCCAATCGATCC
AATGGCAATTATTATTAGATTTCGTTTCGTTTTGTTTGATTCTGATTCTGATTTCTGCTCCCATTC
GGCA

>Chr3R 2441161-14417150 gDNA RT-PCR product (428bp)

CTAATACTAAATGTATCTATATTTGAACCATTACCTTACGCCGAAGTAGGATAGCTGCAAATGT
ATTATGAAAATATAAGAAATAAATAAAAGAACCGAACTTATCGCAGTGGCGTTGTGATTTTTTGA
GTGTCGAAACGCCAAAGCAAATATGATTGGAGTTTTATTTACTCTGGCCCGTCCGATTGGGTC
TTAATTACTTCCGAAAACAATGACAACCGATGACCAGTGGCGATACGAGCCCCGTCGCCGAG
CATTAATCAATTTACAACGAGATTTAAGCAATGACATCGGTAAATAATAAAATAAATACAATTTG
AATACCGGTTTGCCATTTTCTGGCTATCATTCATAAACTTTTGTGCGGCCTGACATGAAAATTA
GCTCTTCAGCCAAGGCAGACAGCGCAGCGATCTCCGAGATCTCGTAGATCGGAGATCGCAGA
TCGGAGTCGTCTAAAGCTGGAACTCCGATCGCTGAAGAGCTCCAGACTCCGGAGTTCGAGTT
GGCGATGGAGATGGAGACCTG

Primer Forward CTTATCGCAGTGGCGTTGTG
Primer Reverse CAGGTCTCCATCTCCATCGC

Appendix VII Comparison of rank of top 30 genes

Comparison of rank of top 30 genes from FlyAtlas and RNA-seq (tubule/whole fly) by fold change

Gene Symbol	RNA-seq				Microarray			
	FC	p-value	FDR	Rank	FC	p-value	FDR	Rank
CG18095	132	0.00E+00	0.00E+00	24	135	2.78E-09	1.30E-07	1
CG15408	98	0.00E+00	0.00E+00	47	108	9.06E-06	9.66E-06	2
CG42235	133	0.00E+0	0.00E+00	23	105	2.01E-07	8.01E-07	3
CG11407	11	2.62E-14	2,32E-13	411	86	9.16E-07	2.04E-07	4
CG5697	60	0.00E+00	0.00E+00	104	74	7.18E-06	5.95E-06	5
Oatp58Da	59	0.00E+00	0.00E+00	111	72	9.36E-07	2.07E-06	6
st	82	0.00E+00	0.00E+00	70	71	1.53E-07	1.53E-07	7
CG13905	139	0.00E+00	0.00E+00	21	71	5.41E-08	4.43E-07	8
Pkg21D	52	6.37E-13	5.08E-12	132	70	3.86E-09	1.49E-07	9
CG3690	72	0.00E+00	0.00E+00	82	68	4.54E-06	5.96E-06	10
CG8028	103	0.00E+00	0.00E+00	42	67	1.73E-08	2.87E-07	11
CG33282	46	0.00E+00	0.00E+00	144	67	1.30E-07	6.40E-07	12
CG18473	51	0.00E +00	0.00E+00	33	67	2.72E-07	9.53E-07	13
Swi2	91	0.00E+00	0.00E+00	55	67	0.000113	6.39E-05	14
CG8620	87	0.00E+0.00	0.00E+00	65	65	1.31E-09	1.08E-07	15
CG32024	82	0.00E+00	0.00E+00	69	65	6.49E-06	7.64E-06	16
CG14606	61	0.00E+00	0.00E+00	103	61	2.01E-12	3.39E-09	17
CG33281	39	0.00E+00	0.00E+00	166	63	1.28E-07	6.37E-07	18
CG17751	73	0.00E+00	0.00E+00	80	63	1.47E-05	1.36E-05	19
CG10006	157	0.00E+00	0.00E+00	17	63	5.17E-08	4.37E-07	20
Sr-CIV	181	0.00E+00	0.00E+00	181	62	1.21E-08	2.46E-07	21
CG8837	47	0.00E+00	0.00E+00	145	61	2.81E-08	3.49E-07	22
CG9270	111	0.00E+00	0.00E+00	33	61	7.46E-09	2.13E-07	23
CG31090	not found				61	1.12E-06	2.34E-06	24
CG1139	47	0.00E+00	0.00E+00	143	58	2.00E-09	1.30E-07	25
CG42235	132	0.00E+00	0.00E+00	23	56	9.92E-07	2.15E-06	26
CG14963	68	0.00E+00	0.00E+00	68	56	1.43E-05	1.33E-05	27
Cyp6a18	68	0.00E+00	0.00E+00	91	55	5.68E-09	1.95E-07	28
CG1736	79	0.00E+00	0.00E+00	73	55	1.63E-07	7.06E-07	29
CG14957	53	0.00E+00	0.00E+00	131	54	0.000788	0.000303	30

Appendix VIII Contributions from the PhD project

Chintapalli, V., Wang, J., and Dow, J. (2007). Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. Nat Genet *39*, 715-720.

Venkateswara R. Chintapalli, Selim Terhzaz, Jing Wang, Mohammed Al Bratty, David G. Watson, Pawel Herzyk, Shireen A. Davies and Julian A. T. Dow (2012). Functional correlates of positional and gender-specific renal asymmetry *in Drosophila*. PLoS ONE 7(4): e32577.

Chintapalli R Venkateswara, Wang Jing, Herzyk Pawel, Davies A Shireen, Dow AT Julian (2013). Data-mining the FlyAtlas online resource to identify core functional motifs across transporting epithelia. BMC Genomics **14**, 518-529

Jing Wang, Pawel Herzyk, Julian A. T. Dow. 3-way analyzes of *Drosophila* RNA-seq, *Drosophila* tiling microarrays and *Drosophila* expression microarrays. (in Prep)

Jing Wang, Venkateswara R. Chintapalli, Pawel Herzyk, Shireen A. Davies and Julian A. T. Dow. New views on the *Drosophila* transcriptome. (in Prep)

References

Abruzzi, K.C., Rodriguez, J., Menet, J.S., Desrochers, J., Zadina, A., Luo, W., Tkachev, S., and Rosbash, M. (2011). Drosophila CLOCK target gene characterization: implications for circadian tissue-specific gene expression. Genes Dev *25*, 2374-2386.

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., *et al.* (2000). The genome sequence of Drosophila melanogaster. Science *287*, 2185-2195.

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F., *et al.* (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. Science *252*, 1651-1656.

Adams, M.D., Kerlavage, A.R., Fields, C., and Venter, J.C. (1993a). 3,400 new expressed sequence tags identify diversity of transcripts in human brain. Nat Genet *4*, 256-267.

Adams, M.D., and Sekelsky, J.J. (2002). From sequence to phenotype: reverse genetics in Drosophila melanogaster. Nat Rev Genet *3*, 189-198.

Adams, M.D., Soares, M.B., Kerlavage, A.R., Fields, C., and Venter, J.C. (1993b). Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. Nat Genet *4*, 373-380.

