IDENTIFICATION AND ANALYSIS OF PROGRAMMED CELL DEATH GENES IN *DROSOPHILA MELANOGASTER* AND HUMAN CANCER USING BIOINFORMATIC ANALYSIS OF GENE EXPRESSION DATA

by

ERIN DAEL PLEASANCE

B.Sc., The University of British Columbia, 2000

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

(Medical Genetics)

THE UNIVERSITY OF BRITISH COLUMBIA

December 2005

© Erin Dael Pleasance, 2005

Abstract

Programmed cell death (PCD), or cell suicide, encompasses multiple pathways including apoptosis and autophagy and is essential for development, cellular homeostasis, and prevention of cancer cell growth. I describe here the development and use of bioinformatic methods to identify and analyze genes involved in PCD, both in the model organism *Drosophila melanogaster* and in human cancer, by analysis of large-scale gene expression data. An approach was developed to correctly identify genes from serial analysis of gene expression (SAGE) data, distinguish the set of genes not accessible to the SAGE method, and determine the optimal set of enzymes for *Drosophila*, *C. elegans*, and human SAGE library construction. In *Drosophila* metamorphosis the salivary gland undergoes autophagic PCD, whereby cellular components are engulfed and degraded by cytoplasmic vacuoles, with additional hallmarks of apoptosis. This is an excellent model in which to study the genes involved in PCD. Transcriptional profiling of this tissue by expressed sequence tags (ESTs) and serial analysis of gene expression (SAGE) identified many genes differentially regulated prior to cell death, including genes known to be death regulators, genes in related pathways, genes of no known function, and potentially novel unannotated genes. The PCD-associated genes found in this analysis were then used to identify similar genes in the human genome that are differentially expressed in cancer, which have the potential to be involved in PCD and in oncogenesis. The pattern of genes expressed suggests a role for autophagy-associated processes in cancer progression. To examine this further, expression of the autophagy gene LC3 was examined in multiple cancer types, subtypes, and stages. LC3 expression is decreased significantly in several cancer types and also during cancer progression, suggesting a tissue- and stage-specific role for autophagy in regulating oncogenesis.

Table of Contents

Abstract ii
Table of Contents iii
List of Tables vi
List of Figures vii
List of Abbreviations viii
Acknowledgements ix
Chapter 1: Introduction 1
1.1 Thesis overview 2
1.2 Programmed cell death 2
1.2.1 Functions and types of programmed cell death 2
1.2.2 Apoptosis 3
1.2.3 Autophagy 7
1.2.4 Autophagic cell death 10
1.3 Cancer and programmed cell death 11
1.3.1 Molecular mechanisms of cancer 11
1.3.2 Apoptosis and cancer 13
1.3.3 Autophagy and cancer 15
1.4 Programmed cell death in the development of *Drosophila melanogaster* 17
1.4.1 *Drosophila* as a model for programmed cell death 17
1.4.2 Molecular mechanisms of programmed cell death in *Drosophila* 18
1.4.3 Programmed cell death in the *Drosophila* salivary gland 20
1.5 Gene expression analysis 21
1.5.1 Gene expression and mRNA levels 21
1.5.2 Microarrays 23
1.5.3 Expressed Sequence Tags 24
1.5.4 Serial Analysis of Gene Expression 25
1.5.4.1 SAGE experimental method 25
1.5.4.2 SAGE tag processing and statistics 30
1.5.4.3 SAGE tag-to-gene mapping 32
1.5.4.4 Advantages, disadvantages, and applications of SAGE 37
1.6 Thesis objectives and chapter summaries 39
Chapter 2: Assessment and analysis of SAGE for transcript identification 44
2.1 Introduction 45
2.2 Methods 47
2.2.1 ACEDB 47
2.2.2 Transcript construction 48
2.2.3 Evaluating tag-to-gene mapping accuracy 50
2.2.4 Choice of tagging and anchoring enzymes, tag length 51
2.2.5 Mapping experimentally derived tags 51
2.2.6 Human tag mapping with RefSeq 52
2.3 Results 52
2.3.1 Tag mapping using genomic sequence 52
2.3.2 Transcript construction and analysis 53

2.3.3 Tag-to-gene mapping 55
2.3.4 Analysis of transcriptomes with SAGE 58
2.3.5 Mapping experimentally derived tags 62
2.3.6 Assessment of human tag mapping with RefSeq 64
2.4 Discussion 66
2.4.1 Ambiguity in SAGE 66
2.4.2 Choice of tag length, anchoring, and tagging enzyme 68
2.4.3 Tags with no gene mapping 72
2.4.4 Limitations of conceptual transcripts 74
2.4.5 Conclusions 75
Chapter 3: SAGE and EST analysis of PCD in the *Drosophila* salivary gland 76
3.1 Introduction 77
3.2 Methods 78
3.2.1 EST sequence processing and clustering 78
3.2.2 BLAST analysis of ESTs 79
3.2.3 SAGE tag processing 79
3.2.4 SAGE tag-to-gene mapping 80
3.2.5 SAGE statistics 80
3.2.6 ACEDB database 81
3.3 Results 81
3.3.1 Processing and clustering of EST sequences 81
3.3.2 Genes identified by ESTs 83
3.3.3 Processing and differential expression of SAGE tags 86
3.3.4 Genes identified by SAGE tags 87
3.4 Discussion 93
3.4.1 Use of EST and SAGE to identify genes 93
3.4.2 Expression of PCD genes 95
3.4.3 Novel putative PCD genes 97
3.4.4 Conclusions 98
Chapter 4: Identification of PCD genes in *Drosophila* and cancer gene expression data 99
4.1 Introduction 100
4.2 Methods 101
4.2.1 *Drosophila* PCD expression data 101
4.2.2 Human cancer expression data 102
4.2.3 Drosophila-human orthology 103
4.2.4 False discovery rate 103
4.2.5 Expected proportion of differentially expressed and PCD genes 104
4.2.6 GO analysis 104
4.3 Results 105
4.3.1 *Drosophila* PCD and human cancer expression 105
4.3.2 Differentially expressed genes 106
4.3.3 Roles of genes in PCD and cancer 108
4.3.4 Functional categorization and overrepresentation 110
4.4 Discussion 113
4.4.1 Cross-species integration 113
4.4.2 Functions of genes in PCD and cancer 114

4.4.3 Conclusions 117
Chapter 5: The autophagy gene MAP1LC3B in cancer 118
5.1 Introduction 119
5.2 Methods 120
5.2.1 Processing of microarray expression data 120
5.2.2 Processing of SAGE expression data 121
5.2.3 Tag mapping and EST analysis of secondary LC3 tag 122
5.3 Results and discussion 123
5.3.1 Overview of LC3 regulation in multiple cancer and normal tissues 123
5.3.2 LC3 in cancer stages 126
5.3.3 LC3 in breast cancer progression 128
5.3.4 LC3 in cancer subtypes 131
5.3.5 LC3 and Beclin 1 expression 133
5.3.6 Conclusions 134
Chapter 6: Summary and Conclusions 136
6.1 Summary 137
6.2 Large-scale gene expression analysis with SAGE and other techniques 137
6.2.1 Measurement and comparison of gene expression 137
6.2.2 Novel genes and alternative transcripts 139
6.3 Genes and pathways associated with programmed cell death 140
6.3.1 In *Drosophila* 140
6.3.2 In cancer 142
6.4 Roles of autophagy in PCD and cancer 143
6.4.1 In *Drosophila* PCD 143
6.4.2 In cancer 144
6.5 Conclusions 145
References 146

List of Tables

Table 1.1 Acquired molecular changes in cancer progression. 13
Table 1.2 Variations in the SAGE procedure affecting SAGE tag length. 30
Table 2.1 SAGE tag mapping accuracy using conceptual transcripts compared to other available sequences. 58
Table 3.1 Summary of sequences matched by representative *Drosophila* ESTs. 84
Table 3.2 Cell death genes identified by ESTs. 84
Table 3.3 SAGE library tag counts and frequency distributions. 87
Table 3.4 Differentially expressed genes associated with salivary gland autophagic cell death in *Drosophila* SAGE libraries. 92
Table 4.1 Tissues and SAGE libraries from CGAP used for analysis of cancer expression. 106
Table 4.2 Functions overrepresented in the pcd-cancer set. 112
Table 5.1 Composition of normal and cancer samples shown in Figures 5.1 and 5.5. 126

List of Figures

Figure 1.1 Mammalian apoptosis pathway. 6
Figure 1.2 Mammalian autophagy pathway. 9
Figure 1.3 Programmed cell death during *Drosophila* development. 18
Figure 1.4 Ecdysone-triggered transcriptional cascade controlling salivary gland PCD. 21
Figure 1.5 Diagram of the SAGE procedure. 29
Figure 2.1 Conceptual transcript construction for SAGE tag mapping. 54
Figure 2.2 UTR size distributions in *Drosophila* and *C. elegans* 55
Figure 2.3 SAGE tag ambiguity varies with tag length. 60
Figure 2.4 Number of genes not resolvable by SAGE varies with anchoring enzyme. 61
Figure 2.5 Experimental SAGE tags mapped to conceptual transcripts. 63
Figure 2.6 SAGE tag mapping with RefSeq. 65
Figure 2.7 Creating shorter LongSAGE tags. 72
Figure 3.1 Processing and clustering of *Drosophila* 3' ESTs. 82
Figure 3.2 ESTs representing known and potentially novel genes. 85
Figure 3.3 Mapping of *Drosophila* SAGE tags to genes, ESTs, and genomic sequence. 90
Figure 3.4 Expression of known salivary gland cell death related genes in SAGE libraries. 91
Figure 4.1 Expression categories of human pcd-cancer genes. 108
Figure 4.2 Categorized roles of human genes in pcd-cancer set as found in the literature. 110
Figure 4.3 Functions associated with human genes in the pcd-cancer set. 112
Figure 5.1 LC3 expression in multiple normal and cancer tissues. 125
Figure 5.2 LC3 expression in liver cancer stages. 128
Figure 5.3 LC3 expression in breast cancer progression. 130
Figure 5.4 LC3 expression in lung and breast cancer subtypes. 132
Figure 5.5 Beclin 1 expression in multiple normal and cancer tissues. 134

List of Abbreviations

ACEDB	A *C. elegans* Database
BDGP	Berkeley *Drosophila* Genome Project
BLAST	Basic Local Alignment Search Tool
BLAT	BLAST-Like Alignment Tool
DCIS	Ductal Carcinoma In Situ
EST	Expressed Sequence Tag
FDR	False Discovery Rate
GFP	Green Fluorescent Protein
GO	Gene Ontology
IAP	Inhibitor of Apoptosis Protein
LC3	MAP1LC3B; microtubule-associated protein 1 light chain 3 beta
MGC	Mammalian Gene Collection
MMP	Matrix Metalloprotease
NCBI	National Center for Biotechnology Information
RT-PCR	Reverse Transcriptase Polymerase Chain Reaction
SAGE	Serial Analysis of Gene Expression
TIMP	Tissue Inhibitor of Metalloproteases
TOR	Target of Rapamycin
UCSC	University of California, Santa Cruz
UTR	Untranslated Region

Acknowledgements

I would like to thank my graduate supervisor Dr. Steven Jones for the opportunity to work in the Genome Sciences Centre's excellent and stimulating research environment, and for his support, encouragement, and guidance during my studies. In particular, I appreciate the opportunities to expand my knowledge and experience as a researcher, through attending conferences and exploring multiple different aspects of my project. I am grateful for salary and travel funding from the Natural Sciences and Engineering Research Council, the Canadian Institutes for Health Research, and the Michael Smith Foundation for Health Research. I would also like to thank others for their guidance and the time they have taken to work with me, particularly Drs. Sharon Gorski and Marco Marra, from whom I have learned a great deal. Thanks also to thesis committee members Drs. Dixie Mager and Aly Karsan for guidance and advice. I have enjoyed and appreciated the friendship and help of fellow bioinformatics graduate students, among them Angelique, Obi, Steve, and Monica. Thanks to Scott, Richard, and Mehrdad, along with others in the Gene Expression group, for help with aspects of SAGE analysis, and to members of the Programmed Cell Death group (Sugi, Doug, Ian, Claire, Melissa, Qadir and others), and the Gene Regulation Informatics group. I am grateful to have worked with so many outstanding people. Finally, I would like to thank my parents, Jane and Will, for their encouragement and my husband Steve for his constant love and support.

Chapter 1: Introduction

Portions of this chapter have been published.

Pleasance, E.D. and Jones, S.J.M. 2005. Evaluation of SAGE tags for transcriptome study. In: Wang, S.M., ed., *SAGE Technologies: Current Technologies and Applications*. Horizon Scientific Press, Norwich, UK.

1.1 Thesis overview

The aim of this thesis was to identify and analyze genes involved in programmed cell death (PCD) in *Drosophila melanogaster* and in human cancer systems, using a bioinformatics approach. Programmed cell death, referring to a number of processes of cellular suicide, is essential for the development of multicellular organisms and the prevention of unwanted cell growth that can lead to cancer, and underlies the pathology of some neurodegenerative and other diseases. Several decades of research have defined complex pathways that control cell death, but many of the genes involved are still unknown, especially in less well studied forms of PCD such as autophagic cell death. In addition, the role of PCD genes in the process of oncogenesis is both important and complex; understanding of the functions of PCD in cancer has the potential to impact cancer research, diagnosis, and treatment. I approached these research questions by first examining the process of PCD during the development of the model organism *Drosophila melanogaster*, and then applying these results, in combination with further studies, to investigate the role of PCD in cancer. The primary method employed was bioinformatic analysis of large-scale gene expression data, with additional use of genomic, expressed sequence, comparative genomic, and functional data. This comprehensive approach takes advantage of the large amount of high-throughput biological data that is being generated at an astounding rate, and applies it to the understanding of essential cellular pathways and their role in the complex causes of cancer.

1.2 Programmed cell death

1.2.1 Functions and types of programmed cell death

Programmed cell death is precisely initiated and executed cellular suicide, and requires the activation of specific genes and pathways. Thus, it is differentiated from non-programmed

necrotic cell death which occurs when cells lose homeostasis often due to environmental insult. PCD is essential in development, where the specific removal of cells acts together with cell proliferation to shape tissues such as digits and the neural tube, and to select the cells of the immune system (reviewed in Baehrecke 2002). It also acts in a protective manner by deleting damaged or diseased cells, such as those that are virally infected. However, excess PCD can be detrimental and play a role in disease, particularly neurodegenerative diseases (reviewed in Dlamini et al. 2004; Marino and Lopez-Otin 2004).

PCD was categorized by Schweichel and Merker (1973) into three types: type I, apoptotic; type II, autophagic; and type III, non-lysosomal programmed cell death. Little research has gone into the study of non-lysosomal programmed cell death, described morphologically as cellular disintegration with no involvement of the lysosomal pathways and no cellular condensation (Clarke 1990), and it will not be discussed further here. Both apoptosis and autophagic cell death are important in development and disease, though autophagic cell death has only recently become a major research focus. Indeed, the relationship between apoptosis, autophagic cell death and the morphologically related process of autophagy is unclear and thus these processes are described separately here.

1.2.2 Apoptosis

Classical apoptosis as described by Kerr et al (1972) is caspase-dependent and marked by chromatin and cytoplasm condensation, and DNA cleavage and fragmentation (reviewed in Baehrecke 2002; Zornig et al. 2001). The cell condenses to form apoptotic bodies which are cleared by phagocytes recognizing the phosphatidylserine which becomes exposed on the plasma membrane. Controlled apoptosis is necessary for development and maintenance of multicellular organisms. Most mice deficient in caspases, the effectors of apoptosis, die during development.

Apoptosis of self-reactive lymphocytes prevents autoimmune diseases, and apoptosis of virally infected cells prevents virus replication and spread. However, inappropriate apoptosis of neurons is thought to play a role in neurodegenerative diseases such as Alzheimer's, Parkinson's, and Huntington's diseases (reviewed in Friedlander 2003). The genes responsible for control of apoptosis were first elucidated in *C. elegans* with the identification and characterization of three *ced* genes (Ellis and Horvitz 1986). Discoveries of functionally homologous genes in insects, mammals, and many other organisms have shown that the apoptosis pathway is essentially conserved in many multicellular eukaryotes. However, the complexity is substantially increased compared to the *C. elegans* system, as there are hundreds of proteins that are involved in mammalian apoptosis.

There are two major apoptotic pathways in mammals (Figure 1.1, reviewed in Hengartner 2000; Zornig et al. 2001). The extrinsic pathway responds to signals from outside the cell such as binding of T-cell Fas ligand, and is mediated by death receptors of the tumor necrosis factor family and a variety of adaptor proteins such as TRADD and FADD. The intrinsic pathway responds to internal indicators such as p53 signals due to DNA damage, and is mediated by the mitochondria which, regulated by pro- and anti-apoptotic Bcl-2 family proteins, can be triggered to release pro-apoptotic molecules including cytochrome c and Smac/DIABLO. Both pathways, which also have means of cross-talk, ultimately result in recruitment, cleavage and activation of cysteine proteases known as caspases. Their proteolytic action on a wide variety of substrates results in the observed apoptotic morphology: DNA fragmentation, protein degradation, and eventual cellular disintegration and phagocytosis. Numerous regulators such as the IAPs (Inhibitor of Apoptosis Proteins) tightly control the activity of apoptosis proteins so that the process is only activated when specific environmental or genetic triggers are present. Control of

apoptosis is also exerted by regulators of growth and differentiation, such as the NFκB, Ras, and Jun kinase pathways. The number of genes with roles in apoptosis is ever-increasing, with some genes only recently described as associated with Bcl-2 proteins (Zhao et al. 2005), death receptors (Kamradt et al. 2005) and control of NFκB (Park and James 2005). Given the complexity of this process, there are undoubtedly more genes involved in regulation of apoptosis yet to be discovered.

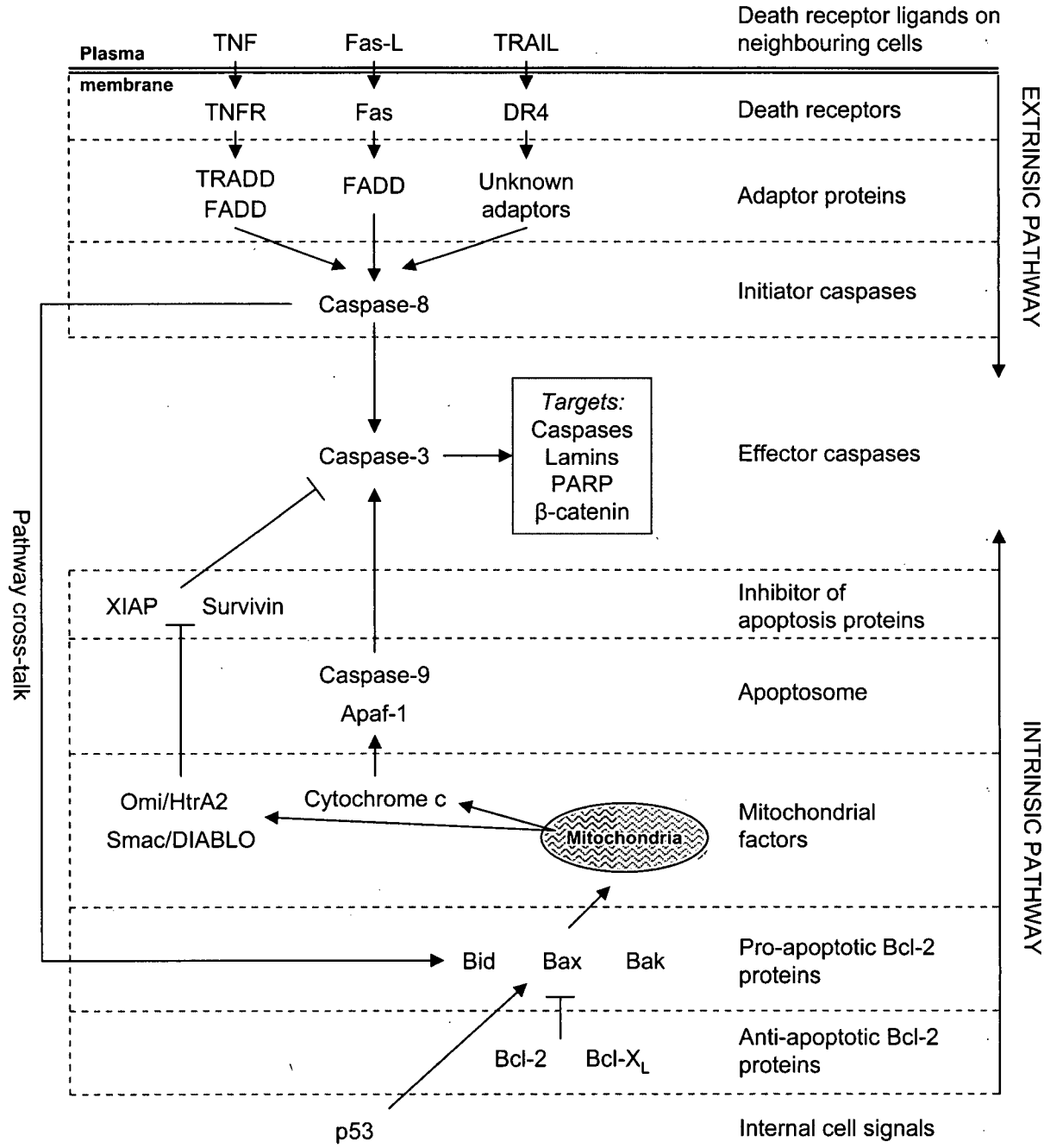

Figure 1.1 Mammalian apoptosis pathway.
Not all proteins known to be involved in apoptosis are shown, only representatives of the major pathway components. The mechanism of p53 effects are multi-fold, but include transcriptional regulation of Bax. Many of the Bcl-2 proteins shown are actually located on the mitochondrial membrane. Initiator caspases cleave effector caspases, which cleave both other caspases and various cellular targets. (Hengartner 2000; Zornig et al. 2001)

1.2.3 Autophagy

The term "autophagy" refers to multiple processes (Cuervo 2004b), but it is the process of macroautophagy which has been most extensively studied and implicated as having a role in development and disease. Macroautophagy, hereafter referred to as simply autophagy, is a mechanism of protein and organelle degradation and turnover that is conserved in organisms as distantly related as yeast and humans (reviewed in Levine and Klionsky 2004). It involves the sequestration of cellular components in double-membraned structures thought to be derived from the endoplasmic reticulum. These structures, called autophagic vacuoles or autophagosomes, fuse with lysosomes to form autolysosomes in which the sequestered cellular components are degraded. It is this pattern of autophagosomal activity that forms the morphological definition of autophagy. Autophagy in yeast is a mechanism to survive nutrient starvation, and it may have a similar function in mammals (Kuma et al. 2004). A role for autophagy in aging is suggested by its necessity for the formation of the long-lived *C. elegans* dauer larva (Melendez et al. 2003) and its association with increased lifespan in mammals (Bergamini et al. 2003). It also functions in cellular responses to bacterial infection and may be involved in cell differentiation (reviewed in Bursch 2001; Gozuacik and Kimchi 2004; Levine and Klionsky 2004). Defects in autophagy are implicated in muscular diseases and in neurological diseases such as Huntington's Disease and Alzheimer's Disease (reviewed in Cuervo 2004a; Marino and Lopez-Otin 2004; Shintani and Klionsky 2004).

The majority of the molecular components of autophagy were first identified using genetic screens in yeast. Three screens identified *aut* (Thumm et al. 1994), *apg* (Tsukada and Ohsumi 1993), and *cvt* (Harding et al. 1995) gene sets, which were subsequently shown to be partially overlapping; the nomenclature has since been unified and all genes are referred to as *atg*

(autophagy-related) genes (Klionsky et al. 2003). An overview of the mammalian autophagy pathway is shown in Figure 1.2. Initiation of autophagy, regulated by nutrient availability and mitogenic signals, is controlled by TOR (target of rapamycin). Early stages of autophagic vacuole formation are dependent on class III PI3K which is found in a complex with Beclin 1 (orthologous to Atg6 in yeast). Beclin 1 can induce and is required for normal autophagy (Liang et al. 1999; Qu et al. 2003; Yue et al. 2003). Expansion of the autophagosomal membrane occurs through the action of two ubiquitin-like systems, with the result that the Atg8 protein is tethered to the autophagosomal membrane. In mammals, several homologs of yeast Atg8 have been identified, but only MAP1LC3B, known as LC3, has a clear role in autophagy (reviewed in Tanida et al. 2004). LC3 is used as a marker for autophagosomes (Mizushima et al. 2004), as it is the only protein known to remain membrane-bound after the autophagosome has been completely formed. Beclin 1 and LC3 are among the most extensively studied mammalian autophagy genes, and both are integrally involved in the autophagic process.

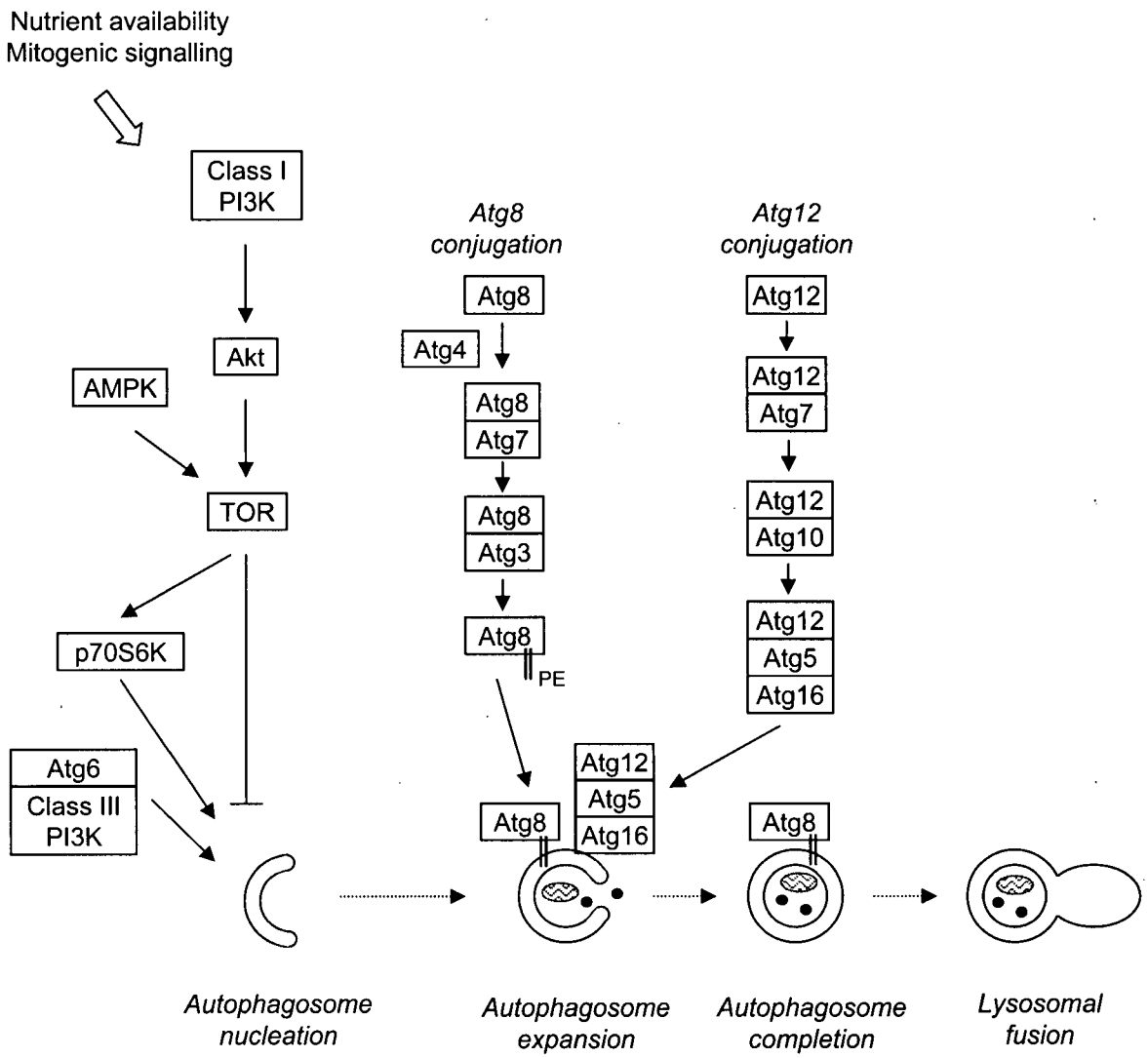

Figure 1.2 Mammalian autophagy pathway.
Atg6 is also known as Beclin 1. In mammals, there are at least three Atg8 homologs, the best studied of which is LC3. There are four Atg4 homologs, known also as Autophagin-1 to -4. PE, phosphatidylethanolamine, is a lipid inserted into the autophagosomal membrane anchoring Atg8. Several other pathways that are not shown, including the Ras pathway, also impact autophagy. (Baehrecke 2005; Levine and Klionsky 2004; Lum et al. 2005; Ng and Huang 2005; Tanida et al. 2004).

1.2.4 Autophagic cell death

Autophagic cell death has morphological features of autophagy, as the formation of large vacuoles which engulf and degrade both cytoplasm and organelles is observed prior to nuclear collapse (reviewed in Bursch 2001). The level of autophagic vacuole formation can be very extensive, such that a majority of the area in the cell is contained within vacuoles. Autophagic PCD is distinct both in morphology and molecular control from apoptosis. In addition to the presence of large autophagic vacuoles and early organelle degradation, autophagic PCD is marked by the preservation of the cytoskeleton, which is cleaved early in apoptosis but is presumably necessary for the movement of autophagic vesicles in type II PCD (Bursch et al. 2000). Molecularly, autophagic PCD is not caspase-dependent, but can be prevented by inhibitors of the autophagy pathway which target the PI3K pathway or prevent the fusion of autophagosomes with lysosomes, and can require autophagy genes such as Beclin 1 (Shimizu et al. 2004; Yu et al. 2004). However, little work has yet been done to define the molecular pathways controlling autophagic PCD, and thus the relationship between autophagy and autophagic PCD is unclear. For instance, autophagy is a protective mechanism of removing potentially detrimental protein aggregates in the brain, but high levels of autophagy are also observed in neurodegenerative disorders associated with cell death (reviewed in Levine and Klionsky 2004). In such cases, the observed autophagy-associated PCD may represent a progression of the autophagy pathway past the point at which the cell can survive, thus leading to PCD.

Many tissues which were previously thought to undergo only apoptosis, such as the tadpole tail, sexual structures during mammalian development, and various organs during insect metamorphosis, have since been shown to have additional hallmarks of autophagy, and thus the

mechanisms of PCD are in some cases being redefined (reviewed in Gozuacik and Kimchi 2004). It appears that autophagic PCD may play an important role when a large number of cells or an entire tissue need to be deleted. This will be discussed in more detail in the context of the *Drosophila* salivary gland in Section 1.4. Autophagic cell death and apoptosis are not mutually exclusive, and there may be some cross-talk between the respective signaling pathways as indicated by the regulation of Beclin 1 by Bcl-2 (Liang et al. 1998; Pattingre et al. 2005), and the triggering of autophagy when apoptosis is prevented, for instance, by Caspase-8 inhibition (Yu et al. 2004). However, as apoptosis can occur with no evidence for autophagy, so can autophagy occur with no evidence for apoptosis. In particular, there is growing evidence that under certain environmental or genetic conditions cancer cells undergo caspase-independent cell death with no apoptotic features, which can be prevented by autophagy inhibitors (Kitanaka et al. 2002). The role of autophagy and autophagic PCD in cancer will be discussed further in Section 1.3.3.

1.3 Cancer and programmed cell death

1.3.1 Molecular mechanisms of cancer

Cancer is a disease of uncontrolled cell growth caused by genetic changes that alter cell physiology. This process is exceptionally complex, and thus only a broad overview of concepts related to cancer progression and associated molecular changes is given here. These are discussed in more detail in Kufe et al (2003). A recent review identified 291 genes demonstrated to be involved in cancer, representing over 1% of human genes (Futreal et al. 2004).

For a cancer cell to survive, it must initially start to grow in an uncontrolled manner, beyond what is specified by various extracellular signals. Concurrently, it must circumvent the normal cellular mechanisms for recognizing and preventing this uncontrolled growth. It then must sustain this growth by obtaining sufficient nutrients and oxygen, even in a large tumor cell

mass. To advance to a metastatic state, it must gain the ability to live outside of its normal cellular milieu. Many different types of genetic mutations contribute to this complex multistep process, which differ from cancer to cancer and from cell to cell. Hanahan and Weinberg (2000) summarize the common molecular characteristics of cancers into six categories which specify the genetic changes a cancer cell must acquire to progress (Table 1.1).

The genes involved in cancer progression can be placed in two broad categories: oncogenes and tumor suppressors. Oncogenes promote cancer growth when active, and are often upregulated in cancers, for instance by increased expression or by mutations that increase their activity. Examples are the growth-promoting genes Ras and Myc. Tumor suppressors inhibit cell growth, and are often downregulated in cancers, for instance by decreased expression, inactivating mutations, or deletion. A very common example is p53, which can restrain the cell cycle or promote apoptosis, and is inactivated in an estimated 50% of cancers. A number of genes which regulate or execute the apoptotic program clearly fall into these categories. The role of autophagy genes in cancer is, however, more enigmatic.

Table 1.1 Acquired molecular changes in cancer progression.

Characteristic of cancer*	Molecular examples
Self-sufficiency in growth signals	• Production and secretion of extracellular growth factor PDGF results in autocrine growth signaling • Mutations in Ras cause constitutive activation of intracellular growth signaling pathways
Insensitivity to growth-inhibitory signals	• Mutation or inhibition of Rb allows cell cycle progression • Overexpression of Myc inhibits cell differentiation
Evasion of programmed cell death	• Bcl-2 upregulation blocks mitochondrial apoptosis pathway • Deletion of p53 prevents apoptosis due to DNA damage
Limitless replicative potential	• Upregulation of telomerase enzyme maintains telomere length which is otherwise lost after many cell divisions; loss of telomeres otherwise results in crisis, characterized by chromosome fusions and cell death
Sustained angiogenesis	• Upregulation of VEGF promotes growth of blood vessels into tumor, providing nutrients and oxygen • Downregulation of angiogenesis inhibitor thrombospondin-1 promotes vessel formation
Tissue invasion and metastasis	• Loss of E-cadherin function removes inhibitory cell-cell signals and connections • Proteases such as MMPs activated due to loss of inhibition by TIMPs degrade the extracellular matrix, promoting metastasis and invasion

* as categorized by Hanahan and Weinberg (2000).

1.3.2 Apoptosis and cancer

Current knowledge of tumorigenesis suggests that all cancers must inhibit apoptosis at some stage in their growth, as the apoptotic response is linked to alterations in many other cellular pathways (reviewed in Evan and Vousden 2001; Zornig et al. 2001). Overexpression of oncogenes such as Myc, in addition to signaling cell proliferation, also triggers apoptosis (Hipfner and Cohen 2004). Similarly, detachment of endothelial cells from the extracellular matrix during metastasis removes survival signals from integrins and cadherins and disrupts the cytoskeleton, resulting in a form of apoptosis called anoikis (Frisch and Screaton 2001). Genomic instability and DNA damage due to mutations and erosion of telomeres also trigger

apoptosis, primarily through p53, and result in cancer cell death (Evan and Vousden 2001). As might be expected, then, it is genes that promote apoptosis (pro-apoptotic) that are generally found to be inactivated or deleted in cancer, and thus are potential tumor suppressors. Genes that prevent apoptosis (anti-apoptotic) are potential oncogenes and are expected to be overexpressed.

Probably the most common mechanism resulting in inhibition of apoptosis is the mutation or deletion of p53, which protects the cell from apoptosis due to DNA damage and oncogene activation (Asker et al. 1999). If p53 is not mutated, its activity may be decreased by overexpression of its inhibitor Mdm2 or downregulation of the Mdm2 inhibitor ARF (Lowe and Lin 2000). Other common mutations seen in the apoptotic pathway include the overexpression of anti-apoptotic Bcl-2 proteins, for instance due to a translocation involving Bcl-2 in lymphoma (Zornig et al. 2001), or the downregulation of pro-apoptotic proteins such as the frameshift mutations of Bax seen in some colon carcinomas (Rampino et al. 1997), either of which can prevent apoptosis via the mitochondrial pathway. While the death receptor pathway is a less common target for oncogenic mutations, changes in Fas and other death receptors have been found in non-Hodgkin's lymphoma and gastric cancer (Gronbaek et al. 1998; Park et al. 2001), and the decoy receptor DcR3 that competes with death receptors for ligand binding is amplified in a significant proportion of lung and color cancers (Pitti et al. 1998). IAPs, especially survivin, are overexpressed in cell lines and many common tumor types (Ambrosini et al. 1997). Direct effects on caspases themselves are rare, possibly because it is difficult to stop the apoptotic program at the late stage of caspase activation, although methylation, deletion, or mutation of the Caspase-8 gene is seen in neuroblastomas (Takita et al. 2000; Teitz et al. 2000) and late stage gastric cancers (Soung et al. 2005). The central role of apoptosis in cancer and the commonality

of apoptosis-related alterations in cancer makes modulation of cell death pathways an attractive target for cancer therapies (Ghobrial et al. 2005).