Agarwal, A., Koppstein, D., Rozowsky, J., Sboner, A., Habegger, L., Hillier, L.W., Sasidharan, R., Reinke, V., Waterston, R.H., and Gerstein, M. (2010). Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. BMC Genomics *11*, 383.

Aggarwal, G., and Ramaswamy, R. (2002). Ab initio gene identification: prokaryote genome annotation with GeneScan and GLIMMER. J Biosci *27*, 7-14.

Akbari, O.S., Oliver, D., Eyer, K., and Pai, C.Y. (2009). An Entry/Gateway cloning system for general expression of genes with molecular tags in Drosophila melanogaster. BMC Cell Biol *10*, 8.

Alba, R., Fei, Z., Payton, P., Liu, Y., Moore, S.L., Debbie, P., Cohn, J., D'Ascenzo, M., Gordon, J.S., Rose, J.K., *et al.* (2004). ESTs, cDNA microarrays, and gene expression profiling: tools for dissecting plant physiology and development. Plant J *39*, 697-714.

Allan, A.K., Du, J., Davies, S.A., and Dow, J.A. (2005). Genome-wide survey of V-ATPase genes in Drosophila reveals a conserved renal phenotype for lethal alleles. Physiol Genomics 22, 128-138.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol *11*, R106.

Arbeitman, M.N., Furlong, E.E., Imam, F., Johnson, E., Null, B.H., Baker, B.S., Krasnow, M.A., Scott, M.P., Davis, R.W., and White, K.P. (2002). Gene expression during the life cycle of Drosophila melanogaster. Science *297*, 2270-2275.

Argani, P., Rosty, C., Reiter, R.E., Wilentz, R.E., Murugesan, S.R., Leach, S.D., Ryu, B., Skinner, H.G., Goggins, M., Jaffee, E.M., *et al.* (2001). Discovery of new markers of cancer through serial analysis of gene expression: prostate stem cell antigen is overexpressed in pancreatic adenocarcinoma. Cancer Res *61*, 4320-4324.

Arya, R., Mallik, M., and Lakhotia, S.C. (2007). Heat shock genes - integrating cell survival and death. J Biosci *32*, 595-610.

Bacolod, M.D., Schemmann, G.S., Giardina, S.F., Paty, P., Notterman, D.A., and Barany, F. (2009). Emerging paradigms in cancer genetics: some important findings from high-density single nucleotide polymorphism array studies. Cancer Res *69*, 723-727.

Bar-Or, C., Bar-Eyal, M., Gal, T.Z., Kapulnik, Y., Czosnek, H., and Koltai, H. (2006). Derivation of species-specific hybridization-like knowledge out of cross-species hybridization results. BMC Genomics *7*, 110.

Bar-Or, C., Czosnek, H., and Koltai, H. (2007). Cross-species microarray hybridizations: a developing tool for studying species diversity. Trends Genet *23*, 200-207.

Bartosiewicz, M., Penn, S., and Buckpitt, A. (2001a). Applications of gene arrays in environmental toxicology: fingerprints of gene regulation associated with cadmium chloride, benzo(a)pyrene, and trichloroethylene. Environ Health Perspect *109*, 71-74.

Bartosiewicz, M.J., Jenkins, D., Penn, S., Emery, J., and Buckpitt, A. (2001b). Unique gene expression patterns in liver and kidney associated with exposure to chemical toxicants. J Pharmacol Exp Ther *297*, 895-905.

Bebenek, K., Abbotts, J., Roberts, J.D., Wilson, S.H., and Kunkel, T.A. (1989). Specificity and mechanism of error-prone replication by human immunodeficiency virus-1 reverse transcriptase. J Biol Chem *264*, 16948-16956.

Bellingham, S.A., Coleman, B.M., and Hill, A.F. (2012). Small RNA deep sequencing reveals a distinct miRNA signature released in exosomes from prion-infected neuronal cells. Nucleic Acids Res *40*, 10937-10949.

Bellis, M. (2013). Estimating the similarity of alternative Affymetrix probe sets using transcriptional networks. BMC Res Notes *6*, 107.

Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N., and Golani, I. (2001). Controlling the false discovery rate in behavior genetics research. Behav Brain Res *125*, 279-284.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. J Roy Stat Soc B Met *57*, 289-300.

Boguski, M.S., Lowe, T.M., and Tolstoshev, C.M. (1993). dbEST--database for "expressed sequence tags". Nat Genet *4*, 332-333.

Boguski, M.S., Tolstoshev, C.M., and Bassett, D.E., Jr. (1994). Gene discovery in dbEST. Science *265*, 1993-1994.

Bonaldo, M.F., Lennon, G., and Soares, M.B. (1996). Normalization and subtraction: two approaches to facilitate gene discovery. Genome Res *6*, 791-806.

Botstein, D., and Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet *33 Suppl*, 228-237.

Bradford, J.R., Hey, Y., Yates, T., Li, Y., Pepper, S.D., and Miller, C.J. (2010). A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling. BMC Genomics *11*, 282.

Brand, A.H., and Perrimon, N. (1993). Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. Development *118*, 401-415.

Bratty, M.A., Chintapalli, V.R., Dow, J.A., Zhang, T., and Watson, D.G. (2012). Metabolomic profiling reveals that Drosophila melanogaster larvae with the y mutation have altered lysine metabolism. FEBS Open Bio 2, 217-221.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., *et al.* (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet *29*, 365-371.

Brown, S.D., and Peters, J. (1996). Combining mutagenesis and genomics in the mouse--closing the phenotype gap. Trends Genet *12*, 433-435.

Bullard, J.H., Purdom, E., Hansen, K.D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics *11*, 94.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev *25*, 1915-1927.

Carpenter, S., Aiello, D., Atianand, M.K., Ricci, E.P., Gandhi, P., Hall, L.L., Byron, M., Monks, B., Henry-Bezy, M., Lawrence, J.B., *et al.* (2013). A long noncoding RNA mediates both activation and repression of immune response genes. Science *341*, 789-792.

Cawley, S., Bekiranov, S., Ng, H.H., Kapranov, P., Sekinger, E.A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A.J., *et al.* (2004). Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell *116*, 499-509.

Chen, J., Hu, Z., Phatak, M., Reichard, J., Freudenberg, J.M., Sivaganesan, S., and Medvedovic, M. (2013). Genome-wide signatures of transcription factor activity: connecting transcription factors, disease, and small molecules. PLoS Comput Biol *9*, e1003198.

Chen, J., Lee, S., Zhou, G., and Wang, S.M. (2002a). High-throughput GLGI procedure for converting a large number of serial analysis of gene expression tag sequences into 3' complementary DNAs. Genes Chromosomes Cancer *33*, 252-261.

Chen, J., Sun, M., Lee, S., Zhou, G., Rowley, J.D., and Wang, S.M. (2002b). Identifying novel transcripts and novel genes in the human genome by using novel SAGE tags. Proc Natl Acad Sci U S A *99*, 12257-12262.