1.3.3 Autophagy and cancer

One of the avenues of research that broadly ignited interest in investigating the role of autophagy in cancer was the study of the autophagy gene Beclin 1. Beclin 1 is monoallelically deleted in 40% of breast cancer cell lines, and its protein expression is reduced in a majority of breast and ovarian tumors (Liang et al. 1999). The MCF-7 breast cancer cell line has undetectable expression of Beclin 1 protein, and reintroduction of expression increases autophagy, decreases cell proliferation, and reduces MCF-7-derived tumor formation in mice (Liang et al. 1999). Additionally, Beclin 1 has been shown to act as a haploinsufficient tumor suppressor, as mice heterozygous for a deletion of *beclin 1* develop lymphomas, liver and lung cancers, and breast hyperplasias, with no evidence for additional mutations in the wild-type Beclin 1 gene (Qu et al. 2003; Yue et al. 2003). Although it is clear that inactivation of Beclin 1 can be an important step in the oncogenic process, it binds the apoptosis protein Bcl-2 (Liang et al. 1998; Pattingre et al. 2005) and may function through mechanisms other than autophagy (Furuya et al. 2005; Yu et al. 2004). Thus, an outstanding question is whether it is reduced autophagy that is responsible for the tumorigenic effects of Beclin 1 deletion mutants.

Recently, a number of studies have demonstrated that autophagy has a role in cancer causation, progression, and therapy. Multiple lines of evidence suggest that autophagy has a role in suppressing oncogenesis, either through its role in autophagic cell death, or through other mechanisms such as regulation of cell growth (reveiwed in Gozuacik and Kimchi 2004). Rat models of hepatocellular carcinomas and pancreatic adenocarcinomas demonstrate decreased levels of autophagy (Schwarze and Seglen 1985; Toth et al. 2002). Autophagic cell death can be

induced by oncogenic forms of Ras in glioma and gastric cancer cell lines (Chi et al. 1999), and may contribute to spontaneous regression of neuroblastomas (Kitanaka et al. 2002). MCF-7 breast cancer cells undergo autophagic cell death in response to tamoxifen (Bursch et al. 1996), as do glioma cells when treated with the chemotherapeutic arsenic trioxide (Kanzawa et al. 2003). These data suggest that autophagy and autophagic cell death can act to prevent cancer, and thus induction of these pathways could be of therapeutic benefit in treatment of at least some cancers.

However, autophagy does not only have an inhibitory role in oncogenesis; indeed, it may in some situations contribute to cancer progression (reviewed in Ogier-Denis and Codogno 2003). Some cancers exhibit high levels of autophagy, which may aid in cancer cell survival under conditions of nutrient starvation and hypoxia which are common in preangiogenic solid tumors. The potential survival advantage of active autophagy in early stage cancers is supported by the observation that although rat models of pancreatic carcinomas show decreased autophagic capacity compared to normal pancreatic cells, premalignant rat pancreatic cancer cells show increased levels of autophagy (Toth et al. 2002). Additionally, autophagy may diminish the effects of chemotherapy or irradiation, potentially by contributing to elimination or sequestration of toxic molecules and damaged organelles (Cuervo 2004a; Ogier-Denis and Codogno 2003; Paglin et al. 2001). Thus, it appears that autophagy may have dual roles in cancer. In certain cancer types or stages, it may aid in cancer cell survival. In other types or stages, or in response to other stimuli such as blocked apoptosis or presence of particular chemotherapeutics, autophagic cell death may be induced (Ogier-Denis and Codogno 2003).

1.4 Programmed cell death in the development of *Drosophila melanogaster*

1.4.1 Drosophila as a model for programmed cell death

During development and metamorphosis of the fruitfly *Drosophila melanogaster*, a number of tissues are subject to programmed cell death (Figure 1.3A, reviewed in Baehrecke 2000; Rusconi et al. 2000). Tissues such as the salivary gland and midgut are removed completely during metamorphosis by autophagic cell death to make way for the adult organs (Figure 1.3B). In other tissues, such as the eye, it is only individual cells which are deleted by apoptosis in order to shape the tissue in a precise manner (reviewed in Brachmann and Cagan 2003). As discussed above, the pathways responsible for programmed cell death are conserved in many organisms, and many human PCD genes have homologs in the *Drosophila* system. Importantly, both the apoptotic and autophagic systems are in place in the fly. However, the pathways in *Drosophila* are less complex and less redundant; for instance, there have been at least 14 caspases identified in mammals (Zhang et al. 2003), and only seven in *Drosophila* (Kornbluth and White 2005). The conservation of PCD elements between species implies that novel genes found in this model system are likely to have relevant functional homologues in mammalian systems, but at the same time will be easier to study in *Drosophila* both because of the relative ease of genetic studies in this model, and the reduced complexity of the system (Richardson and Kumar 2002).

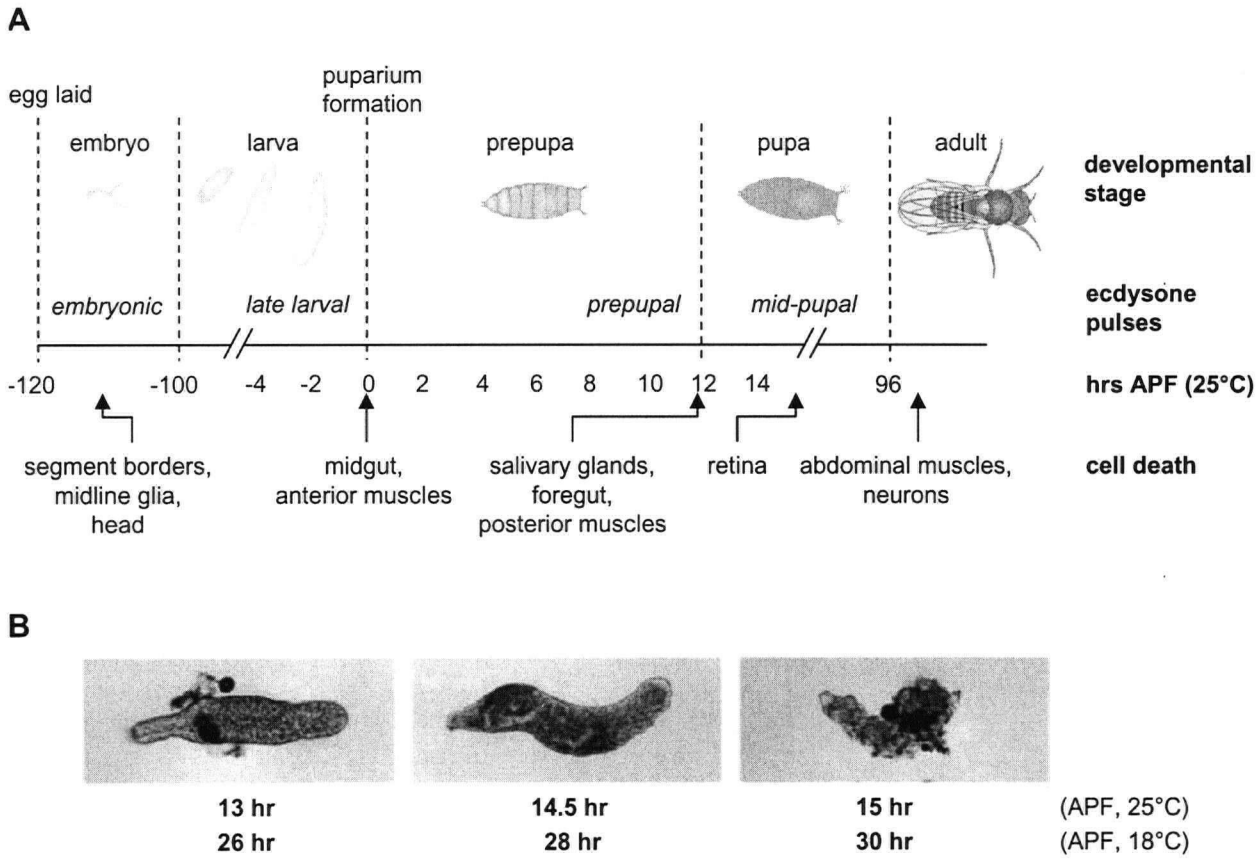

Figure 1.3 Programmed cell death during *Drosophila* development.
(A) Pulses of the steroid hormone ecdysone punctuate development. The late larval pulse in the 3rd instar larva triggers destruction of the midgut, and the prepupal pulse triggers destruction of the salivary gland; both are destroyed by autophagic cell death. The time scale is expanded in the prepupal stage. Images from (Griffiths et al. 1999). (B) The prepupal ecdysone pulse triggers death and removal of the salivary gland by autophagy. Images from (Jiang et al. 1997). APF, after puparium formation.

1.4.2 Molecular mechanisms of programmed cell death in Drosophila

Apoptosis in *Drosophila*, as in the mammalian system, is executed by caspases. The caspase-inhibitory proteins, IAPs such as Diap1, have a central role in controlling apoptosis as lack of these proteins results in increased apoptosis and lethality (Hay et al. 1995). Primary control of IAPs is exerted by three IAP-inhibiting pro-death proteins, Hid, Reaper, and Grim (Goyal et al. 2000), and also by two recently discovered proteins which appear to have similar function, Sickle and Jafrac2 (Srinivasula et al. 2002; Tenev et al. 2002). Though these genes do

not have direct mammalian homologs, several can induce apoptosis by activating caspases in mammalian cells (Varghese et al. 2002) and they are analogous in function to mammalian Smac/DIABLO and Omi/HtrA2 (Du et al. 2000; Hegde et al. 2002). Expression of these proteins allows caspases to be activated and thus triggers death. A mitochondrial apoptosis pathway also exists, involving the Bcl-2 homologues dBORG-1 and 2 and the Apaf-1 homologue Dark, which initiates activation of Dronc and downstream caspases (Kumar and Doumanis 2000). Several important PCD-related proteins, such as the Bcl-2 homologues (Colussi et al. 2000) and several more caspases (Doumanis et al. 2001; Harvey et al. 2001), have only recently been recognized, and there are undoubtedly more apoptosis regulators to be identified in *Drosophila*. For instance, the initiator caspase Dredd binds a protein very similar to human FADD (Hu and Yang 2000) and thus resembles in function human Caspase-8, but the putative death receptors which may initiate this pathway have yet to be uncovered. In general, few upstream regulators of cell death have been identified in *Drosophila*, and thus research into potential upstream regulators of caspases, transcriptional regulators of Reaper, Hid and Grim, and other as yet unidentified control mechanisms has the potential to contribute significantly to the understanding of cell death in *Drosophila* (Hay et al. 2004).

Although 11 putative homologs of autophagy genes have been identified in *Drosophila* (Baehrecke 2003), most have not been studied in detail. *Drosophila* Atg4 (Apg4/Aut2) can complement its yeast counterpart, but mutants have no phenotype in *Drosophila*; genetic studies have demonstrated that it is a modifier of Notch signaling but no studies on its role in autophagy have yet been done (Thumm and Kadowaki 2001). Atg3 (Apg3/Aut1), on the other hand, is required for autophagy in *Drosophila* fat body cells and mutants die during metamorphosis, demonstrating this gene is part of the *Drosophila* autophagy pathway and that this pathway is

required for normal development (Juhasz et al. 2003). Several other putative autophagy genes have recently been shown by loss of function to be necessary for autophagy in *Drosophila* fat bodies (Scott et al. 2004). As in the human system, the exact role of autophagy in programmed cell death is unclear but intriguing. The *Drosophila* salivary gland has been a focus of research in this area.

1.4.3 Programmed cell death in the Drosophila salivary gland

The *Drosophila* salivary gland is an excellent model in which to study gene regulation during PCD, because nearly every cell in the gland undergoes PCD within a well-defined, short time period (Figure 1.3B). This provides a large number of cells which are all in the same known stage of PCD for analysis. Developmental PCD in the salivary gland is triggered by a pulse of the steroid hormone ecdysone at the prepupal-pupal stage transition. Upon ecdysone binding to the ecdysone receptor, this complex along with βFTZ-F1 initiates a transcriptional cascade (Figure 1.4) by activating the early genes BR-C, E74A and E93, which in turn regulate downstream genes including the IAP inhibitors Hid and Reaper (Jiang et al. 1997; Jiang et al. 2000; Lee et al. 2002a; Martin and Baehrecke 2004). This results in activation of caspases such as Dronc and death of the salivary gland.

Intriguingly, although genes involved in the apoptotic pathway are required for salivary gland death, and the death has hallmarks of apoptosis such as DNA fragmentation (Jiang et al. 1997; Martin and Baehrecke 2004), the primary morphology of salivary gland death is autophagic (Lee and Baehrecke 2001). Large autophagic vacuoles are formed, and changes in the cytoskeleton and increased acid phosphatase activity indicate vesicular movement and protease degradation (Jochova et al. 1997). Inhibition of caspases by baculovirus p35 does prevent DNA fragmentation and death of the salivary glands, but they progress to a late stage of autophagic

activity. Mutants in the early genes BR-C and E74A similarly progress to late stages but salivary glands are not destroyed, whereas βFTZ-F1 and E93 mutants show early defects in vacuolarization (Lee and Baehrecke 2001). Our work (Gorski et al. 2003) as discussed in Chapter 3, along with a similar large-scale study published concurrently (Lee et al. 2003), has demonstrated that multiple autophagy genes are expressed during autophagy in the salivary gland, suggesting that the canonical autophagy pathway is involved in salivary gland death. Thus, programmed cell death in the salivary gland appears to be autophagic but with hallmarks of apoptosis, possibly with some level of overlap between pathways in the genes involved, and is a model for understanding the relationship between these very different types of cell death.

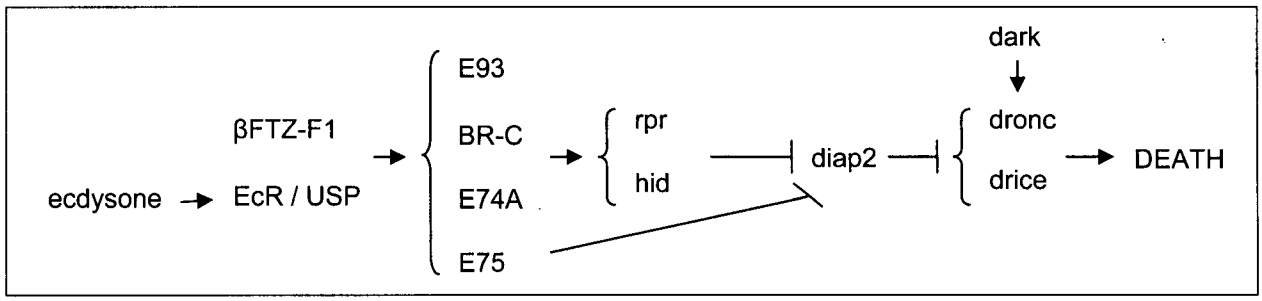

Figure 1.4 Ecdysone-triggered transcriptional cascade controlling salivary gland PCD.
Binding of ecdysone to the ecdysone receptor, in the presence of the competency factor βFTZ-F1, initiates transcription of early genes, which in turn act as transcription factors to express late genes. Repression of Diap2 allows caspases including Dronc and Drice to cleave downstream substrates.

1.5 Gene expression analysis

1.5.1 Gene expression and mRNA levels

Messenger RNA is the molecular intermediate between DNA and protein, and the quantity of mRNA for a gene present in a cell is often used as a surrogate measure of the activity of the functional gene product, generally a protein. mRNA levels are determined by several

levels of regulation, including transcription, splicing, and mRNA stability (Akker et al. 2001; Day and Tuite 1998). Although correlations between mRNA and protein levels vary, there is generally high consistency for abundantly expressed genes (Chen et al. 2002a; Gygi et al. 1999; Kern et al. 2003). Additionally, genes whose protein products are involved in the same protein complex are expressed similarly (Greenbaum et al. 2002), and it is generally thought that changes in mRNA levels, because they are specifically regulated, are indicative of changes in cellular function. Cellular activities are often measured at the mRNA level because, unlike protein abundance, mRNA abundance is relatively easy to assay. Both low throughput methods such as Northern blots and high-throughput methods such as microarrays rely on the natural RNA base-pairing to probe for the presence of particular mRNA molecules; other methods such as ESTs and SAGE rely on reverse transcription and sequencing of mRNAs. Several such high-throughput methods for gene expression analysis, that allow simultaneous assay of large numbers of mRNA molecules and thus generate large datasets requiring computational algorithms for analysis, are described below in more detail.

Analysis of mRNA expression measurements can identify genes that change in expression between conditions and thus have a potential role in the system under study. The utility and potential power of these studies is indicated by the size of publicly available gene expression repositories such as the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/), containing tens of thousands of experiments representing hundreds of millions of measurements (Barrett et al. 2005). In the study of cancer, it is common to compare tumor cells to nontumor cells, compare types of cancers, or compare expression across a timecourse such as cancer progression or drug effects in cell culture. A subset of genes, including the autophagy gene Beclin 1 (Liang et al. 1999), are regulated at the protein level and thus are not identifiable

through mRNA expression profiling. In cancers, however, this subset may be reduced as genes normally regulated only at the protein level may be perturbed due to mutations causing deletions, over-, or under-expression. Large-scale analysis of gene expression is particularly powerful for indicating cellular pathways that are perturbed in the process under study, and for identifying new genes involved in the process without the necessity of prior functional knowledge.

1.5.2 Microarrays

Microarray methods, using either cDNA or oligonucleotide-based chips, make use of hybridization rather than sequencing to measure mRNA expression levels. cDNAs or oligonucleotide sequences are printed or constructed in known locations on a glass slide or other substrate. mRNA is reverse transcribed to cDNA and labeled, and then allowed to hybridize with the DNA on the chip. After washing, the measured intensity of the fluorescence is related to the expression level of the mRNA in the sample. Two-color microarrays, as described originally (Schena et al. 1995), use long PCR products or complete cDNAs to measure relative hybridization levels of two samples. Oligonucleotide microarrays, such as those produced by Affymetrix (Lipshutz et al. 1999), use shorter sequences of 25-70 nucleotides designed to precisely and specifically match known genes and thus permit more direct rather than relative quantitiation. These methods are discussed in more detail in (Butte 2002; Lipshutz et al. 1999).

Microarrays have significant advantages over other large-scale expression measurement methods as they are relatively inexpensive, widely available, and produce results quickly, but they also have disadvantages that can affect the accuracy of the resulting measurements. Probe design is a major determining factor in the efficacy of microarray analysis. Prior knowledge of the genes to be profiled is necessary to place the probes on the array, and thus microarrays, unless specifically designed to do so (Shoemaker et al. 2001), do not have the ability to detect

novel genes. Additionally, genes which are very similar in sequence may hybridize to each others' probes (cross-hybridization) and thus interfere with correct expression measurement (Evertsz et al. 2001; Zhang et al. 2005); this is a more significant problem with longer probes such as those used on cDNA arrays. Analysis of microarray data to determine true expression levels is also complicated by issues such as differences in cDNA labeling efficiencies and probe affinities, and requires sophisticated computational tools for steps including image processing, background subtraction, and normalization (discussed in Quackenbush 2001).

Many cellular systems in a wide range of organisms have been subjected to microarray analysis, and microarrays have been used extensively in the study of cancer. Gene expression during *Drosophila* development from embryo to adult, and in specific tissues including the salivary gland and midgut, has been profiled with microarray time courses (Arbeitman et al. 2002; Li and White 2003). The applications of microarray analysis to cancer include using gene expression profiles to define subtypes of breast cancer (Sorlie et al. 2003), and differentiate patients with poor prognoses (van 't Veer et al. 2002). Meta-analysis of multiple microarray experiments can identify sets of genes commonly deregulated in tumors (Rhodes et al. 2004). Because they are relatively economical compared to other large-scale gene expression methods, it is thought that microarrays may become practical for use in diagnosis. All large-scale gene expression methods have the potential to identify genes important in cancer progression that could potentially be markers or targets for treatment.

1.5.3 Expressed Sequence Tags

Expressed sequence tags, or ESTs, are single-pass sequences representing the 3' or 5' ends of genes in a cDNA library that is constructed by reverse transcription of mRNAs extracted from a cell population. Each sequence read corresponds to one gene, and thus ESTs are usually

300-500 nucleotides long. As a result, ESTs are good for gene identification and genome annotation as they can be compared to other genes and to genome sequence, but are expensive and not best suited for quantitative large-scale expression profiling.

Nearly 30 million ESTs have been sequenced and stored in the dbEST database (http://www.ncbi.nlm.nih.gov/dbEST/, Boguski et al. 1993), of which the largest proportion are human. Use of ESTs for quantitative profiling is limited by the common practice of normalizing cDNA libraries to avoid repeatedly sequencing the same abundant transcripts, and by some skewing of expression levels (Haverty et al. 2004). However, EST sequences are very valuable for identifying alternative splice variants, and combined with expression levels can identify tumor-associated variants (Hui et al. 2004). Additionally, ESTs are widely used in gene annotation (Curwen et al. 2004). Over 200,000 *Drosophila* ESTs have been sequenced, and used both for identification of full-length clones for the majority of *Drosophila* genes (Stapleton et al. 2002) and for gene prediction (Misra et al. 2002).

1.5.4 Serial Analysis of Gene Expression

Serial analysis of gene expression (SAGE) is a sequence-based rather than a hybridization-based approach to expression profiling developed in 1995 (Velculescu et al. 1995) and now widely used. It is both more efficient and more quantitative than EST-based profiling. As this technique and the associated analysis are central to this thesis, the experimental method and computational analysis of SAGE data are described here in more detail.

1.5.4.1 SAGE experimental method

Serial analysis of gene expression, or SAGE, involves extraction of short sequences called SAGE tags from polyadenylated RNA by conversion to cDNA followed by a series of restriction

digestions. Tags are then PCR amplified, concatenated, and sequenced (Figure 1.5, Velculescu et al. 1995). The SAGE method is designed such that the frequency of SAGE tags in the final sequences should be directly proportional to the abundance of their parent mRNA molecules. The position in the transcript from which the tag is extracted is determined by the 4-cutter anchoring enzyme used, commonly *Nla*III which recognizes "CATG", but also occasionally *Sau*3A which recognizes "GATC". As the 3' end of the cDNA is immobilized and all non-immobilized fragments are removed after cleavage with the anchoring enzyme, only the portion of the transcript 3' to the anchoring enzyme site is retained. After the linker sequence containing the tagging enzyme recognition site and PCR primers is ligated, tagging enzyme cleavage results in a tag representing the sequence adjacent to the 3'-most anchoring enzyme site. *BsmF*I is the most common tagging enzyme used, but others such as *Mme*I are also employed, and this defines the length of the extracted SAGE tag (Table 1.2).

There are several sources of bias in the SAGE method that can skew the resulting tag frequencies and lead to misinterpretation of gene expression levels. The efficiency of the blunt-end ligation used in the standard SAGE procedure varies dependent on the terminal nucleotides of the sequences to be ligated (Yamamoto et al. 2001). In LongSAGE, 2 bp overhangs are retained for ligation. Both of these approaches can result in nonrandom pairing of tags for amplification and therefore influence overall tag frequency. An independent bias in GC content has also been observed whereby room-temperature storage of ditags results in denaturation and loss of AT-rich tags (Margulies et al. 2001). Finally, PCR amplification has an inherent potential for bias, as not all sequences are amplified with equal efficiency (Warnecke et al. 1997). AT-rich ditags may produce as much as 7-fold more amplicon than GC-rich tags (Spinella et al. 1999). The potential for PCR bias was recognized as a concern in the original description of SAGE

(Velculescu et al. 1995), and it is partly to account for this problem that ditags are introduced, rather than simply amplifying and ligating tags individually. Because ditags are formed before amplification, and there are generally an extremely large number of different tag species, it is expected that only rarely will two tags be coupled to produce the same ditag more than once by chance. Therefore, if two identical ditags ("duplicate ditags") are observed in the final SAGE sequences, such ditags are proposed to have arisen from biased PCR amplification (Velculescu et al. 1995). Duplicate ditags are observed in essentially every SAGE library, commonly at frequencies of 5-10% but as high as 50% in less complex tissues (Gorski et al. 2003).

Several experimental issues can produce tags other than those expected. Only SAGE tags derived from the 3'-most anchoring enzyme (e.g. *Nla*III) site are canonical or expected tags in SAGE libraries. During cDNA synthesis, extended strings of A's can bind the oligo-dT primer, resulting in cDNA production starting at sites upstream of the poly-A tail, and thus can produce SAGE tags from farther upstream in the transcript. Similarly, tags derived from 2, 3 or even 4 anchoring enzyme sites upstream from the 3'-most are observed due to incomplete digestion by the tagging enzyme. Additionally, some proportion of SAGE tag sequences produced are incorrect. Sequence error rates in SAGE libraries have been estimated at 0.7%-1% per base, or 6.8%-9.6% per 10 bp tag (Colinge and Feger 2001a; Velculescu et al. 1997). Less frequently, similar errors in SAGE tags can result from misincorporation of bases during PCR. Formation of linker dimers during tag cleavage and ligation can produce artifactual SAGE tags whose sequence is derived from the linker sequence itself and not the transcript. Linker contamination in SAGE libraries usually ranges from less than 1% to 5% of sequenced tags (Cheng and Porter 2002; Velculescu et al. 1999). Incorrect "quasi-ditags" can also arise from genomic contamination or random nucleotide combinations, and are most likely to appear in clones with

only one or two ditags (Anisimov and Sharov 2004). All of these issues of sequence bias, sequence errors, and non-canonical tags must be dealt with where possible in the computational analysis of SAGE libraries.

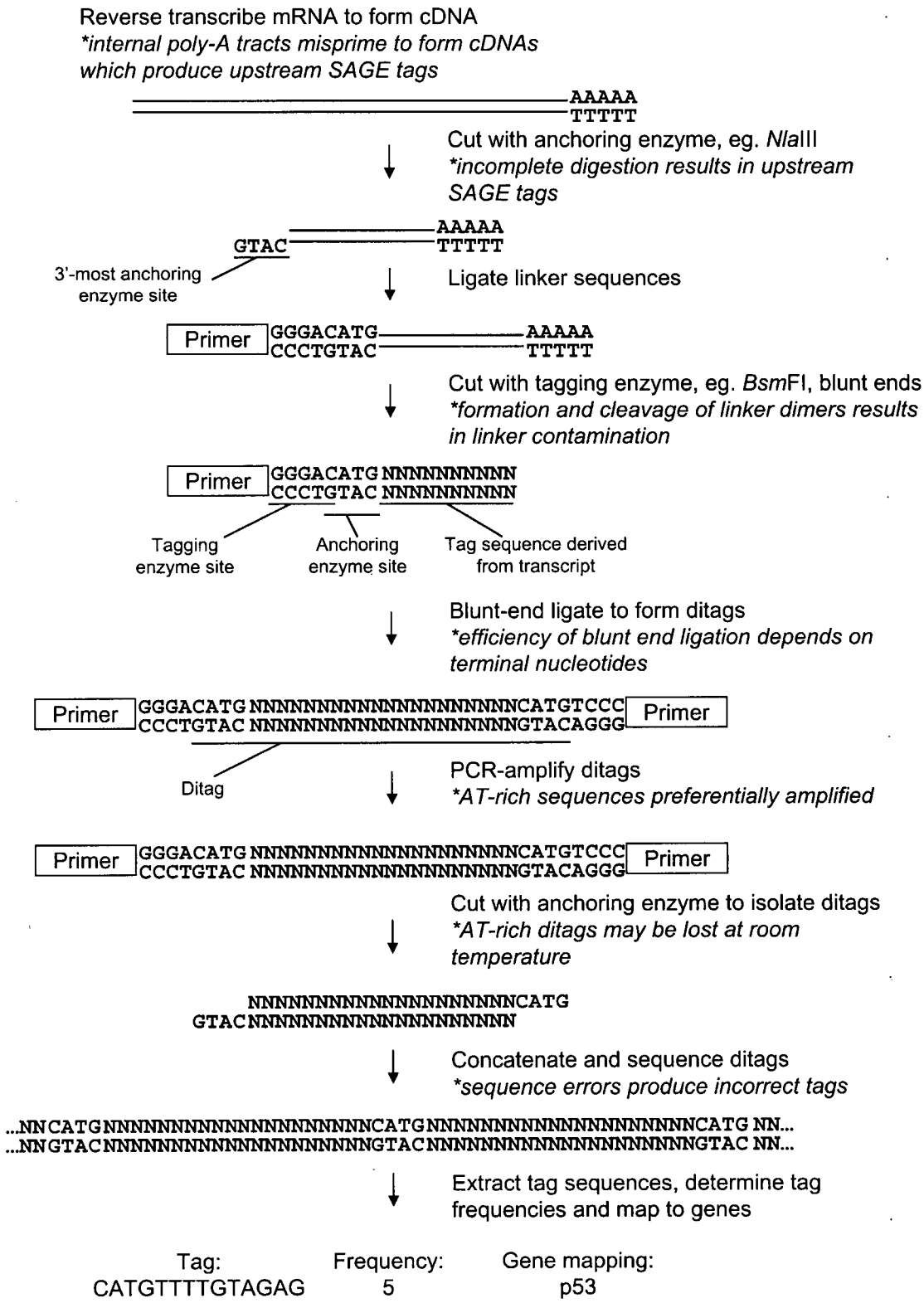

Figure 1.5 Diagram of the SAGE procedure.
Steps marked in italics are subject to errors and biases. For more detailed procedure, see (Velculescu et al. 1995) or www.sagenet.org. From Pleasance and Jones (2005).

Table 1.2 Variations in the SAGE procedure affecting SAGE tag length.

Procedure	Anchoring enzyme	Tagging enzyme	Tag length	Reference
SAGE (original)	*Nla*III	*Fok*I	13 bp	(Velculescu et al. 1995)
SAGE (now in practice)	*Nla*III	*Bsm*FI	14 bp	(Velculescu et al. 1997)
SADE	*Sau*3A	*Bsm*FI	14 bp	(Virlon et al. 1999)
Modified SAGE	*Nla*III	*Rsa*I	18 bp	(Ryo et al. 2000)
LongSAGE	*Nla*III	*Mme*I	21 bp	(Saha et al. 2002)
SuperSAGE	*Nla*III	*EcoP15*I	26 bp	(Matsumura et al. 2003)

1.5.4.2 SAGE tag processing and statistics

Sequence reads from SAGE libraries consist of a series of ditag sequences separated by anchoring enzyme sites, and it is commonly assumed that all tags are the same length. However, variation in cleavage by the *Bsm*FI tagging enzyme produces tags ranging from 13 to 17 bp including the anchoring enzyme site (Yamamoto et al. 2001). Therefore, there is no way to determine if a particular ditag of 28 bp corresponds to ligation of two 14 bp tags, or a 13 bp tag and a 15 bp tag. Approximately 88% of ditags are 30 or 31 bp long, corresponding to tags of 15-16 bp (Colinge and Feger 2001b), so the likelihood of one base of a longer tag being erroneously assigned to a 13 bp tag is small and therefore 14 bp tags are commonly extracted. This variation in tag length, however, provides an opportunity to extend an extracted SAGE tag to the 15th base if a longer tag is needed to distinguish between two possible gene assignments. This extension can be done by determining the most common 15th base for a particular tag using statistical methods to determine confidence (Colinge and Feger 2001b), but still increases the likelihood of misassigning a base to the wrong tag and so should be used with caution.

Many of the issues discussed in Section 1.5.4.1 can be mitigated by appropriate computational processing to remove incorrect SAGE tags. The most common way to deal with PCR bias is to remove all but one copy of repeated ditags (Lash et al. 2000; Velculescu et al.

1995). However, if genes of extremely high abundance are present, for instance those representing >1% of total tags, they may pair in a significant fraction of ditags to produce non-artifactual duplicate ditags in the final sequence. In such cases, it has been suggested that, based on observed tag frequencies, a certain calculable proportion of duplicate ditags should not be removed from analysis as they may represent true transcript expression (Vilain et al. 2003). Non-canonical tags produced due to mispriming or incomplete digestion can be included in an analysis if they can be correctly assigned to genes (see Section 1.5.4.3), in which case their frequency can be added to that of the canonical tag for the same gene. Tags resulting from incorporation of linker sequence can be easily identified computationally and simply removed. Sequence errors can be reduced by selecting tags only of high sequence quality (Gorski et al. 2003; Pleasance et al. 2003). However, there is no way to detect PCR errors through Phred scores. For relatively small libraries, the chance of seeing the same sequence or PCR error twice is very low and thus the approach of removing all tags seen only once ("singletons") will alleviate almost all such errors (Lash et al. 2000). For tags with expression high enough to produce multiple redundant errors, all possible permutations of a tag must be determined, and tags corresponding to these permutations identified as potentially affected by sequence error. These tags can either be excluded from the library if their expression is not significantly above that expected due to error (Velculescu et al. 1999), or adjusted in frequency by estimating the original tag count without sequencing error (Akmaev and Wang 2004; Beissbarth et al. 2004; Colinge and Feger 2001a).

As SAGE tag counts are absolute, SAGE libraries from different tissues or conditions can be directly compared once all sequence processing is complete and tag frequencies are determined. A number of statistical tests for SAGE library comparison have been developed.

Tests generally involve the comparison of two libraries, where the frequency of a tag in one library is compared to the frequency of that tag in another and the probability (p-value) that the observed difference in frequencies would have occurred by chance due to sampling error is determined. Some statistical tests have been developed particularly for SAGE analysis (Audic and Claverie 1997), while others are more generic tests applied to SAGE data (Man et al. 2000). A comparison of the number of tags found to be differentially expressed in various situations with each of these tests revealed that most of them produce equivalent results (Ruijter et al. 2002). Several web-based tools have been developed to apply statistical tests to experimental SAGE libraries (Lal et al. 1999; Pylouster et al. 2005).

1.5.4.3 SAGE tag-to-gene mapping

An essential step in SAGE is the unambiguous correlation of the 14 bp SAGE tag with the transcript from which it is derived, a process termed "tag mapping". It generally involves automated searching for the tag sequence in sequence databases. This process is complicated by variation in gene structure and sequence including alternative splicing, alternative polyadenylation site usage, polyadenylation cleavage heterogeneity, and single nucleotide polymorphisms (SNPs). It is estimated that between 35% and 75% of human genes are alternatively spliced (Johnson et al. 2003; Xie et al. 2002; Xu et al. 2002), producing different transcripts by use of different exons. Computational extraction of SAGE tags from alternatively spliced transcripts suggests 38% of human genes produce more than one tag due to alternative splicing (Unneberg et al. 2003), and 24% of *Caenorhabditis elegans* genes produce multiple SAGE tags (Jones et al. 2001). 30-50% of genes are likely to use multiple different polyadenylation signals, producing 3' ends that vary in length by up to several kb (Beaudoing et al. 2000; Iseli et al. 2002). As this variation is focused at the 3' end of the transcript, and

generally changes the length of the transcript by several hundred nucleotides, use of different polyadenylation signals can result in extraction of different SAGE tags. In addition, the exact site of polyadenylation cleavage varies from 10-25 nucleotides downstream of the polyadenylation signal. The approximately 3% of human SAGE tags affected by cleavage heterogeneity (Pauws et al. 2001) can be difficult to map to genes as part of their sequence is often derived from the poly-A tail itself, and thus a one-base-pair variation in cleavage (eg. GGTCGAAAAA vs. GGTCAAAAAA) can result in a tag that no longer matches any known sequence. Finally, SNPs result in slight changes in a gene's nucleotide sequence, and thus can produce a SAGE tag that does not exactly match the expected sequence for that gene; as many as 8.6% of genes may produce such tags (Silva et al. 2004). It can also be difficult to distinguish true gene variations from errors in SAGE tags due to sequencing errors and incomplete digestion.

Full-length cDNA sequences which exactly match the transcripts from which SAGE tags are experimentally derived would be ideal for making the most accurate assignment of SAGE tags to genes. A number of full-length cDNA sequencing projects are ongoing (Ota et al. 2004; Stapleton et al. 2002; Strausberg et al. 1999), but complete transcript sequences of all gene variants are not currently available for any organism. Sequences that are not full-length will result in tags that cannot be assigned to genes. For instance, in 54% of *Drosophila* full-length cDNAs, SAGE tags are derived from either the 5' or 3' UTR and thus full-length coding sequences alone are not sufficient for tag mapping (Pleasance et al. 2003). Having an incomplete set of gene sequences can also be very misleading, as some tags will not be assignable, and others may be assigned to the wrong gene. A common issue in SAGE tag mapping is that, because of the limited length of a SAGE tag, two or more different genes may produce identical SAGE tags and thus a tag sequence may not have a unique gene assignment. If not all gene

sequences are known, a tag may appear to match one gene uniquely, when in reality it also matches another gene as well and thus its frequency may represent the expression of one or both genes. Choosing the sequences and methods used for SAGE tag mapping has a significant effect on the completeness and accuracy of the resulting gene expression profiles.

For simpler organisms such as yeast, simply mapping to ORFs and 500 bp of downstream sequence can assign 88% of tags, and mapping the remaining tags to genomic sequence assigns in total 93% (Velculescu et al. 1997). For more complex and larger genomes, however, this method is not applicable. Instead, a number of different methods have been developed for assigning SAGE tags to genes, most of which use a combination of different sequence databases.

The first comprehensive human SAGE tag mapping resource to be developed was SAGEmap, available from NCBI (Lal et al. 1999; Lash et al. 2000). SAGEmap extracts SAGE tags from GenBank mRNA and 3' EST sequences that are grouped into UniGene clusters. Tags extracted from the 3'-most *Nla*III or *Sau*3A site are assigned reliability based on the type of sequence they are derived from, where mRNAs are more reliable than ESTs, and ESTs with polyadenylation signals and poly-A tails are more reliable than those without as they can be correctly assigned an orientation. There are a number of difficulties with using ESTs for tag mapping. Because ESTs are single-pass sequences, they have an error rate of approximately 1% and thus an overall tag error rate of ~10% (Lash et al. 2000). In addition, clustering of ESTs into gene groups as is done in the UniGene database (Schuler 1997) is a difficult process due to sequence errors, sequence artifacts such as chimeras, incomplete gene sequences, and the existence of highly similar genes. ESTs belonging to different genes may be clustered together, or the same gene may be represented in different EST clusters. Thus, EST-based tag mapping can result in erroneous or highly nonunique SAGE tag assignments. To accommodate the

expected EST sequence errors, SAGEmap removes the 10% of least commonly observed tag-gene pairs to create a "reliable" mapping. In this reliable mapping, 87% of tags can be assigned uniquely to one gene and 59% of genes have only one tag, compared to mappings using mRNAs only from which 96% of tags were unambiguous, and 80% of genes had a single tag.