Chen, R., Mias, G.I., Li-Pook-Than, J., Jiang, L., Lam, H.Y., Miriami, E., Karczewski, K.J., Hariharan, M., Dewey, F.E., Cheng, Y., *et al.* (2012). Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell *148*, 1293-1307.

Cherbas, L., Willingham, A., Zhang, D., Yang, L., Zou, Y., Eads, B.D., Carlson, J.W., Landolin, J.M., Kapranov, P., Dumais, J., *et al.* (2011). The transcriptional diversity of 25 Drosophila cell lines. Genome Res *21*, 301-314.

Chien, S., Reiter, L.T., Bier, E., and Gribskov, M. (2002). Homophila: human disease gene cognates in Drosophila. Nucleic Acids Res *30*, 149-151.

Chintapalli, V.R., Terhzaz, S., Wang, J., Al Bratty, M., Watson, D.G., Herzyk, P., Davies, S.A., and Dow, J.A. (2012). Functional correlates of positional and gender-specific renal asymmetry in Drosophila. PLoS One *7*, e32577.

Chintapalli, V.R., Wang, J., and Dow, J.A. (2007). Using FlyAtlas to identify better Drosophila melanogaster models of human disease. Nat Genet *39*, 715-720.

Clark, A.G., Eisen, M.B., Smith, D.R., Bergman, C.M., Oliver, B., Markow, T.A., Kaufman, T.C., Kellis, M., Gelbart, W., Iyer, V.N., *et al.* (2007). Evolution of genes and genomes on the Drosophila phylogeny. Nature *450*, 203-218.

Cui, P., Lin, Q., Ding, F., Xin, C., Gong, W., Zhang, L., Geng, J., Zhang, B., Yu, X., Yang, J., *et al.* (2010). A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing. Genomics *96*, 259-265.

Cui, X., and Loraine, A.E. (2009). Consistency analysis of redundant probe sets on affymetrix three-prime expression arrays and applications to differential mRNA processing. PLoS One *4*, e4229.

Daines, B., Wang, H., Wang, L., Li, Y., Han, Y., Emmert, D., Gelbart, W., Wang, X., Li, W., Gibbs, R., *et al.* (2011). The Drosophila melanogaster transcriptome by paired-end RNA sequencing. Genome Res *21*, 315-324.

Datson, N.A., van der Perk-de Jong, J., van den Berg, M.P., de Kloet, E.R., and Vreugdenhil, E. (1999). MicroSAGE: a modified procedure for serial analysis of gene expression in limited amounts of tissue. Nucleic Acids Res *27*, 1300-1307.

Davey, M.W., Graham, N.S., Vanholme, B., Swennen, R., May, S.T., and Keulemans, J. (2009). Heterologous oligonucleotide microarrays for transcriptomics in a non-model species; a proof-of-concept study of drought stress in Musa. BMC Genomics *10*, 436.

Davies, S.A., Cabrero, P., Povsic, M., Johnston, N.R., Terhzaz, S., and Dow, J.A. (2013). Signaling by Drosophila capa neuropeptides. Gen Comp Endocrinol *188*, 60-66.

Davies, S.A., Goodwin, S.F., Kelly, D.C., Wang, Z., Sozen, M.A., Kaiser, K., and Dow, J.A. (1996). Analysis and inactivation of vha55, the gene encoding the vacuolar ATPase B-subunit in Drosophila melanogaster reveals a larval lethal phenotype. J Biol Chem *271*, 30677-30684.

Davies, S.A., Overend, G., Sebastian, S., Cundall, M., Cabrero, P., Dow, J.A., and Terhzaz, S. (2012). Immune and stress response 'cross-talk' in the Drosophila Malpighian tubule. J Insect Physiol *58*, 488-497.

Davies, S.A., and Terhzaz, S. (2009). Organellar calcium signalling mechanisms in Drosophila epithelial function. J Exp Biol *212*, 387-400.

Day, J.P., Wan, S., Allan, A.K., Kean, L., Davies, S.A., Gray, J.V., and Dow, J.A. (2008). Identification of two partners from the bacterial Kef exchanger family for the apical plasma membrane V-ATPase of Metazoa. J Cell Sci *121*, 2612-2619.

DeRisi, J., Penland, L., Brown, P.O., Bittner, M.L., Meltzer, P.S., Ray, M., Chen, Y., Su, Y.A., and Trent, J.M. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet *14*, 457-460.

Dillies, M.A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., *et al.* (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform *14*, 671-683.

Dow, J.A. (2007). Integrative physiology, functional genomics and the phenotype gap: a guide for comparative physiologists. J Exp Biol *210*, 1632-1640.

Dow, J.A. (2012). The versatile stellate cell - more than just a space-filler. J Insect Physiol *58*, 467-472.

Dow, J.A., and Davies, S.A. (2003). Integrative physiology and functional genomics of epithelial function in a genetic model organism. Physiol Rev *83*, 687-729.

Dow, J.A., Maddrell, S.H., Gortz, A., Skaer, N.J., Brogan, S., and Kaiser, K. (1994). The malpighian tubules of Drosophila melanogaster: a novel phenotype for studies of fluid secretion and its control. J Exp Biol *197*, 421-428.

Dow, J.A., and Romero, M.F. (2010). Drosophila provides rapid modeling of renal development, function, and disease. Am J Physiol Renal Physiol *299*, F1237-1244.

Downey, T. (2006). Analysis of a multifactor microarray study using Partek genomics solution. Methods Enzymol *411*, 256-270.

Drysdale, R. (2008). FlyBase : a database for the Drosophila research community. Methods Mol Biol *420*, 45-59.

Dube, K., McDonald, D.G., and O'Donnell, M.J. (2000). Calcium transport by isolated anterior and posterior Malpighian tubules of Drosophila melanogaster: roles of sequestration and secretion. J Insect Physiol *46*, 1449-1460.

Duffy, J.B. (2002). GAL4 system in Drosophila: a fly geneticist's Swiss army knife. Genesis *34*, 1-15.

Eddy, S.R. (2001). Non-coding RNA genes and the modern RNA world. Nat Rev Genet 2, 919-929.

Eilers, M., Patel, A.B., Liu, W., and Smith, S.O. (2002). Comparison of helix interactions in membrane and soluble alpha-bundle proteins. Biophys J *82*, 2720-2736.

Elliott, D.A., and Brand, A.H. (2008). The GAL4 system : a versatile system for the expression of genes. Methods Mol Biol *420*, 79-95.

Estrada, B., and Michelson, A.M. (2008). A genomic approach to myoblast fusion in Drosophila. Methods Mol Biol *475*, 299-314.

Evans, J.M., Allan, A.K., Davies, S.A., and Dow, J.A. (2005). Sulphonylurea sensitivity and enriched expression implicate inward rectifier K+ channels in Drosophila melanogaster renal function. J Exp Biol *208*, 3771-3783.

Fan, H., Xiao, Y., Yang, Y., Xia, W., Mason, A.S., Xia, Z., Qiao, F., Zhao, S., and Tang, H. (2013). RNA-Seq analysis of Cocos nucifera: transcriptome sequencing and de novo assembly for subsequent functional genomics approaches. PLoS One *8*, e59997.

Fickett, J.W. (1996). Finding genes by computer: the state of the art. Trends Genet *12*, 316-320.

Franzen, O., Jerlstrom-Hultqvist, J., Einarsson, E., Ankarklev, J., Ferella, M., Andersson, B., and Svard, S.G. (2013). Transcriptome profiling of Giardia intestinalis using strand-specific RNA-seq. PLoS Comput Biol *9*, e1003000.

Fu, X., Fu, N., Guo, S., Yan, Z., Xu, Y., Hu, H., Menzel, C., Chen, W., Li, Y., Zeng, R., *et al.* (2009). Estimating accuracy of RNA-Seq and microarrays with proteomics. BMC Genomics *10*, 161.

Garber, M., Grabherr, M.G., Guttman, M., and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods *8*, 469-477.

Garrigan, D., Kingan, S.B., Geneva, A.J., Andolfatto, P., Clark, A.G., Thornton, K.R., and Presgraves, D.C. (2012). Genome sequencing reveals complex speciation in the Drosophila simulans clade. Genome Res *22*, 1499-1511.

Glücksmann-Kuis and Schneider (1995). Polycystic kidney disease: the complete structure of the PKD1 gene and its protein. . Cell *81*, 289-298.

Gorke, B., and Vogel, J. (2008). Noncoding RNA control of the making and breaking of sugars. Genes Dev 22, 2914-2925.

Graham, N.S., Clutterbuck, A.L., James, N., Lea, R.G., Mobasheri, A., Broadley, M.R., and May, S.T. (2010). Equine transcriptome quantification using human GeneChip arrays can be improved using genomic DNA hybridisation and probe selection. Vet J *186*, 323-327.

Graham, N.S., May, S.T., Daniel, Z.C., Emmerson, Z.F., Brameld, J.M., and Parr, T. (2011). Use of the Affymetrix Human GeneChip array and genomic DNA hybridisation probe selection to study ovine transcriptomes. Animal *5*, 861-866.

Graveley, B.R. (2008). Molecular biology: power sequencing. Nature *453*, 1197-1198.

Graveley, B.R., Brooks, A.N., Carlson, J.W., Duff, M.O., Landolin, J.M., Yang, L., Artieri, C.G., van Baren, M.J., Boley, N., Booth, B.W., *et al.* (2011). The developmental transcriptome of Drosophila melanogaster. Nature *471*, 473-479.

Gros, F. (2003). From the messenger RNA saga to the transcriptome era. C R Biol *326*, 893-900.

Gu, L., Xu, D., You, T., Li, X., Yao, S., Chen, S., Zhao, J., Lan, H., and Zhang, F. (2011). Analysis of gene expression by ESTs from suppression subtractive hybridization library in Chenopodium album L. under salt stress. Mol Biol Rep *38*, 5285-5295.

Guida, A., Lindstadt, C., Maguire, S.L., Ding, C., Higgins, D.G., Corton, N.J., Berriman, M., and Butler, G. (2011). Using RNA-seq to determine the transcriptional landscape and the hypoxic response of the pathogenic yeast Candida parapsilosis. BMC Genomics *12*, 628.

Guo, Y. (1996). Cloning, characterisation and site-selected mutagenesis of genes encoding V-ATPase in *Drosophila (phD thesis)*. Universiy of Glasgow

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., *et al.* (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223-227.

Guttman, M., Garber, M., Levin, J.Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M.J., Gnirke, A., Nusbaum, C., *et al.* (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol *28*, 503-510.

Haley, B., Tang, G., and Zamore, P.D. (2003). In vitro analysis of RNA interference in Drosophila melanogaster. Methods *30*, 330-336.

Hammond, J.P., Bowen, H.C., White, P.J., Mills, V., Pyke, K.A., Baker, A.J., Whiting, S.N., May, S.T., and Broadley, M.R. (2006). A comparison of the Thlaspi caerulescens and Thlaspi arvense shoot transcriptomes. New Phytol *170*, 239-260.

Hammond, J.P., Broadley, M.R., Craigon, D.J., Higgins, J., Emmerson, Z.F., Townsend, H.J., White, P.J., and May, S.T. (2005). Using genomic DNA-based probe-selection to improve the sensitivity of high-density oligonucleotide arrays when applied to heterologous species. Plant Methods *1*, 10.

Hardy, S., Legagneux, V., Audic, Y., and Paillard, L. (2010). Reverse genetics in eukaryotes. Biol Cell *102*, 561-580.

Harvey, W.R., and Wieczorek, H. (1997). Animal plasma membrane energization by chemiosmotic H+ V-ATPases. J Exp Biol *200*, 203-216.

Hawkins, R.D., Hon, G.C., and Ren, B. (2010). Next-generation genomics: an integrative approach. Nat Rev Genet *11*, 476-486.

Hiller, M., Findeiss, S., Lein, S., Marz, M., Nickel, C., Rose, D., Schulz, C., Backofen, R., Prohaska, S.J., Reuter, G., *et al.* (2009). Conserved introns reveal novel transcripts in Drosophila melanogaster. Genome Res *19*, 1289-1300.

Hitzemann, R., Bottomly, D., Darakjian, P., Walter, N., Iancu, O., Searles, R., Wilmot, B., and McWeeney, S. (2013). Genes, behavior and next-generation RNA sequencing. Genes Brain Behav *12*, 1-12.

Hong-Bin Zhang, C.W. (2001). BAC as tools for genome sequencing. Plant Physiology and Biochemistry *39*, 195-209.

Horton, P., Park, K.J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C.J., and Nakai, K. (2007). WoLF PSORT: protein localization predictor. Nucleic Acids Res *35*, W585-587.

Hu, M., and Polyak, K. (2006). Serial analysis of gene expression. Nat Protoc *1*, 1743-1760.

Hughes, J., Ward, C.J., Peral, B., Aspinwall, R., Clark, K., San Millan, J.L., Gamble, V., and Harris, P.C. (1995). The polycystic kidney disease 1 (PKD1) gene encodes a novel protein with multiple cell recognition domains. Nat Genet *10*, 151-160.

Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B., and Speed, T.P. (2003). Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res *31*, e15.

Issac, B., and Raghava, G.P. (2004). EGPred: prediction of eukaryotic genes using ab initio methods after combining with sequence similarity approaches. Genome Res *14*, 1756-1766.

Ji, W., Zhou, W., Gregg, K., Yu, N., and Davis, S. (2004). A method for cross-species gene expression analysis with high-density oligonucleotide arrays. Nucleic Acids Res *32*, e93.

Jiang, Z.F., Croshaw, D.A., Wang, Y., Hey, J., and Machado, C.A. (2011). Enrichment of mRNA-like noncoding RNAs in the divergence of Drosophila males. Mol Biol Evol *28*, 1339-1348.

Joseph Sambrook, D.R. (2001). Molecular Cloning: A Laboratory Manual (Third Edition), 3 edn (CSH Press).