The SAGE Genie tool refines the SAGEmap approach by incorporating full-length sequences from the Mammalian Gene Collection (MGC, Strausberg et al. 1999) and the Reference Sequence (RefSeq) database at NCBI (Maglott et al. 2000), and devising a datasource scoring method based on experimentally observed SAGE tags (Boon et al. 2002). Similar approaches have since been applied by others (Bala et al. 2005). For SAGE Genie, tag databases were created from MGC, RefSeq, mitochondrial, clustered mRNA, clustered EST, and unclustered EST sequences, divided based on presence or absence of polyadenylation signals and poly-A tails. Tags were extracted from the four 3'-most *Nla*III sites, as well as sites upstream of poly-A tracts if they were confirmed by EST or cDNA evidence to be potential sites of internal priming. A confident SAGE tag list of 194,126 different experimental tags was also produced, using 6.8 million tags from 171 SAGE libraries filtered to remove linker sequences, sequence errors, and singletons. Each of the tag databases was then scored based on the percentage of computationally extracted tags that were represented in the confident SAGE tag list, with the expected result that MGC sequences were the most reliable, followed by RefSeq, clustered sequences, and finally unclustered ESTs. For tag-to-gene mapping, the final "best" assignment is then made based on the database the tag was extracted from, the position relative to the 3' end, whether the site is internally primed, and the expression level of the gene based on the confident tag list where more highly expressed genes are considered more reliable. A limited manual

confirmation of 77 tag-to-gene assignments suggests the method produces fairly accurate tag mappings.

As more eukaryotic genome sequences are completed (Adams et al. 2000; Arabidopsis Genome Initiative 2000; *C. elegans* Sequencing Consortium 1998; Lander et al. 2001; Waterston et al. 2002), it is possible to assign SAGE tags directly to genomic sequence. This eliminates the problems arising from incomplete representations of the transcriptome by sequenced cDNAs, difficulties with clustering sequences, and effects of sequence errors, as finished genome sequences have error rates typically below 1/10,000 bp (*C. elegans* Sequencing Consortium 1998). However, SAGE tags derived from sites in the transcript that cross exon splice junctions will not exist directly in genomic sequence, and are very computationally difficult to search for. Tags with sequence derived partly from the poly-A tail will also not be found in genomic sequence, although these tags are difficult to map by any means due to slight variations in polyadenylation cleavage sites as discussed above. Finally, depending on SAGE tag length and the size of the genome, it is often not possible to uniquely map SAGE tags to genomic sequence. For instance, only 59% of *Drosophila* and *C. elegans* SAGE tags are unique in their genomes, even though these organisms only have genome sizes of 120 MB and 100 MB respectively (Pleasance et al. 2003). Statistically, it is expected that essentially every 14 bp SAGE tag will be found more than once in the human genome making tag mapping to genomic sequence infeasible (Saha et al. 2002). It is also important to include the mitochondrial genome in SAGE tag mappings, as mitochondrial genes produce a significant proportion of SAGE tags (Boon et al. 2002; Jones et al. 2001). Thus, genomic sequence is best used as part of an iterative process of tag mapping using multiple data sources (Gorski et al. 2003; Robinson et al. 2004).

1.5.4.4 Advantages, disadvantages, and applications of SAGE

SAGE is highly reproducible in the expression patterns it produces, and has reasonable concordance with other measurements of gene expression. SAGE libraries produced from the same RNA sample show correlations of 0.96 (Dinel et al. 2005). Similarly, MicroSAGE libraries constructed from the same RNA, the same ditag ligation, or the same pool of sequences in sequencing vectors, have correlations of between 0.94 and 0.98, with all of the differences due only to variability in tag sampling. As this sampling variability has a larger effect on tags of lower frequency, tags with counts under 100 show correlations closer to 0.8 (Blackshaw et al. 2003). Comparisons to RT-PCR and to oligonucleotide microarrays show correlations generally between 0.4 and 0.8 (Huminiecki et al. 2003; Ishii et al. 2000; Kim 2003). The biases and errors in SAGE discussed above can largely be accounted for by appropriate computational analysis, which contributes to the high reproducibility of SAGE. As SAGE does not depend on hybridization, issues such as cross-hybridization and saturation, which are more difficult to measure and resolve, do not influence observed expression measurements.

The ability of SAGE to determine transcript abundance is dependent on the depth to which the library is sequenced, which may typically be as few as 20,000-30,000 tags or as many as over 100,000 tags from a single RNA sample. SAGE is much more sensitive than EST sequencing for identifying low abundance transcripts (Sun et al. 2004), and at moderate library sizes it has similar detection efficiency to Affymetrix oligonucleotide arrays (Evans et al. 2002). Deep sequencing makes detection of very low abundance transcripts possible, which is particularly important in understanding complex tissues and transcriptional regulation (Boheler and Stern 2003). However, SAGE library construction is substantially more expensive than microarray profiling, especially when many tags are sequenced to produce large libraries, and

thus the number of libraries constructed is often more limited. This is offset by the increased utility of SAGE libraries for multiple experiments and comparisons. SAGE tag counts are absolute rather than relative as many hybridization-based methods are, and are not dependent on array design or prior knowledge of sequences. Thus, SAGE tag frequencies from different tissues or conditions can be directly compared, even across experiments. Additionally, as more genome and gene sequence information becomes available, tag-to-gene mappings can be updated and thus available SAGE data will be able to be reanalyzed and used for years to come.

Like ESTs and unlike most microarrays, SAGE does not require prior knowledge of the genes to be profiled, and thus it is an ideal method for novel gene identification. While the rate of novel sequence discovery by ESTs has decreased over time, a significant proportion of newly sequenced SAGE tags are unique and do not represent known genes (Boheler and Stern 2003). 35% of SAGE tags from mouse embryonic stem cells cannot be mapped to any known gene (Anisimov et al. 2002a; Anisimov et al. 2002b), suggesting there could be 16,000-39,000 novel genes and transcripts yet to be identified (Boheler and Stern 2003). Additionally, SAGE has been used to identify antisense transcripts (Quere et al. 2004), and the procedure modified to produce 5' SAGE tags for genome annotation (Wei et al. 2004). The ability of SAGE to profile expression of a gene is dependent on the presence of specific sequence motifs corresponding to endonuclease recognition sites, particularly of the commonly used anchoring enzyme *Nla*III and the occasionally employed *Sau*3A (Virlon et al. 1999) and *Rsa*I (Ryo et al. 2000). Additionally, gene profiling is dependent on unique representation of genes by SAGE tags. These issues are examined and discussed in Chapter 2.

Tens of millions of SAGE tags have been generated for hundreds of studies of normal and diseased tissues, cell lines, and model organisms. A limited number of studies have produced

Drosophila SAGE libraries, including experiments on the JNK signaling pathway (Jasper et al. 2001) and sex-specific differences (Fujii and Amrein 2002), as well as our work on programmed cell death (Gorski et al. 2003). SAGE libraries produced from multiple *Drosophila* tissues have been used to identify many novel low-abundance transcripts in *Drosophila*, demonstrating the utility of SAGE for gene identification (Lee et al. 2005). As is the case with microarray technologies, the SAGE method is commonly applied to studies of cancer. Many of these studies are collected in the Cancer Genome Anatomy Project (http://cgap.nci.nih.gov/) catalog of SAGE tags from hundreds of libraries representing close to 30 tissues, which also provides tools for identification of genes differentially expressed between normal and cancer samples (Boon et al. 2002).

1.6 Thesis objectives and chapter summaries

Programmed cell death is a complex cellular response that encompasses multiple pathways including apoptosis and autophagy. This response is essential in development and cellular homeostasis, but when misregulated can result in diseases including cancer. The death of the *Drosophila* salivary gland during development is an excellent model for studying PCD due to the precisely timed and complete nature of the destruction of this tissue; additionally, this death has hallmarks of both autophagy and apoptosis, and the genes involved are regulated at the transcriptional level. The mechanisms of programmed cell death, and the role of autophagy in PCD, are not completely understood, and identification of genes and pathways involved in these processes would contribute to our understanding both of PCD and its role in cancer. Genome-scale gene expression profiling and comparison of expression between species is a powerful approach that has the potential to identify such genes and pathways. Analysis of expression

datasets requires development of computational methods and approaches to examine the patterns of gene expression in cells undergoing cell death or tumorigenesis.

The primary aim of this thesis was to identify, using gene expression data, genes and associated pathways involved in programmed cell death in two systems, the *Drosophila* salivary gland and human cancer. All of the analyses described here are computational, involving development of methods, algorithms and databases appropriate to the assembly of data from disparate sources, and analysis for the purpose of extracting biological meaning. The first part of this analysis involved the development of methods to analyze gene expression data, particularly serial analysis of gene expression data, for the purpose of gene discovery and monitoring differential expression and also to define the limitations of the technique. Using these methods, I analyzed SAGE and EST data from the *Drosophila* salivary gland to identify genes expressed and regulated prior to programmed cell death in this tissue. The PCD-associated genes found in this analysis were then used to identify similar genes in the human genome that are differentially expressed in cancer, which have the potential to be involved in PCD and its regulation in oncogenesis. As both of these analyses pointed toward an important role for autophagy and associated processes in cancer, I subsequently examined the expression of an autophagy marker in various cancers, which indicated that autophagy is likely to be altered in cancer in a tissue-specific manner. A summary of these analyses as presented in Chapters 2 through 5 are given here.

In Chapter 2 (as published in Pleasance and Jones 2005; Pleasance et al. 2003), I describe an approach to constructing full-length "conceptual transcripts" from genome, EST, and gene prediction data, which I implemented in the *Drosophila* and *C. elegans* genomes. I hypothesized that using these transcripts constructed from multiple data sources would allow for more

effective assignment of SAGE tags to genes. The method I developed was indeed superior to previous tag-to-gene mapping methods available for *Drosophila*, and also allowed the assessment of SAGE for its ability to correctly, completely, and unambiguously identify genes. I demonstrated the effects in the *Drosophila*, *C. elegans*, and human genomes of varying the restriction enzymes used in SAGE that affect SAGE tag length and position. This analysis provides SAGE tag-to-gene mappings for these organisms, defines the limitations of SAGE, and suggests optimal choices of enzymes for a SAGE experiment.

In Chapter 3, I describe the analysis of an EST library and three SAGE libraries derived from developmental timepoints leading up to programmed cell death in the *Drosophila* salivary gland (Gorski et al. 2003). ESTs were clustered and assigned to genes, and novel ESTs identified. SAGE tags were assigned to genes using the method described in Chapter 2, and novel tags assigned to ESTs and genomic locations. Confirming the hypotheses that genes involved in PCD could be identified by gene expression analysis and that both apoptosis and autophagy genes are regulated in *Drosophila* PCD, genes which change in expression between SAGE libraries were found to include genes known to be part of the salivary gland PCD transcriptional hierarchy, genes belonging to other cellular pathways, and genes of no known function. These genes are potential regulators or effectors of programmed cell death in *Drosophila*.

Chapter 4 describes my approach based on the hypothesis that similar genes are involved in programmed cell death in cancer and in *Drosophila*, and that these genes can be identified based on a cross-species gene expression filter. Human orthologs of genes differentially expressed in the *Drosophila* salivary gland were identified, and the expression of these genes examined in multiple human cancers based on a database of SAGE data. Analysis of *Drosophila*

and human datasets identified a subset of genes which are regulated in both systems. These candidates for involvement in PCD in cancer were examined for known roles in PCD or cancer, and overall functions determined.

Chapter 5 (Pleasance et al, manuscript submitted) focuses on the analysis of the human LC3 autophagy gene in cancer based on multiple expression datasets, from both SAGE and microarray technologies. Consistent with the hypothesis that LC3 expression is altered in cancers, potentially reflecting the impact of autophagy on oncogenesis, differential expression of this gene was observed in multiple cancers compared to corresponding normal tissues. Differences in expression were also seen in cancer subtypes and stages of cancer progression. My observations suggest potential roles for the process of autophagy in cancer which are tissue-specific.

The bioinformatic methods used in these analyses can be applied to studies of other cellular processes, and the results suggest roles for multiple processes including autophagy in programmed cell death and cancer. This work contributes to our overall understanding of the functions of cell death in oncogenesis, and to the goal of a comprehensive view of the genetic changes that occur in cancer.

In addition to the work described in this thesis, I have been involved in several other collaborative projects at the Genome Sciences Centre which have been described in publications or submitted manuscripts. I worked with Obi Griffith and others to identify human genes which are coexpressed across many conditions by comparing and combining SAGE and microarray expression datasets, as described in Griffith et al (2005). I was also involved in work done by Peter Huang, examining the expression of genes in operons in *C. elegans* based on large-scale

GFP and SAGE data (Huang, Pleasance, et al, manuscript submitted). Finally, I am currently working on both the collection of known regulatory elements in a public database (Montgomery et al, manuscript accepted in *Bioinformatics*) and the discovery of novel regulatory elements in *Drosophila* based on genome sequence data as part of a pipeline and database system known as cisRED (Robertson et al, manuscript accepted in *Nucl Acids Res.*).

Chapter 2: Assessment and analysis of SAGE for transcript identification

A version of this chapter has been published.

Pleasance, E.D., Marra, M.A., and Jones, S.J.M. 2003. Assessment of SAGE in transcript identification. *Genome Res* **13**: 1203-1215.

Pleasance, E.D. and Jones, S.J.M. 2005. Evaluation of SAGE tags for transcriptome study. In: Wang SM, ed., *SAGE Technologies: Current Technologies and Applications*. Horizon Scientific Press, Norwich, UK.

2.1 Introduction

An essential step in Serial Analysis of Gene Expression (SAGE) is tag-to-gene mapping, or tag mapping, which refers to the unambiguous determination of the gene represented by a SAGE tag. At the time of the work described in this chapter, the importance of full-length, complete transcript sets for SAGE tag mapping had not been discussed in the literature. The primary tag mapping resource at the time of our analysis was SAGEmap (Lash et al. 2000), which makes use of EST and mRNA sequences, with the corresponding drawbacks as described in Section 1.5.4.3. Additionally, this resource was only available for mammalian and *Arabidopsis thaliana* sequences, while the SAGE technique had been increasingly applied to other organisms (Jasper et al. 2001; Jones et al. 2001; Steen et al. 2002). In particular, there was no publicly available tag mapping resource for *Drosophila*, and such a resource was necessary for our analysis of *Drosophila* SAGE libraries.

Many of the difficulties encountered when tag mapping based solely on expressed sequences could be addressed by using genomic sequence and annotation. As genomic sequence is more accurate, with a low estimated error rate of <1/10,000 (Adams et al. 2000; *C. elegans* Sequencing Consortium 1998), the potential for results to be obfuscated due to sequence errors in the tag mappings is reduced. Also, annotated gene sets are potentially more complete than expressed sequence sets alone due to the use of additional information such as genefinder predictions and exons detected through conserved protein similarity, and the curation of this data to form the annotation (FlyBase Consortium 2002; Stein et al. 2001). Basing tag mappings entirely on predicted genes from genomic databases is problematic, however, as gene annotations rarely include untranslated regions (UTRs) of expressed transcripts. Since SAGE tags will correlate to the 3'-most anchoring enzyme site, typically that of the four-cutter restriction

enzyme *Nla*III which occurs on average every 256 bp, many SAGE tags are expected to be derived from 3' UTR sequence.

It is essential when using the SAGE technique to be aware of the transcripts that will not be identifiable using this method, due to shared sequence resulting in ambiguity or due to lack of an appropriate anchoring enzyme site, as the expression of such transcripts will not be accurately determined. Identification and quantification of these refractory transcripts would potentially permit the choice of appropriate modifications to the SAGE procedure, which could minimize the number of transcripts that will not be profiled using SAGE and increase the utility of this expression profiling method. Thus, in addition to developing a tag mapping method and resource, I assessed for the first time the ability of SAGE to uniquely and completely identify transcripts.

To overcome some of the limitations in SAGE tag mapping and to assess the utility and restrictions of the SAGE approach in transcript identification, I devised a method for constructing a complete predicted transcript set ("conceptual transcripts") and deriving SAGE tags from it. These conceptual transcripts are based on a combination of genomic sequence, annotated predicted genes, and expressed sequences, and thus are sufficiently complete to allow this assessment. I applied this method to the model organisms *Drosophila* and *C. elegans*, as more mature genomic annotation resources were available for these organisms than for the human genome. Also, although the SAGE technique has been utilized in these organisms, there were at the time of our work no publicly available tag mappings. In addition to providing tag mappings for these organisms, I was able to determine the number of genes lacking a suitable anchoring enzyme site, and establish the extent to which SAGE tags will be correlated ambiguously to genes. I also determined the most efficient anchoring enzyme choice for each

particular organism. I applied a similar analysis, though more limited as it did not include construction of conceptual transcripts, to the human transcriptome. The method described here permits assessment of the efficacy of the SAGE approach in transcript identification by distinguishing genes refractory to this profiling method, and facilitates SAGE analysis in both *Drosophila* and *C. elegans* as well as in other organisms to which this method can potentially be applied.

2.2 Methods

2.2.1 ACEDB

An ACEDB (Durbin and Thierry-Mieg 1991) database is available for *C. elegans* (Stein et al. 2001) and contains a large variety of data including genomic, gene, expression, and sequence similarity information. This database is vital in the construction of the conceptual transcript set for tag mapping, as it can be automatically queried using existing software. It was accessed through a remote connection to WormBase (http://www.wormbase.org) version WS78 (May 2002). This database includes EST alignments determined by BLAT (Kent 2002).

The *Drosophila* databases GadFly (http://www.bdgp.org/annot/index.html) and FlyBase (2002), contain similar information, but do not allow the direct automated manipulation of data in the manner necessary for the integration of sequence information into conceptual transcripts. Accordingly, I constructed an ACEDB database for *Drosophila,* based on the GadFly resources, which contains data on the genomic sequence as well as predicted and known genes, and EST and cDNA alignments. This database is directly queryable in an automated manner. Genomic sequence and predicted gene coding sequence positions were obtained from the GadFly Release 2 mySQL database gadb2c on Apr 8 2002, accessible from headcase.lbl.gov. EST and cDNA sequences were obtained from the BDGP at ftp.fruitfly.org/pub/genomic/fasta/ on April 11 2002.