Kampa, D., Cheng, J., Kapranov, P., Yamanaka, M., Brubaker, S., Cawley, S., Drenkow, J., Piccolboni, A., Bekiranov, S., Helt, G., *et al.* (2004). Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. Genome Res *14*, 331-342.

Kapranov, P., Cawley, S.E., Drenkow, J., Bekiranov, S., Strausberg, R.L., Fodor, S.P., and Gingeras, T.R. (2002). Large-scale transcriptional activity in chromosomes 21 and 22. Science *296*, 916-919.

Kapranov, P., Cheng, J., Dike, S., Nix, D.A., Duttagupta, R., Willingham, A.T., Stadler, P.F., Hertel, J., Hackermuller, J., Hofacker, I.L., *et al.* (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science *316*, 1484-1488.

Kaun, K.R., Azanchi, R., Maung, Z., Hirsh, J., and Heberlein, U. (2011). A Drosophila model for alcohol reward. Nat Neurosci *14*, 612-619.

Keller, A., Backes, C., Leidinger, P., Kefer, N., Boisguerin, V., Barbacioru, C., Vogel, B., Matzas, M., Huwer, H., Katus, H.A., *et al.* (2011). Next-generation sequencing identifies novel microRNAs in peripheral blood of lung cancer patients. Mol Biosyst *7*, 3187-3199.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature *423*, 241-254.

Kerr, M., Davies, S.A., and Dow, J.A. (2004). Cell-specific manipulation of second messengers; a toolbox for integrative physiology in Drosophila. Curr Biol *14*, 1468-1474.

Kitano, H. (2002). Systems biology: a brief overview. Science *295*, 1662-1664.

Kogenaru, S., Qing, Y., Guo, Y., and Wang, N. (2012). RNA-seq and microarray complement each other in transcriptome profiling. BMC Genomics *13*, 629.

Kondo, T., Inagaki, S., Yasuda, K., and Kageyama, Y. (2006). Rapid construction of Drosophila RNAi transgenes using pRISE, a P-element-mediated transformation vector exploiting an in vitro recombination system. Genes Genet Syst *81*, 129-134.

Kong, L., Zhang, Y., Ye, Z.Q., Liu, X.Q., Zhao, S.Q., Wei, L., and Gao, G. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res *35*, W345-349.

Kullberg, M., Hallstrom, B., Arnason, U., and Janke, A. (2007). Expressed sequence tags as a tool for phylogenetic analysis of placental mammal evolution. PLoS One *2*, e775.

Lam, G., and Thummel, C.S. (2000). Inducible expression of double-stranded RNA directs specific genetic interference in Drosophila. Curr Biol *10*, 957-963.

Lander, E.S., and Doyle, M. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10,* R25.

Lashkari, D.A., DeRisi, J.L., McCusker, J.H., Namath, A.F., Gentile, C., Hwang, S.Y., Brown, P.O., and Davis, R.W. (1997). Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc Natl Acad Sci U S A *94,* 13057-13062.

Lebovitz, R.M., Takeyasu, K., and Fambrough, D.M. (1989). Molecular characterization and expression of the (Na+ + K+)-ATPase alpha-subunit in Drosophila melanogaster. EMBO J *8*, 193-202.

Lee, Y.S., and Carthew, R.W. (2003). Making a better RNAi vector for Drosophila: use of intron spacers. Methods *30*, 322-329.

Leimena, M.M., Wels, M., Bongers, R.S., Smid, E.J., Zoetendal, E.G., and Kleerebezem, M. (2012). Comparative analysis of Lactobacillus plantarum WCFS1 transcriptomes by using DNA microarray and next-generation sequencing technologies. Appl Environ Microbiol *78*, 4141-4148.

Levin, J.Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D.A., Friedman, N., Gnirke, A., and Regev, A. (2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat Methods *7*, 709-715.

Li, C., and Hung Wong, W. (2001). Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biol *2*, RESEARCH0032.

Li, C., and Wong, W.H. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci U S A *98*, 31-36.

Liu, F., Jenssen, T.K., Trimarchi, J., Punzo, C., Cepko, C.L., Ohno-Machado, L., Hovig, E., and Kuo, W.P. (2007). Comparison of hybridization-based and sequencing-based gene expression technologies on biological replicates. BMC Genomics *8*, 153.

Liu, S., Lin, L., Jiang, P., Wang, D., and Xing, Y. (2011a). A comparison of RNA-Seq and high-density exon array for detecting differential gene expression between closely related species. Nucleic Acids Res *39*, 578-588.

Liu, W., Zhao, Y., Cui, P., Lin, Q., Ding, F., Xin, C., Tan, X., Song, S., Yu, J., and Hu, S. (2011b). Thousands of Novel Transcripts Identified in Mouse Cerebrum, Testis, and ES Cells Based on ribo-minus RNA Sequencing. Front Genet *2*, 93.

Macdonald, S.J., and Long, A.D. (2006). Fine scale structural variants distinguish the genomes of Drosophila melanogaster and D. pseudoobscura. Genome Biol *7*, R67.

Malone, J.H., and Oliver, B. (2011). Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biol *9*, 34.

Manak, J.R., Dike, S., Sementchenko, V., Kapranov, P., Biemar, F., Long, J., Cheng, J., Bell, I., Ghosh, S., Piccolboni, A., *et al.* (2006). Biological function of unannotated transcription during the early development of Drosophila melanogaster. Nat Genet *38*, 1151-1158.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res *18*, 1509-1517.

Martin, K.J., and Pardee, A.B. (2000). Identifying expressed genes. Proc Natl Acad Sci U S A *97*, 3789-3791.

Martinelli, G., Iacobucci, I., Papayannidis, C., and Soverini, S. (2009). New targets for Ph+ leukaemia therapy. Best Pract Res Clin Haematol *22*, 445-454.

Mathe, C., Sagot, M.F., Schiex, T., and Rouze, P. (2002). Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res *30*, 4103-4117.

Matsumura, H., Bin Nasir, K.H., Yoshida, K., Ito, A., Kahl, G., Kruger, D.H., and Terauchi, R. (2006). SuperSAGE array: the direct use of 26-base-pair transcript tags in oligonucleotide arrays. Nat Methods *3*, 469-474.

Matsumura, H., Kruger, D.H., Kahl, G., and Terauchi, R. (2008). SuperSAGE: a modern platform for genome-wide quantitative transcript profiling. Curr Pharm Biotechnol *9*, 368-374.

Matsumura, H., Reich, S., Ito, A., Saitoh, H., Kamoun, S., Winter, P., Kahl, G., Reuter, M., Kruger, D.H., and Terauchi, R. (2003). Gene expression analysis of plant host-pathogen interactions by SuperSAGE. Proc Natl Acad Sci U S A *100*, 15718-15723.

Matsumura, H., Yoshida, K., Luo, S., Kimura, E., Fujibe, T., Albertyn, Z., Barrero, R.A., Kruger, D.H., Kahl, G., Schroth, G.P., *et al.* (2010). High-throughput SuperSAGE for digital gene expression analysis of multiple samples using next generation sequencing. PLoS One *5*, e12010.