ESTs and cDNAs were aligned to genomic segments identified as having similarity by BLASTN (Altschul et al. 1990) (v. 2.0.14) (E value < 1e-100, word size 32) using the dynamic, intron-aware EST_GENOME (Mott 1997) alignment program (min. score 100). EST_GENOME output was converted into ACEDB format using custom Perl scripts. In addition, the gene represented by each of the ESTs was determined by BLASTN similarity (E value < 1e-3, word size 16) processed by MSPcrunch (Sonnhammer and Durbin 1994) (v. 2.3), requiring that an EST match no more than one gene locus. All other input/output processing was performed using custom Perl scripts. This *Drosophila* ACEDB database is available for download from http://sage.bcgsc.ca/tagmapping/. Besides its utility in constructing the conceptual transcript set, the *Drosophila* ACEDB database is useful for viewing gene structure, expression and similarity information. In these respects, it is similar to FlyBase's (2002) GeneSeen (http://www.flybase.org/annot/geneseen-launch-static.html). However, it has the added advantages that it can incorporate user-specific data and be queried directly or through the AcePerl (Stein and Thierry-Mieg 1998) interface to extract sequences and associated information, allowing further whole-genome analysis.

2.2.2 Transcript construction

Transcripts were constructed using custom Perl scripts for each organism which interact with the ACEDB database described above through the AcePerl modules (Stein and Thierry-Mieg 1998). UTRs were added to predicted genes using the genomic sequence as determined by alignments with ESTs and cDNAs for genes where such data is available. If expressed sequences were unavailable, UTRs were estimated to be a length that would encompass 95% of known UTR sequences based on empirical UTR size distributions and polyadenylation signals.

In *C. elegans*, UTRs were added based on EST evidence if the EST was assigned to the gene (in the Matching_cDNA field of the gene in ACEDB) and was in the correct orientation (e.g. only 3' ESTs used to construct 3' UTRs) and started within 1 kb of the end of the gene (for reasons of efficiency; UTRs in *C. elegans* are not expected to extend this far). UTRs were extended to encompass the furthest EST from the gene.

In *Drosophila*, UTRs were added based on EST/cDNA evidence if the EST/cDNA both overlapped with the gene's coding region (by EST_GENOME) and was more similar to that gene than any other (by BLAST). As *Drosophila* UTRs are occasionally spliced, breaks in the EST alignments corresponding to introns were excluded from the final transcript sequence. Only ESTs or cDNAs starting within 20 kb of genomic sequence from the end of the gene were considered. As in *C. elegans*, ESTs were required to be in the correct orientation.

For genes that did not have expressed sequence corresponding to their UTRs, 5' UTRs were extended to 836 bp in *Drosophila* and 388 bp in *C. elegans*, and 3' UTRs were extended to 1039 and 574 bp respectively. This corresponds to a length > 95% of the known UTRs as determined by EST/cDNA alignments. If the most common polyadenylation signal AATAAA (Graber et al. 1999; Riddle et al. 1997) was found in the estimated 3' UTR, the UTR was truncated 35 bp downstream of the end of this signal. Also, if there was a gene nearby, the estimated UTRs were truncated so as not to overlap ESTs associated with another gene nor extend over one-half the distance to the next nearest coding region, so as to prevent adjacent UTRs from overlapping.

A 30 bp poly-A sequence was added to the 3' end of all constructed transcripts to represent the poly-A tail present on mRNAs, as occasionally SAGE tags contain part of this poly-A sequence. 30 bp was chosen because it extends further than the longest SAGE tags (25

bp) used in this analysis, thus permitting SAGE tags to extend their full length into the poly-A sequence if necessary. SAGE tags derived from poly-A sequence will complicate SAGE analysis in any case. All such tags will end in a varying number of As, and thus will be more likely to be ambiguous. Also, the variance in polyadenylation cleavage sites will result in multiple tags derived from the same gene (Pauws et al. 2001).

Conceptual transcript sequences and corresponding tag mappings, to both gene loci and individual alternative transcripts, are available at http://sage.bcgsc.ca/tagmapping/. Perl code is made available upon request.

2.2.3 Evaluating tag-to-gene mapping accuracy

To evaluate the accuracy of tag-to-gene mappings derived from the conceptual transcripts, I used a set of "test" SAGE tags extracted from the 3'-most position of 6614 full-length *Drosophila* cDNA sequences (Stapleton et al. 2002) obtained from BDGP (http://www.bdgp.org/EST/index.shtml; entire available set as of April 11 2002). The accuracy of mapping was calculated as the percentage of tags that could be correctly and unambiguously assigned to genes, as this is the goal of tag-to-gene mapping. First, I attempted to assign each of the cDNA and 252,362 EST sequences (from BDGP sequencing project, http://www.bdgp.org/EST/index.shtml) to one of the 13,489 *Drosophila* predicted genes by BLASTN (E value< e-50); conceptual transcripts, as they are constructed based on predicted gene sequences, are already associated with genes. In order to correctly compare the accuracy of EST mapping approaches, only cDNAs and genes to which EST sequences could be mapped were considered. Also removed were ESTs that did not match predicted genes from genomic annotation. I thus used for the analysis 5606 SAGE tags extracted from cDNA sequences, 13,489 conceptual transcripts, and 204,380 EST sequences, each assigned to a gene.

Each of the sequence sets described in Table 2.1 were then used to map the "test" SAGE tags to genes. I constructed sets of conceptual transcripts with and without using EST or cDNA sequences to determine UTRs; if no ESTs or cDNAs were used, the UTRs were estimated as described above. In mapping to EST sequences, tags were assigned to ESTs which were grouped based on gene assignments so that clustering ESTs based on sequence similarity was unnecessary. Ambiguous mappings would only occur in cases where two ESTs from different genes contained the same SAGE tag. The number of tags assigned to a single gene using these various sequence sets and the number of those assignments that were correct were determined.

2.2.4 Choice of tagging and anchoring enzymes, tag length

All tags of a given size for a given enzyme site were extracted from all conceptual transcripts for all loci in the genomes using Perl scripts. A locus was considered not to contain an enzyme site if none of its alternative transcripts contained that site, and a locus was considered ambiguous if any of its transcripts shared a tag at the 3'-most enzyme site with a transcript from a different locus. If, for instance, a particular tag was found in the 3'-most site in one transcript and a site closer to the 5' end in another transcript, that was not considered to be an ambiguity and that tag would be assigned to the gene in which it was found at the 3'-most site. Probability of uniqueness was calculated from the formula used in (Saha et al. 2002), assuming 13,489 tags of 14 bp each.

2.2.5 Mapping experimentally derived tags

4007 different experimentally derived *Drosophila* tags (Gorski et al. 2003) and 9159 different *C. elegans* tags (Halaschek-Wiener et al. 2005), of at least 99% quality (1% chance of sequence error) and occurring more than once, were derived using software under development

in our laboratory (R. Varhol and S. Zuyderduyn, unpublished). All tags of 14 bp were extracted using Perl scripts from *Nla*III enzyme sites at every position in the *Drosophila* and *C. elegans* genomes derived from ACEDB. Experimental tags were compared to genomic tags, and the number of occurrences of each tag determined. The same libraries of experimental tags, with varying frequency cutoffs, were mapped to conceptual transcripts and the percentage mapped unambiguously and ambiguously determined.

2.2.6 Human tag mapping with RefSeq

Human nucleotide sequences from the RefSeq database were downloaded Oct 10 2002 from the NCBI FTP server (ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/). All 15,274 mRNA sequences (those beginning with NM_) then available were used in analysis. Tags were extracted from the 3'-most enzyme site, with varying length and varying enzymes as described in Section 2.2.4. To estimate the actual level of ambiguity based on the incomplete set of RefSeq sequences, random subsamples of the sequence set of varying sizes were chosen and ambiguity calculated for each subsample. This was done in triplicate and the average ambiguity for each subsample size determined.

2.3 Results

2.3.1 Tag mapping using genomic sequence

Mapping SAGE tags directly to genomic sequence could potentially deal with the issues of sequence quality and incomplete transcript sets. However, the genome sizes of complex eukaryotes are large enough that tag sequences may be present more than once by chance. To determine the impact of this issue, I mapped experimental *C. elegans* and *Drosophila* SAGE tags to the corresponding genomes, which are ~100 MB and ~120 MB respectively (Adams et al.

2000; *C. elegans* Sequencing Consortium 1998). Only 59% of mapped SAGE tags occur unambiguously in the *C. elegans* genome, and 20% of *C. elegans* SAGE tags occur 3 or more times in the genome. 59% of *Drosophila* SAGE tags are also unambiguous, and 18% occur 3 or more times in the genome. Also, only approximately 60% of SAGE tags map to the *Drosophila* or *C. elegans* genomes at all, due in part to some SAGE tags crossing splice boundaries and thus not being present in the genome. These results suggest that it is necessary to use transcript sequences rather than genomic sequence for SAGE tag-to-gene mapping.

2.3.2 Transcript construction and analysis

In order to organize and access the data required to produce the conceptual transcript set, I first constructed a queryable ACEDB (Durbin and Thierry-Mieg 1991) database containing genomic, gene, and expressed sequence information for *Drosophila* (see Methods). A similar database is already available for *C. elegans* (Stein et al. 2001). Using these databases, conceptual transcripts were constructed as described in Methods (Figure 2.1). In *Drosophila*, a total of 8213/14335 (57%) conceptual transcripts have 3' UTRs constructed based directly on expressed sequence evidence, with an average size of 343 bp and a median of 224 bp. The remaining transcripts have predicted 3' ends based on empirical size distributions (Figure 2.2A and B) indicating that 95% of 3' UTRs are 1039 bp or less. In *C. elegans*, the 6608/20448 transcripts (32%) found to have 3' UTRs constructed based on direct sequence evidence were an average of 195 bp and a median of 137 bp, while the remaining 3' UTRs were estimated based on empirical size distributions (Figure 2.2C and D) that show 95% of known UTRs to be 574 bp or less.

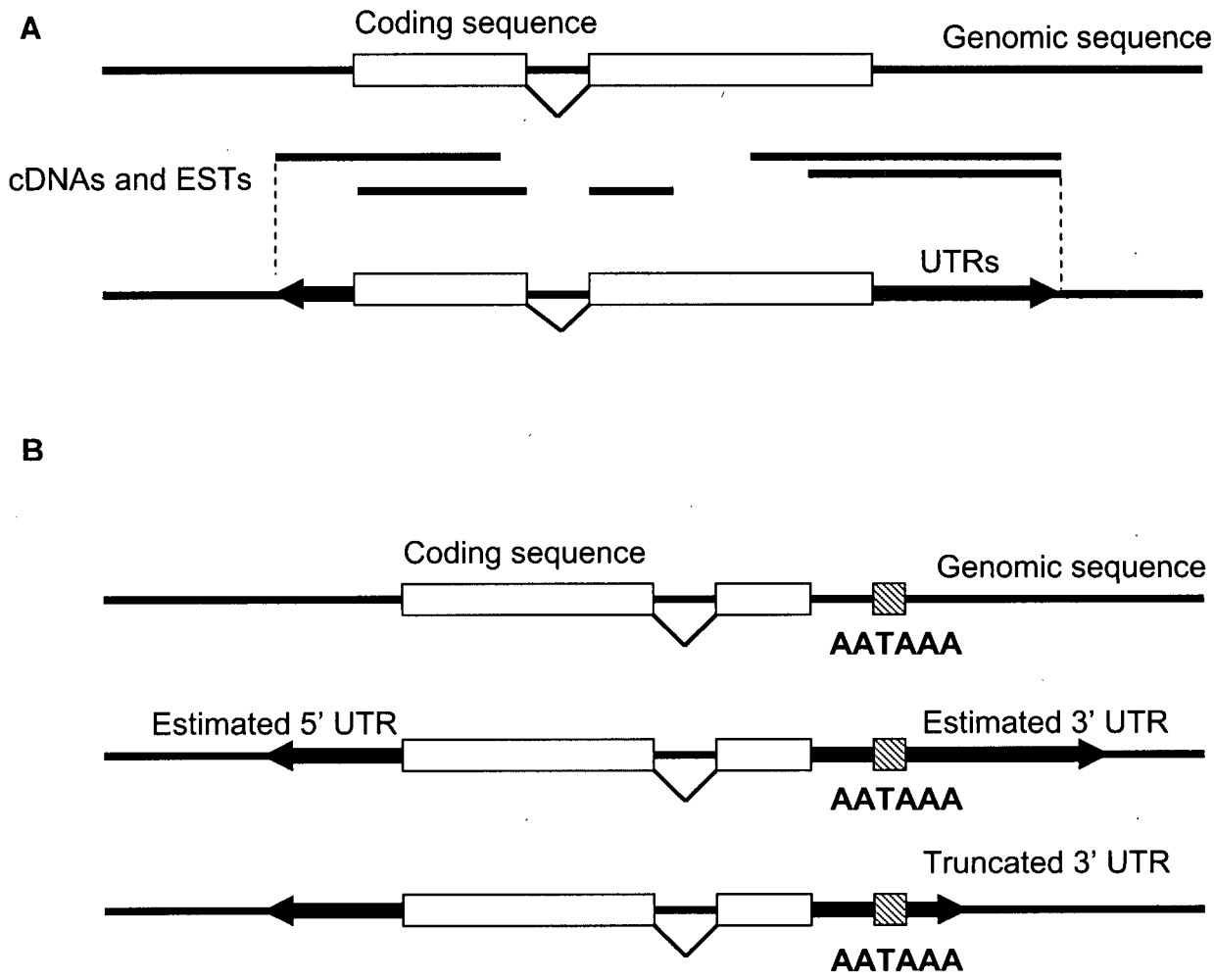

Figure 2.1 Conceptual transcript construction for SAGE tag mapping.
Genomic sequence was used to form conceptual transcripts, taking advantage of its high sequence quality. Coding sequences were derived from current genomic annotation. (A) If expressed sequences that extend beyond the predicted coding sequence were available, the alignment position of these sequences was used to determine the extent of the UTR. (B) If no expressed sequences were available, *Drosophila* 5' and 3' UTRs were extended to 836 bp and 1039 bp respectively, and *C. elegans* UTRs were extended to 388 bp and 574 bp. These UTR size estimates encompass 95% of known UTRs determined in (A). If the polyadenylation signal "AATAAA" was found within the estimated 3' UTR, the UTR was truncated 35 bp downstream of the signal.

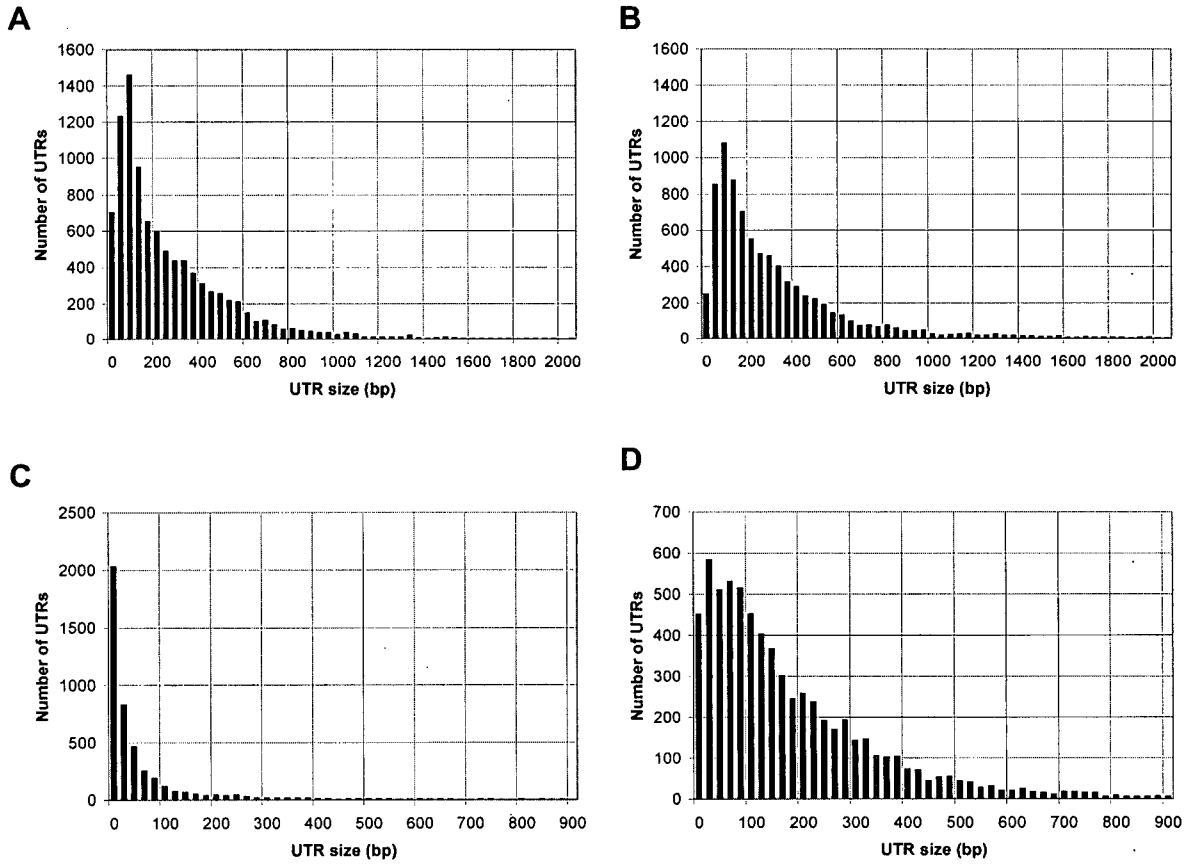

Figure 2.2 UTR size distributions in *Drosophila* and *C. elegans*.
Lengths of conceptual transcript UTRs were determined based on expressed sequence alignments (Figure 2.1). (A) *Drosophila* 5' UTR. (B) *Drosophila* 3' UTR. (C) *C. elegans* 5' UTR. (D) *C. elegans* 3' UTR.

2.3.3 Tag-to-gene mapping

Based on this conceptual transcript set, tag-to-gene mappings were derived by extracting all tags from the transcripts, that is, the 10 bp downstream of each *Nla*III (CATG) site for a total of 14 bp for each tag. It is important for this tag mapping method to consider all anchoring enzyme sites and not just the 3'-most, for two reasons. First, gene prediction programs often have difficulty correctly defining the ends of genes (Rogic et al. 2001), so that the apparent 3'-most tag site may not be correct. Second, estimated UTRs added when no expressed sequence is available are a length that would encompass 95% of known UTRs, and are thus known to be

overestimates of real UTR length. In many cases, therefore, the true 3' end of the transcript and the associated 3'-most SAGE tag will be upstream of the estimated one. SAGE tags are then assigned to the transcript containing the correct tag sequence. In cases where a SAGE tag is found in more than one transcript, we take advantage of the fact that the SAGE procedure is expected to derive the SAGE tag from the 3'-proximal position (Figure 1.5), and resolve this ambiguity where possible by assigning the correct tag as the one closest to the 3' end of the transcript. If a tag is found in the same relative position in more than one transcript, that tag cannot be resolved and is considered to be ambiguous. In this tag to gene mapping, tags derived from alternative transcripts are consolidated and assigned to a single gene locus, to avoid labeling unique assignment to a single locus "ambiguous" when a tag matches two alternative transcripts. This also means that the same gene may be represented unambiguously by two or more different SAGE tags from different alternative transcripts.

These tag mappings, as they are derived from accurate (<1/10,000 error rate) (Adams et al. 2000; *C. elegans* Sequencing Consortium 1998) genomic sequence and incorporate expressed sequence information with the predicted transcriptome, are expected to be more complete than mappings from expressed sequences alone. To assess our tag mapping method, a test set of 5606 tags were extracted from sequences of full-length cDNAs, which are assumed to be accurate representations of true expressed transcripts. The accuracy of each mapping method, including mapping to ESTs or to conceptual transcripts, was determined by its ability to correctly assign tags to genes from this validated test set (Table 2.1). As relatively few full-length cDNAs are available for *C. elegans*, this comparison was done for *Drosophila* only, although the results are expected to be similar for both organisms. I found that conceptual transcripts constructed using EST data in conjuction with gene predictions produced significantly more accurate tag mappings

(85% correct) compared to EST sequences alone (70% correct). There were also slightly more tags assigned ambiguously when mapping to EST sequences alone (6% of tags compared to 4%), most likely to due sequencing errors in the EST sequences that produce extra, erroneous tags, or due to chimeric ESTs. Simulating the situation in which a set of genes have no associated ESTs and thus only predicted protein coding genes from genomic annotation are available for tag mapping, we found that tag mappings derived from such sequences alone (48% correct) were significantly less accurate than tag mappings from conceptual transcripts constructed with estimated UTRs (81% correct). This is not unexpected, given that 56% of SAGE tags extracted from *Drosophila* cDNAs are derived from the UTR sequence, which is lacking in the predicted protein coding sequences. The high proportion of tags derived from UTR sequence emphasizes the importance of estimating UTRs when no expressed sequence evidence is available. Overall, between 81% and 85% of tag mappings derived from conceptual transcripts are correct when full-length cDNA data is not incorporated, a significant improvement over tag mapping with ESTs or genomic annotations alone. When full-length cDNA data is incorporated into the conceptual transcripts, as is normally the case where possible, 93% of test tags are mapped correctly (Table 2.1); the remaining incorrectly mapped tags are due to gene prediction errors.

Table 2.1 SAGE tag mapping accuracy using conceptual transcripts compared to other available sequences.

Test tags assigned to genes by mapping to ...	Tags mapped unambiguously	Unambiguously mapped tags correctly assigned to genes
Full-length cDNAs (5606)	99%	100%
EST sequences (204,380)	94%	70%
Predicted protein coding sequences from genomic annotation (13,489)	97%	48%
Conceptual transcripts constructed from predicted protein coding sequences from genomic annotation, UTRs estimated (13,489)	96%	81%
Conceptual transcripts constructed from predicted protein coding sequences from genomic annotation, UTRs derived from EST data or estimated if ESTs unavailable (13,489)	96%	85%
Conceptual transcripts constructed from predicted protein coding sequences from genomic annotation, UTRs derived from EST and cDNA data or estimated if ESTs and cDNAs unavailable (13,489)	96%	93%

2.3.4 Analysis of transcriptomes with SAGE

As the conceptual transcript sets for *Drosophila* and *C. elegans* are based on essentially complete genomic sequence, representing all known and predicted genes, quantification of the genes that are not amenable to SAGE in these organisms is possible. For this analysis, alternative transcripts were considered to be a single gene, so as not to erroneously assign tags as ambiguous if they are derived from multiple alternative transcripts belonging to the same locus, as described above. Tag mappings are also available (see Methods) that assign tags to individual alternative transcripts.

As described above, genes that produce ambiguous tags, shared with other genes, will not be uniquely identifiable in a SAGE library. 6% of *Drosophila* genes and 12% of *C. elegans* genes fall in this category, in the common situation of 14 bp SAGE tags extracted using the *Nla*III anchoring enzyme. As recent work has yielded a SAGE procedure that produces a 21 bp

tag (Saha et al. 2002), it is relevant to ask how SAGE tag length influences the ambiguity of the tag mappings for each organism. We observe (Figure 2.3A and B) that increasing SAGE tag length by 2-3 bp decreases ambiguity in tag assignments, after which increasing length has little effect. This information can be used to make an informed decision about the ideal tag length for a particular SAGE experiment.

One drawback of the SAGE technique is that genes that lack the appropriate anchoring enzyme recognition sequence will not be represented in SAGE libraries. For instance, genes lacking a "CATG" site will not be represented in SAGE libraries constructed with *Nla*III. However, judicious choice of anchoring enzyme could improve the utility of the SAGE approach. Based on the conceptual transcript set, there are 261 genes (2% of the annotated transcriptome) lacking *Nla*III sites in *Drosophila* and 563 (3%) in *C. elegans*. However, when tags are extracted with other four-cutter anchoring enzymes, differing numbers of genes contain no anchoring enzyme site or produce ambiguous SAGE tags (Figure 2.4A and B). *Sau*3A, which has occasionally been used in SAGE library construction (Virlon et al. 1999), allows recovery of more genes unambiguously in *Drosophila* and *C. elegans* than *Nla*III. Interestingly, however, at higher tag length in *Drosophila*, *Nla*III has this property (Figure 2.3). It is important to consider, however, that *Nla*III is compatible with the *Bsm*FI tagging enzyme in such a way that an extra base pair of tag length can be obtained, resulting in the potential for a 15 bp rather than 14 bp tag (see Section 1.5.4.2). The *Aci*I (CCGC) enzyme, which has not been used in SAGE library construction, also has the same property and allows more genes to be resolved in *Drosophila* than *Nla*III. This knowledge allows the preselection of an anchoring enzyme, which may be organism-specific, which produces the best representation of the expressed genes within an mRNA population.

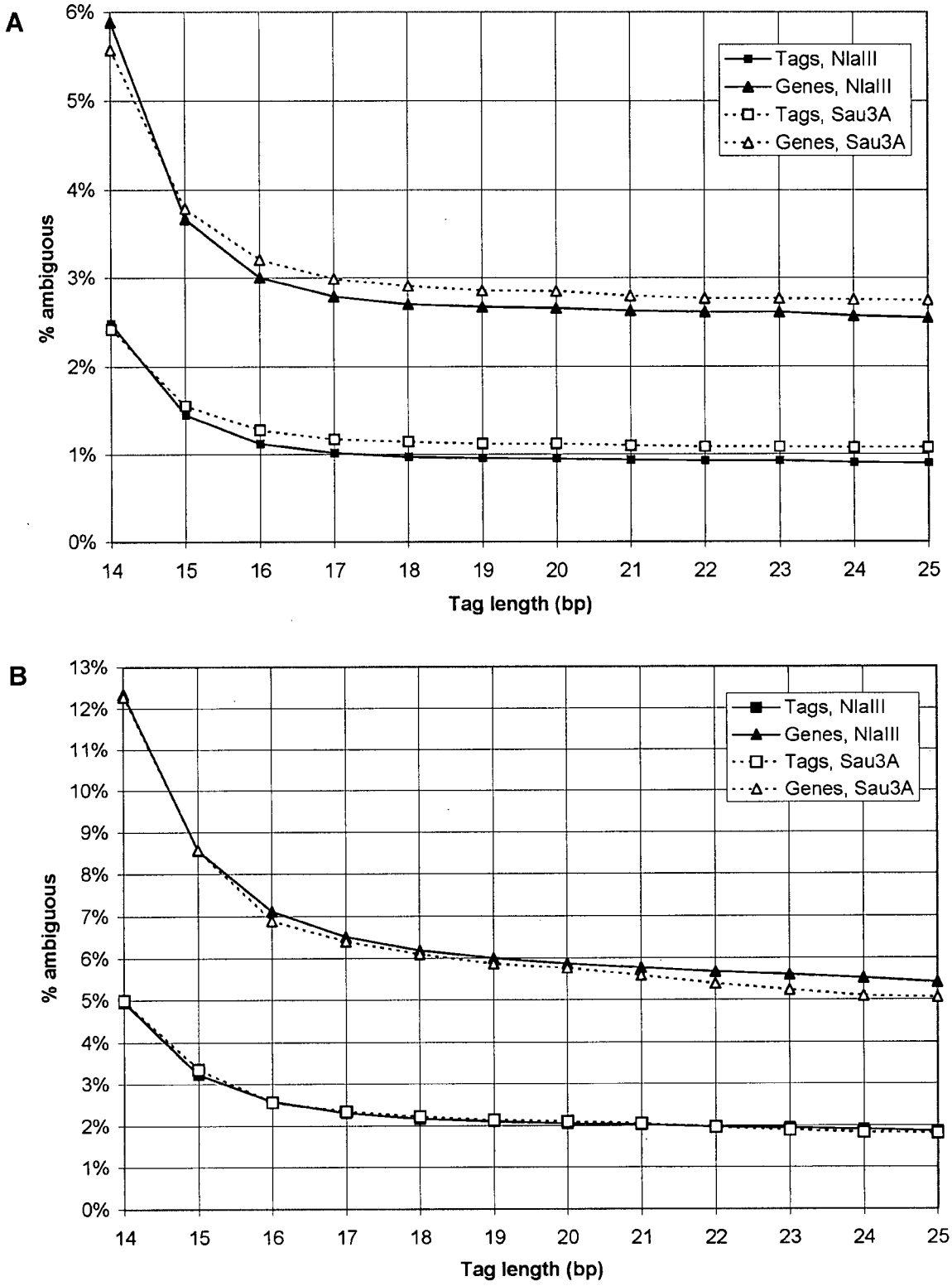

Figure 2.3 SAGE tag ambiguity varies with tag length.
Number of ambiguous genes and number of ambiguous tags derived from conceptual transcripts is shown with varying tag length (length includes 4 bp anchoring enzyme site) and anchoring enzyme. (A) *Drosophila*, total 13,489 gene loci (B) *C. elegans*, total 19,432 gene loci.

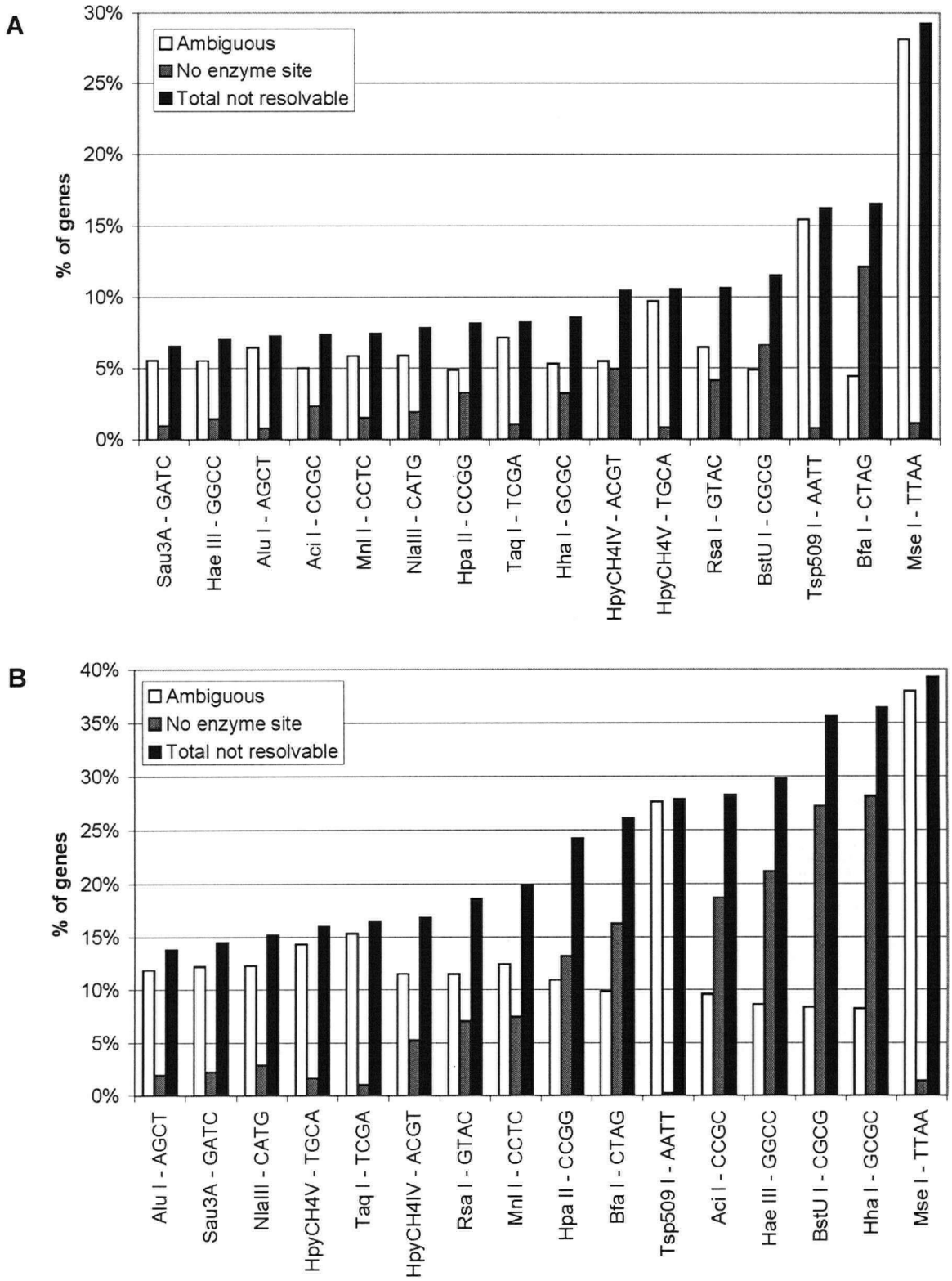

Figure 2.4 Number of genes not resolvable by SAGE varies with anchoring enzyme.
Number of ambiguous genes and genes with no anchoring enzyme site are shown for various restriction enzymes used as the anchoring enzyme. (A) *Drosophila*, total 13,489 gene loci (B) *C. elegans*, total 19,432 gene loci.

2.3.5 Mapping experimentally derived tags

To further analyse the tag-to-gene mappings and compare predicted levels of ambiguity with experimental levels, I mapped *C. elegans* and *Drosophila* experimental SAGE tags using this method of tag-to-gene mapping (Figure 2.5). In this comparison, the number of ambiguous genes cannot be compared, as we do not know how many genes are represented by an ambiguous experimental SAGE tag. Instead, the number of SAGE *tags* that are ambiguous is compared. In both organisms, an even higher proportion of experimental SAGE tags map ambiguously to genes than expected. For *C. elegans*, 5.0% of 14 bp tags were predicted to be ambiguous (Figure 2.3B), and 7.5% of experimental tags were ambiguous; for *Drosophila*, the predicted (Figure 2.3A) and experimental ambiguities were 2.5% and 3.6% respectively. It is also notable that the proportion of tags that are mapped to genes increases with expression level, for both organisms.

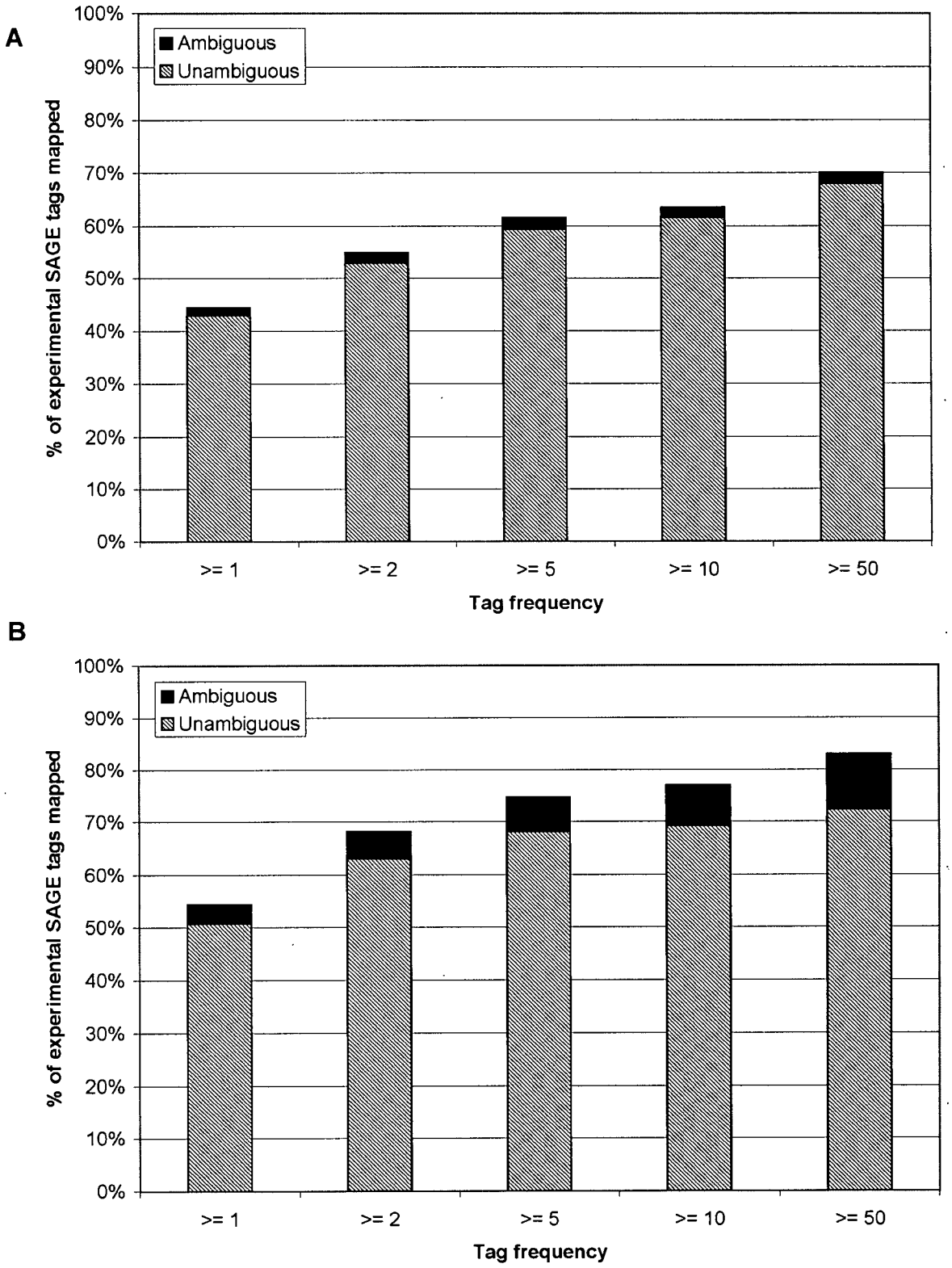

Figure 2.5 Experimental SAGE tags mapped to conceptual transcripts.
Subsets of experimental SAGE tags with varying minumum frequency were mapped to conceptual transcripts, and the percentage that mapped ambiguously and unambiguously determined. (A) *Drosophila* (B) *C. elegans*.