Mattick, J.S. (2009). The genetic signatures of noncoding RNAs. PLoS Genet *5*, e1000459.

McQuilton, P., St Pierre, S.E., and Thurmond, J. (2012). FlyBase 101--the basics of navigating FlyBase. Nucleic Acids Res *40*, D706-714.

Meldolesi, J. (2004). The development of Ca2+ indicators: a breakthrough in pharmacological research. Trends Pharmacol Sci *25*, 172-174.

Menssen, A., and Hermeking, H. (2002). Characterization of the c-MYC-regulated transcriptome by SAGE: identification and analysis of c-MYC target genes. Proc Natl Acad Sci U S A *99*, 6274-6279.

Metzker, M.L. (2005). Emerging technologies in DNA sequencing. Genome Res *15*, 1767-1776.

Michaelson, J.J., Shi, Y., Gujral, M., Zheng, H., Malhotra, D., Jin, X., Jian, M., Liu, G., Greer, D., Bhandari, A., *et al.* (2012). Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell *151*, 1431-1442.

Mockler, T.C., Chan, S., Sundaresan, A., Chen, H., Jacobsen, S.E., and Ecker, J.R. (2005). Applications of DNA tiling arrays for whole-genome analysis. Genomics *85*, 1-15.

Mooney, M., Bond, J., Monks, N., Eugster, E., Cherba, D., Berlinski, P., Kamerling, S., Marotti, K., Simpson, H., Rusk, T., *et al.* (2013). Comparative RNA-Seq and microarray analysis of gene expression changes in B-cell lymphomas of Canis familiaris. PLoS One *8*, e61088.

Morin, R.D., O'Connor, M.D., Griffith, M., Kuchenbauer, F., Delaney, A., Prabhu, A.L., Zhao, Y., McDonald, H., Zeng, T., Hirst, M., *et al.* (2008). Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome Res *18*, 610-621.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods *5*, 621-628.

Nagai, T., Ibata, K., Park, E.S., Kubota, M., Mikoshiba, K., and Miyawaki, A. (2002). A variant of yellow fluorescent protein with fast and efficient maturation for cell-biological applications. Nat Biotechnol *20*, 87-90.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science *320*, 1344-1349.

Nakai, K., and Horton, P. (1999). PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci *24*, 34-36.

Nishiyama, T., Fujita, T., Shin, I.T., Seki, M., Nishide, H., Uchiyama, I., Kamiya, A., Carninci, P., Hayashizaki, Y., Shinozaki, K., *et al.* (2003). Comparative genomics of Physcomitrella patens gametophytic transcriptome and Arabidopsis thaliana: implication for land plant evolution. Proc Natl Acad Sci U S A *100*, 8007-8012.

Nookaew, I., Papini, M., Pornputtapong, N., Scalcinati, G., Fagerberg, L., Uhlen, M., and Nielsen, J. (2012). A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. Nucleic Acids Res *40*, 10084-10097.

Northrup, D.L., and Zhao, K. (2011). Application of ChIP-Seq and related techniques to the study of immune function. Immunity *34*, 830-842.

Nuzhdin SV, W.M., Harmon KL, McIntyre LM (2004). Common pattern of evolution of gene expression level and protein sequence in Drosophila. Mol Biol Evol 2*1*, 1308-1317.

Okamura, H., Yasuhara, J.C., Fambrough, D.M., and Takeyasu, K. (2003). P-type ATPases in Caenorhabditis and Drosophila: implications for evolution of the P-type ATPase subunit families with special reference to the Na,K-ATPase and H,K-ATPase subgroup. J Membr Biol *191*, 13-24.

Palmieri, N., Nolte, V., Suvorov, A., Kosiol, C., and Schlotterer, C. (2012). Evaluation of different reference based annotation strategies using RNA-Seq - a case study in Drososphila pseudoobscura. PLoS One *7*, e46415.

Parkhomchuk, D., Borodina, T., Amstislavskiy, V., Banaru, M., Hallen, L., Krobitsch, S., Lehrach, H., and Soldatov, A. (2009). Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res *37*, e123.

Pavlidis, P. (2003). Using ANOVA for gene selection from microarray studies of the nervous system. Methods *31*, 282-289.

Pennisi, E. (2000). Stealth genome rocks rice researchers. Science *288*, 239-241.

Peters, B.A., St Croix, B., Sjoblom, T., Cummins, J.M., Silliman, N., Ptak, J., Saha, S., Kinzler, K.W., Hatzis, C., and Velculescu, V.E. (2007). Large-scale identification of novel transcripts in the human genome. Genome Res *17*, 287-292.

Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., Veyrieras, J.B., Stephens, M., Gilad, Y., and Pritchard, J.K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature *464*, 768-772.

Polyak, K., and Riggins, G.J. (2001). Gene discovery using the serial analysis of gene expression technique: implications for cancer research. J Clin Oncol *19*, 2948-2958.

Ponting, C.P., Hofmann, K., and Bork, P. (1999). A latrophilin/CL-1-like GPS domain in polycystin-1. Curr Biol *9*, R585-588.

Ponting, C.P., Oliver, P.L., and Reik, W. (2009). Evolution and functions of long noncoding RNAs. Cell *136*, 629-641.

Quackenbush, J. (2001). Computational analysis of microarray data. Nat Rev Genet *2*, 418-427.

Quackenbush, J. (2002). Microarray data normalization and transformation. Nat Genet *32 Suppl*, 496-501.

Raghavachari, N., Barb, J., Yang, Y., Liu, P., Woodhouse, K., Levy, D., O'Donnell, C.J., Munson, P.J., and Kato, G.J. (2012). A systematic comparison and evaluation of high density exon arrays and RNA-seq technology used to unravel the peripheral blood transcriptome of sickle cell disease. BMC Med Genomics *5*, 28.

Ranz, J.M., Castillo-Davis, C.I., Meiklejohn, C.D., and Hartl, D.L. (2003). Sex-dependent gene expression and evolution of the Drosophila transcriptome. Science *300*, 1742-1745.

Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., *et al.* (2006). Global variation in copy number in the human genome. Nature *444*, 444-454.

Reese, M.G., Hartzell, G., Harris, N.L., Ohler, U., Abril, J.F., and Lewis, S.E. (2000). Genome annotation assessment in Drosophila melanogaster. Genome Res *10*, 483-501.

Reiter, L.T., Potocki, L., Chien, S., Gribskov, M., and Bier, E. (2001). A systematic analysis of human disease-associated gene sequences in Drosophila melanogaster. Genome Res *11*, 1114-1125.

Richards, S., Liu, Y., Bettencourt, B.R., Hradecky, P., Letovsky, S., Nielsen, R., Thornton, K., Hubisz, M.J., Chen, R., Meisel, R.P., *et al.* (2005). Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. Genome Res *15*, 1-18.

Riggins, G.J., and Strausberg, R.L. (2001). Genome and genetic resources from the Cancer Genome Anatomy Project. Hum Mol Genet *10*, 663-667.

Roberts, A., Pimentel, H., Trapnell, C., and Pachter, L. (2011). Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics *27*, 2325-2329.

Rogic, S., Mackworth, A.K., and Ouellette, F.B. (2001). Evaluation of gene-finding programs on mammalian sequences. Genome Res *11*, 817-832.

Rosay, P., Davies, S.A., Yu, Y., Sozen, M.A., Kaiser, K., and Dow, J.A. (1997). Cell-type specific calcium signalling in a Drosophila epithelium. J Cell Sci *110 (Pt 15)*, 1683-1692.

Roy, S., Ernst, J., Kharchenko, P.V., Kheradpour, P., Negre, N., Eaton, M.L., Landolin, J.M., Bristow, C.A., Ma, L., Lin, M.F., *et al.* (2010). Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science *330*, 1787-1797.

Rubin, C.J., Zody, M.C., Eriksson, J., Meadows, J.R., Sherwood, E., Webster, M.T., Jiang, L., Ingman, M., Sharpe, T., Ka, S., *et al.* (2010). Whole-genome resequencing reveals loci under selection during chicken domestication. Nature *464*, 587-591.

Rubin, G.M., Hong, L., Brokstein, P., Evans-Holm, M., Frise, E., Stapleton, M., and Harvey, D.A. (2000a). A Drosophila complementary DNA resource. Science *287*, 2222-2224.

Rubin, G.M., and Spradling, A.C. (1982). Genetic transformation of Drosophila with transposable element vectors. Science *218*, 348-353.

Rubin, G.M., Yandell, M.D., Wortman, J.R., Gabor Miklos, G.L., Nelson, C.R., Hariharan, I.K., Fortini, M.E., Li, P.W., Apweiler, R., Fleischmann, W., *et al.* (2000b). Comparative genomics of the eukaryotes. Science *287*, 2204-2215.

Saha, S., Sparks, A.B., Rago, C., Akmaev, V., Wang, C.J., Vogelstein, B., Kinzler, K.W., and Velculescu, V.E. (2002). Using the transcriptome to annotate the genome. Nat Biotechnol *20,* 508-512.

Sasidharan, R., Agarwal, A., Rozowsky, J., and Gerstein, M. (2009). An approach to comparing tiling array and high throughput sequencing technologies for genomic transcript mapping. BMC Res Notes *2*, 150.

Schena, M., Heller, R.A., Theriault, T.P., Konrad, K., Lachenmeier, E., and Davis, R.W. (1998). Microarrays: biotechnology's discovery platform for functional genomics. Trends Biotechnol *16*, 301-306.

Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science *270*, 467-470.

Service, R.F. (2006). Gene sequencing. The race for the $1000 genome. Science *311*, 1544-1546.

Shendure, J. (2008). The beginning of the end for microarrays? Nat Methods *5*, 585-587.

Smith, S.T., Wickramasinghe, P., Olson, A., Loukinov, D., Lin, L., Deng, J., Xiong, Y., Rux, J., Sachidanandam, R., Sun, H., *et al.* (2009). Genome wide ChIP-chip analyses reveal important roles for CTCF in Drosophila genome organization. Dev Biol *328*, 518-528.

Soshnev, A.A., Ishimoto, H., McAllister, B.F., Li, X., Wehling, M.D., Kitamoto, T., and Geyer, P.K. (2011). A conserved long noncoding RNA affects sleep behavior in Drosophila. Genetics *189*, 455-468.

Southall, T.D., Terhzaz, S., Cabrero, P., Chintapalli, V.R., Evans, J.M., Dow, J.A., and Davies, S.A. (2006). Novel subcellular locations and functions for secretory pathway Ca2+/Mn2+-ATPases. Physiol Genomics *26*, 35-45.

Sozen, M.A., Armstrong, J.D., Yang, M., Kaiser, K., and Dow, J.A. (1997). Functional domains are specified to single-cell resolution in a Drosophila epithelium. Proc Natl Acad Sci U S A *94*, 5207-5212.

Spradling, A.C. (2006). Learning the common language of genetics. Genetics *174*, 1-3.

Spradling, A.C., and Rubin, G.M. (1982). Transposition of cloned P elements into Drosophila germ line chromosomes. Science *218*, 341-347.

Stapleton, M., Liao, G., Brokstein, P., Hong, L., Carninci, P., Shiraki, T., Hayashizaki, Y., Champe, M., Pacleb, J., Wan, K., *et al.* (2002). The Drosophila gene collection: identification of putative full-length cDNAs for 70% of D. melanogaster genes. Genome Res *12*, 1294-1300.

Stark, A., Lin, M.F., Kheradpour, P., Pedersen, J.S., Parts, L., Carlson, J.W., Crosby, M.A., Rasmussen, M.D., Roy, S., Deoras, A.N., *et al.* (2007). Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature *450*, 219-232.

Storey, J.D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. Ann Stat *31*, 2013-2035.

Stratton, M. (2008). Genome resequencing and genetic variation. Nat Biotechnol *26*, 65-66.

t Hoen, P.A., Ariyurek, Y., Thygesen, H.H., Vreugdenhil, E., Vossen, R.H., de Menezes, R.X., Boer, J.M., van Ommen, G.J., and den Dunnen, J.T. (2008). Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. Nucleic Acids Res *36*, e141.

Tang, T., Francois, N., Glatigny, A., Agier, N., Mucchielli, M.H., Aggerbeck, L., and Delacroix, H. (2007). Expression ratio evaluation in two-colour microarray experiments is significantly improved by correcting image misalignment. Bioinformatics *23*, 2686-2691.

Tariq, M.A., Kim, H.J., Jejelowo, O., and Pourmand, N. (2011). Whole-transcriptome RNAseq analysis from minute amount of total RNA. Nucleic Acids Res *39*, e120.

Terhzaz, S., Overend, G., Sebastian, S., Dow, J.A., and Davies, S.A. (2013). The D. melanogaster capa-1 neuropeptide activates renal NF-kB signaling. Peptides.

Torales, S.L., Rivarola, M., Pomponio, M.F., Gonzalez, S., Acuna, C.V., Fernandez, P., Lauenstein, D.L., Verga, A.R., Hopp, H.E., Paniego, N.B., *et al.* (2013). De novo assembly and characterization of leaf transcriptome for the development of functional molecular markers of the extremophile multipurpose tree species Prosopis alba. BMC Genomics *14*, 705.

Torrie, L.S., Radford, J.C., Southall, T.D., Kean, L., Dinsmore, A.J., Davies, S.A., and Dow, J.A. (2004). Resolution of the insect ouabain paradox. Proc Natl Acad Sci U S A *101*, 13689-13693.

Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol *31*, 46-53.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc *7*, 562-578.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol *28*, 511-515.

Tripathi, V., Ellis, J.D., Shen, Z., Song, D.Y., Pan, Q., Watt, A.T., Freier, S.M., Bennett, C.F., Sharma, A., Bubulya, P.A., *et al.* (2010). The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. Mol Cell *39*, 925-938.

Tublitz, N.J., and Truman, J.W. (1985a). Identification of neurones containing cardioacceleratory peptides (CAPs) in the ventral nerve cord of the tobacco hawkmoth, Manduca sexta. J Exp Biol *116*, 395-410.

Tublitz, N.J., and Truman, J.W. (1985b). Insect cardioactive peptides. I. Distribution and molecular characteristics of two cardioacceleratory peptides in the tobacco hawkmoth, Manduca sexta. J Exp Biol *114*, 365-379.

Tublitz, N.J., and Truman, J.W. (1985c). Insect cardioactive peptides. II. Neurohormonal control of heart activity by two cardioacceleratory peptides in the tobacco hawkmoth, Manduca sexta. J Exp Biol *114*, 381-395.

Van Adelsberg, J.S., and Frank, D. (1995). The PKD1 gene produces a developmentally regulated protein in mesenchyme and vasculature. Nat Med *1*, 359-364.

van Bakel, H., Nislow, C., Blencowe, B.J., and Hughes, T.R. (2010). Most "dark matter" transcripts are associated with known genes. PLoS Biol *8*, e1000371.

van den Berg, A., van der Leij, J., and Poppema, S. (1999). Serial analysis of gene expression: rapid RT-PCR analysis of unknown SAGE tags. Nucleic Acids Res *27*, e17.

van Iterson, M., t Hoen, P.A., Pedotti, P., Hooiveld, G.J., den Dunnen, J.T., van Ommen, G.J., Boer, J.M., and Menezes, R.X. (2009). Relative power and sample size analysis on gene expression profiling data. BMC Genomics *10*, 439.

Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. (1995). Serial analysis of gene expression. Science *270*, 484-487.

Venken, K.J., and Bellen, H.J. (2007). Transgenesis upgrades for Drosophila melanogaster. Development *134*, 3571-3584.

Verdun, R.E., Di Paolo, N., Urmenyi, T.P., Rondinelli, E., Frasch, A.C., and Sanchez, D.O. (1998). Gene discovery through expressed sequence Tag sequencing in Trypanosoma cruzi. Infect Immun *66*, 5393-5398.

Vidal, E.A., Moyano, T.C., Krouk, G., Katari, M.S., Tanurdzic, M., McCombie, W.R., Coruzzi, G.M., and Gutierrez, R.A. (2013). Integrated RNA-seq and sRNA-seq analysis identifies novel nitrate-responsive genes in Arabidopsis thaliana roots. BMC Genomics *14*, 701.

Wahl, M.B., Heinzmann, U., and Imai, K. (2005a). LongSAGE analysis revealed the presence of a large number of novel antisense genes in the mouse genome. Bioinformatics *21*, 1389-1392.

Wahl, M.B., Heinzmann, U., and Imai, K. (2005b). LongSAGE analysis significantly improves genome annotation: identifications of novel genes and alternative transcripts in the mouse. Bioinformatics *21*, 1393-1400.

Walker, S.J., Wang, Y., Grant, K.A., Chan, F., and Hellmann, G.M. (2006). Long versus short oligonucleotide microarrays for the study of gene expression in nonhuman primates. J Neurosci Methods *152*, 179-189.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470-476.

Wang, J., Kean, L., Yang, J., Allan, A.K., Davies, S.A., Herzyk, P., and Dow, J.A. (2004). Function-informed transcriptome analysis of Drosophila renal tubule. Genome Biol *5*, R69.

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

White, K.P., Rifkin, S.A., Hurban, P., and Hogness, D.S. (1999). Microarray analysis of Drosophila development during metamorphosis. Science *286*, 2179-2184.

Wieczorek, H., Huss, M., Merzendorfer, H., Reineke, S., Vitavska, O., and Zeiske, W. (2003). The insect plasma membrane H+ V-ATPase: intra-, inter-, and supramolecular aspects. J Bioenerg Biomembr *35*, 359-366.

Wilhelm, B.T., Marguerat, S., Goodhead, I., and Bahler, J. (2010). Defining transcribed regions using RNA-seq. Nat Protoc *5*, 255-266.

Wilhelm, B.T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C.J., Rogers, J., and Bahler, J. (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature *453*, 1239-1243.

Willis, D.K., Wang, J., Lindholm, J.R., Orth, A., and Goodman, W.G. (2010). Microarray analysis of juvenile hormone response in Drosophila melanogaster S2 cells. J Insect Sci *10*, 66.

Wilson, D.J. (2012). Insights from genomics into bacterial pathogen populations. PLoS Pathog *8*, e1002874.

Wilson, R.J., Goodman, J.L., and Strelets, V.B. (2008). FlyBase: integration and improvements to query tools. Nucleic Acids Res *36*, D588-593.

Wu, T.D., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics *26*, 873-881.

Yamamoto, M., Wakatsuki, T., Hada, A., and Ryo, A. (2001). Use of serial analysis of gene expression (SAGE) technology. J Immunol Methods *250*, 45-66.

Yang, J., McCart, C., Woods, D.J., Terhzaz, S., Greenwood, K.G., ffrench-Constant, R.H., and Dow, J.A. (2007). A Drosophila systems approach to xenobiotic metabolism. Physiol Genomics *30*, 223-231.

Yassour, M., Kaplan, T., Fraser, H.B., Levin, J.Z., Pfiffner, J., Adiconis, X., Schroth, G., Luo, S., Khrebtukova, I., Gnirke, A., *et al.* (2009). Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. Proc Natl Acad Sci U S A *106*, 3264-3269.

Yassour, M., Pfiffner, J., Levin, J.Z., Adiconis, X., Gnirke, A., Nusbaum, C., Thompson, D.A., Friedman, N., and Regev, A. (2010). Strand-specific RNA sequencing reveals extensive regulated long antisense transcripts that are conserved across yeast species. Genome Biol *11*, R87.

Young, R.S., Marques, A.C., Tibbit, C., Haerty, W., Bassett, A.R., Liu, J.L., and Ponting, C.P. (2012). Identification and properties of 1,119 candidate lincRNA loci in the Drosophila melanogaster genome. Genome Biol Evol *4*, 427-442.

Zhang, J., Finney, R.P., Clifford, R.J., Derr, L.K., and Buetow, K.H. (2005). Detecting false expression signals in high-density oligonucleotide arrays by an in silico approach. Genomics *85*, 297-308.

Zhou, X., Ren, L., Meng, Q., Li, Y., Yu, Y., and Yu, J. (2010). The next-generation sequencing technology and application. Protein Cell *1*, 520-536.

Zhu, F., Ding, H., and Zhu, B. (2013a). Transcriptional profiling of Drosophila S2 cells in early response to Drosophila C virus. Virol J *10*, 210.

Zhu, Q.H., Stephen, S., Taylor, J., Helliwell, C.A., and Wang, M.B. (2013b). Long noncoding RNAs responsive to Fusarium oxysporum infection in Arabidopsis thaliana. New Phytol.