2.3.6 Assessment of human tag mapping with RefSeq

A complete gene set is not available for the human transcriptome, and thus SAGE tag ambiguity cannot be directly determined as for *Drosophila* and *C. elegans*. However, ambiguity can be estimated from the incomplete set of full-length, high-quality sequences in the RefSeq database by examining the increase in ambiguity with the number of sequences in a transcriptome (Figure 2.6A). 5000 sequences produce ~4.5% ambiguous tags, while the full set of 15,274 sequences produces ~7.5% ambiguous tags. Extrapolation from this trend suggests that 10% of 14 bp tags will be ambiguous in a set of 30,000 transcripts, 15% in 50,000 transcripts, 22% in 75,000 transcripts, and 27% in 100,000 transcripts. The actual transcriptome size is not yet known. As expected, increasing tag size decreases SAGE tag ambiguity, by approximately 50% for a 21 bp tag (Figure 2.6B), and less thereafter. Less than 1% of human RefSeq sequences lack an *Nla*III site, compared to nearly 3% which lack a *Sau*3A site, and *Sau*3A also derives tags with higher overall ambiguity (Figure 2.6C). In fact, of all the restriction enzymes that are compatible in recognition sequence with the *Bsm*FI tagging enzyme (see Section 2.3.4), *Nla*III appears to allow the most human genes to be resolved by SAGE.

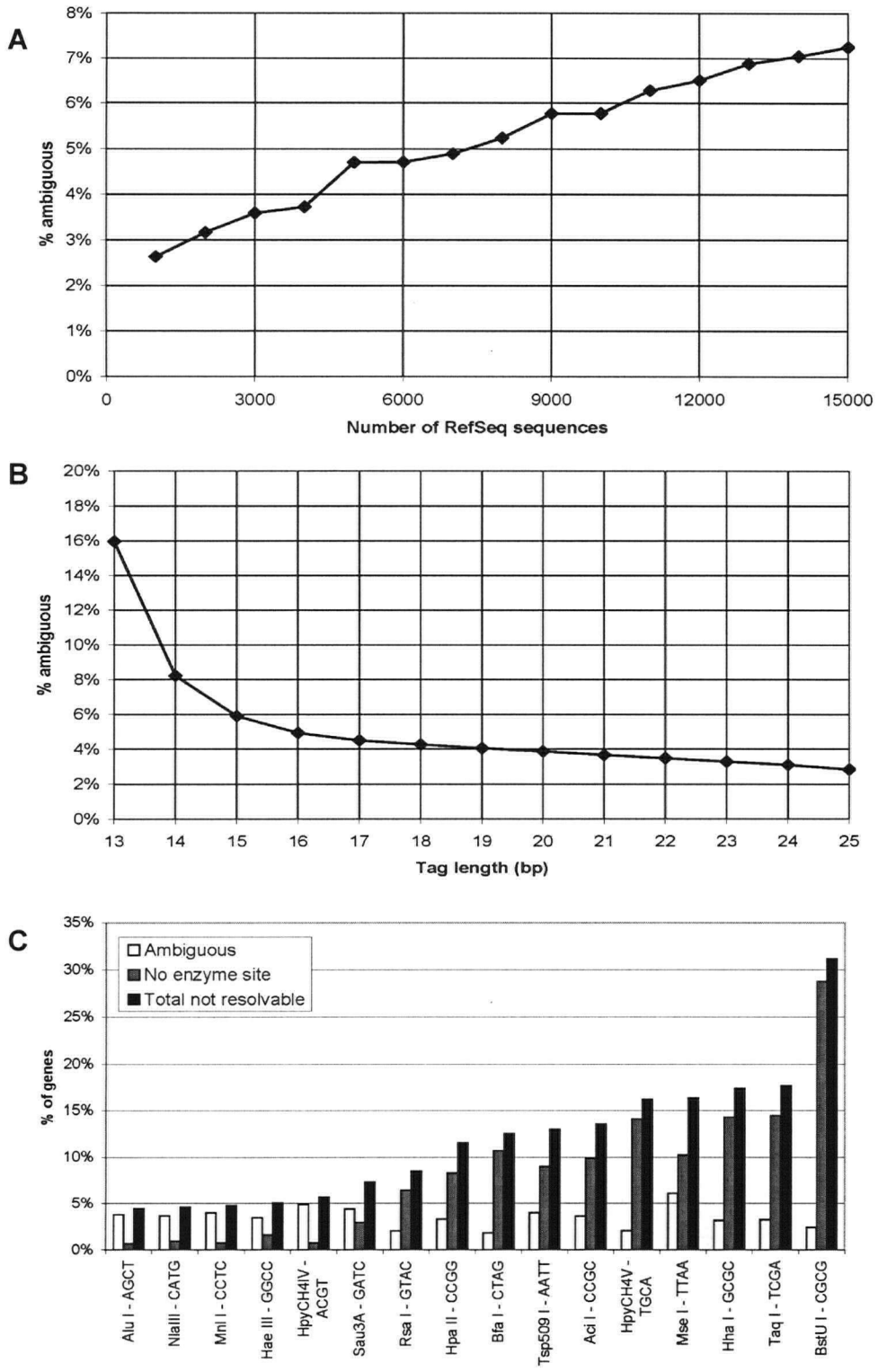

Figure 2.6 SAGE tag mapping with RefSeq.
(A) Effect of transcriptome size on unique SAGE tag assignment. Varying size random subsets of RefSeq sequences were sampled and the percentage of genes producing ambiguous 14 bp tags determined. (B) Percentage of 15,274 RefSeq transcripts that produce ambiguous SAGE tags. (C) Percentage of RefSeq transcripts that produce ambiguous tags or have no anchoring enzyme site for various choices of anchoring enzyme.

2.4 Discussion

This work has shown that utilizing genomic annotation provides a more accurate strategy for the assignment of SAGE tags to gene transcripts. I have also estimated the total resolving power of the SAGE approach, determining that 93% and 86% of genes can be detected in *Drosophila* and *C. elegans* respectively, when optimum anchoring enzymes are utilized. Extrapolations for the human transcriptome suggest that between 75% and 90% of genes can be resolved by SAGE. These findings suggest that the use of *Nla*III, most commonly used in SAGE experiments, is most advantageous for studies in human cells but is potentially sub-optimal for other organisms.

2.4.1 Ambiguity in SAGE

The observed rate of SAGE tag ambiguity is higher than would have been expected based on statistical calculations that predict 13,489 genes would produce tags with a 98.7% probability of being unique (see Methods). The observed increase in ambiguity is likely due to the non-independent nature of gene sequences, as there is similarity within gene families and even between distantly related genes. In addition, the presence of repetitive elements in the 3' UTRs of genes can contribute to increased rates of ambiguity. The 6% increase in ambiguity seen in *C. elegans* compared to *Drosophila* may be due to a combination of increased gene number and expansion of gene families (Friedman and Hughes 2001). Similarly, increasing transcriptome size as determined by subsampling of human sequences yields higher levels of ambiguity. The same factors of non-independent and repetitive sequence are likely to be responsible for the higher than expected rates of over 40% ambiguity observed when mapping directly to genomic sequence, which makes such an approach less feasible even for relatively small genomes of ~100

MB. For the human genome of 3,000 MB, 14 bp tags cannot be usefully mapped to the genome, as even random sequence would result in each tag occurring by chance 10 times in the genome.

The ambiguity predicted based on the conceptual transcripts is likely to be an underestimate of the true ambiguity. This is demonstrated by the higher than predicted proportion of ambiguous tags seen in experimental data. There are two factors that contribute to this. First, the genomic annotations remain under constant revision and our current understanding of both transcription and gene prediction suggest that complete transcript predictions will not be available for some time. Thus, the total gene count may be an underestimate as not all genes have necessarily been identified. EST data in *Drosophila* suggests there may be as many as 10-20% more genes than currently predicted (Andrews et al. 2000; Gorski et al. 2003), and 5% of full-length cDNAs do not match known or predicted genes (data not shown). Increased gene number increases the chance of shared sequence and thus shared SAGE tags. Also, increasing numbers of alternative transcripts, which are poorly predicted by genefinders and are constantly being discovered based on expressed sequences, can also increase ambiguity if an alternative transcript produces a tag that is already present in a transcript from a different locus. Second, 43-68% of 3' UTRs predicted in this analysis are estimated and thus are likely to be longer than the actual UTRs, and so will extend into less conserved intergenic sequence. It is expected that this sequence, as it is more random, will be less likely to be shared between genes, and thus less likely to yield ambiguous tags. As this hypothesis would predict, updating the transcripts with increasing amounts of expressed sequence data has resulted in shorter UTRs and increased ambiguity as more tags are derived from conserved sequence (data not shown).

The increase in ambiguity with expression level that occurs in *C. elegans* is likely because ambiguous tags have the potential to represent the sum of expression of several genes,

thus increasing the observed tag frequency. For instance, the tag "AAAAAAAAAA" which can be derived from the poly-A tail of many transcripts is seen at a relatively high frequency in most SAGE libraries (data not shown). Such an increase may not be visible for *Drosophila* in Figure 2.5 because of the lower overall ambiguity, and fewer tags available with high frequencies in the libraries under study (less than 200 tags with expression > 50 tags).

2.4.2 Choice of tag length, anchoring, and tagging enzyme

Based on Figure 2.3, the ideal tag lengths are approximately 16 bp for *Drosophila* and 17 bp for *C. elegans*. It is at these points that the curves of ambiguity vs. tag length level off and little is gained from further tag length. The residual level of ambiguity is most likely due to high similarity within gene families and repeats in 3' UTRs that result in sequences identical over regions > 25 bp. This suggests that approximately half of the ambiguity seen for 14 bp tags is due to random production of identical tags by dissimilar genes, which can be distinguished by longer tags. Human sequences similarly produce tags that are less ambiguous at higher tag lengths. However, unlike the curves of *Drosophila* and *C. elegans*, the human curve in Figure 2.6 does not level off completely, suggesting that tags of 21, 25 bp or longer continue to allow an increase in resolving power.

SAGE procedures that produce tags of 14 bp (Velculescu et al. 1995) and 21 bp (Saha et al. 2002) are commonly in use, and the SuperSAGE procedure produces tags of 26 bp (Matsumura et al. 2003). Although the currently described procedures cannot produce tags of arbitrary length, it is possible to modify any of these methods to produce shorter tags. Shorter tags may be desirable for organisms such as *Drosophila* and *C. elegans* where tags of 21 bp do not provide more resolving power than tags of 16 or 17 bp, and are more error-prone and expensive to produce as discussed below. Normally, the linker sequence used in SAGE consists

of a PCR primer attached to the tagging enzyme recognition site, where the recognition site is placed at the far 3' end of the linker to maximize the resulting tag length (Figure 2.7A). However, if one or more bases are appended to the 3' end of the linker, the tagging enzyme site will be further from the tag sequence and thus a shorter tag will be extracted (Figure 2.7B), allowing tags of arbitrary length up to 26 bp to be extracted. An important additional factor to consider when comparing SAGE tag lengths is the crucial role of longer SAGE tags in novel gene discovery, as longer tags can be mapped directly to the genome. 99.8% of 21 bp tags are expected to occur only once in the human genome, though mappings indicate that only 70% of tags are unique in practise, due to repeats and related genes. Even this lower proportion is still practical, and allows tag mapping to be done independent of the cDNA sequences available in current databases.

Approximately 7.5% of human RefSeq full-length mRNA sequences produce ambiguous 14 bp SAGE tags, while 4% produce ambiguous 21 bp tags. As RefSeq is incomplete, these levels are underestimates; extrapolation from ambiguity levels in smaller sets of sequences suggests that ambiguity is between 10% (in a transcriptome of 25,000) and 27% (in a transcriptome of 100,000). These results are in line with those published after our work, estimating that 9% of RefSeq sequences produce ambiguous 14 bp *Nla*III-derived SAGE tags, but increasing to a tag of 34 bp only decreases ambiguity to 6% (Unneberg et al. 2003). In contrast to tags extracted from RefSeq, tags extracted from UniGene clusters appear to have extremely high ambiguity, as 89% of 14 bp tags are ambiguous, and even tags of 39 bp are 80% nonunique (Clark et al. 2002; Lee et al. 2002b). Using only high quality non-EST sequences in this mapping decreases ambiguity to 56%, still significantly above the RefSeq estimates. The actual number of unique genes is likely to be somewhere between these estimates, as RefSeq

represents only a fraction of the transcriptome and so will underestimate ambiguity, while UniGene, due to problems with clustering and other artifacts, may overestimate ambiguity. An intermediate level of ambiguity along the lines of our estimates is also suggested by a model incorporating sequence errors and the nonrandom nature of DNA sequences, which predicts 6% of 14 bp tags will be ambiguous in a set of 15,000 sequences, similar to the size of RefSeq, while 25% will be ambiguous in a set of 78,600 sequences, more similar to the size of UniGene (Stollberg et al. 2000). The true level of ambiguity will be dependent on the actual number of genes transcribed.

It is important also to consider that increasing tag length decreases the efficiency of sequencing SAGE tags due to both longer SAGE tag length and thus fewer tags per sequence read, and also an increased per-tag error rate. At 1% sequence error, 13% of 14 bp tags are expected to contain sequence errors, while 21 bp tags will be erroneous in 19% of cases. A more recent detailed analysis of sequencing errors in SAGE data has indeed shown that 17.3% of LongSAGE tags contain one or more incorrect bases due to sequencing and PCR amplification errors (Akmaev and Wang 2004). Thus, for *Drosophila* and *C. elegans*, a tag length of 14 bp is probably an efficient and cost-effective choice, unless a more complete snapshot of the transcriptome is desired or a particular gene or set of genes of interest is identified as ambiguous with a 14 bp tag and not with a longer tag. Alternatively, if a tag of particular interest is found to be ambiguous, it can be resolved with further experimental procedures such as GLGI (Chen et al. 2000) which derive a larger portion of the expressed sequence represented by a SAGE tag allowing more likelihood of unambiguous gene assignment.

As shown in Figure 2.4 and Figure 2.6A, there can be a decrease in the number of genes that cannot be accurately analyzed using SAGE depending on the anchoring enzyme chosen. The

number of genes that are not unambiguously identifiable with a 14 bp tag can be decreased by 12% (883 vs. 1001) in *Drosophila* and 5% (2823 vs. 2962) in *C. elegans* by using *Sau*3A instead of *Nla*III. Conversely, *Nla*III performs 37% better than Sau3A in the human transcriptome (629 vs. 1001 not unambiguously identifiable). Other work has since confirmed our findings that less than 1% of human genes represented in RefSeq do not contain *Nla*III sites and 3% do not contain *Sau*3A sites (Boon et al. 2002; Unneberg et al. 2003). Notably, 7% of RefSeq genes do not contain *Rsa*I sites, an enzyme used in an alternative SAGE procedure (Ryo et al. 2000), suggesting this enzyme is not a good choice. Making use of the extra base pair of tag length provided by *Nla*III by using a 15 bp tag in analysis would further increase the effectiveness of *Nla*III. This would not involve a change in the SAGE procedure, only a change in the way in which SAGE tags are extracted from raw sequence derived from serially ligated tags. The tradeoff of such an analysis would be the loss of the small proportion of tags which, due to the slight variation in cutting position of *Bsm*FI, would only be 14 bp even with *Nla*III. My analysis of the effects of the choice of anchoring enzyme allows the potential for tailoring of the SAGE procedure to produce the most comprehensive results for each organism individually.

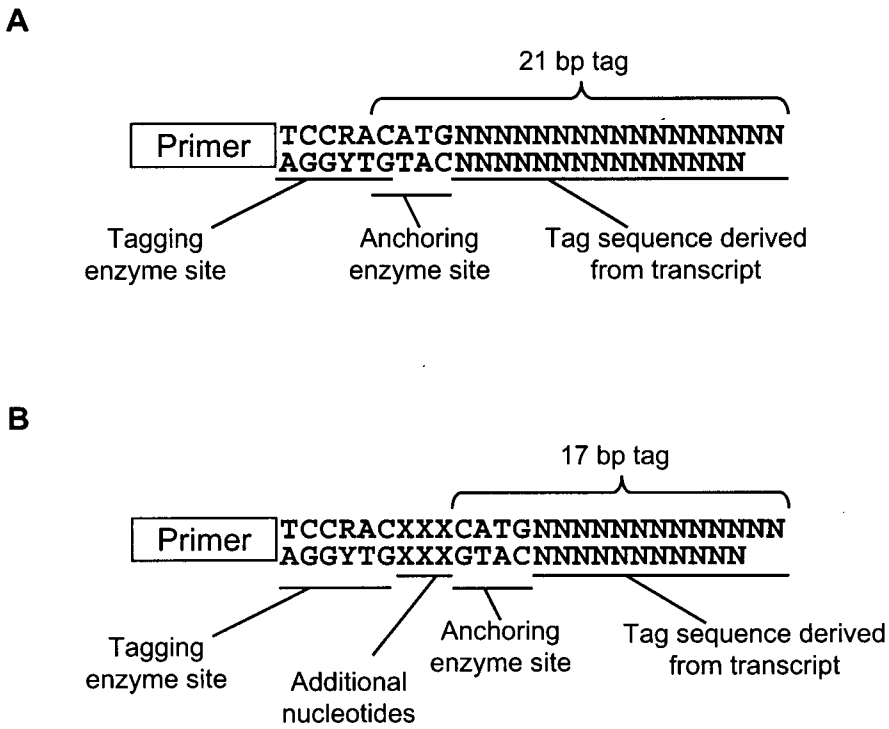

Figure 2.7 Creating shorter LongSAGE tags.
(A) Current LongSAGE procedure produces tags of 21 bp (Saha et al. 2002). (B) Designing the linker sequence with the appropriate number of nucleotides added to the 3' end could produce shorter tags of arbitrary length under 21 bp.

2.4.3 Tags with no gene mapping

It is notable that the proportion of experimental SAGE tags which cannot be mapped to genes or genomic sequence is significant (Figure 2.5). This is due in part to the presence of sequencing errors in experimental SAGE data. In Figure 2.5A, 55% of all experimental SAGE tags are mapped to genes, whereas 68% of tags that occur 2 or more times are mapped. This suggests that a significant proportion of the unmapped tags are singletons (occur only once) in the SAGE library under consideration. A similar trend occurs in Figure 2.5B. Singleton tags are more likely to represent sequencing errors, which generally produce tags that do not match

known sequences. In rare cases, a sequencing error may produce a tag that, by chance, matches another gene, which can confound analysis. Using more sequences for tag-to-gene mapping, such as a large set of EST sequences or clusters, increases the chance of this occurrence because there is a larger body of sequence to which a match may occur. Using the conceptual transcripts for mapping makes this unlikely, and thus a significant proportion of infrequently occurring, unmapped tags may be due to sequencing errors. Many of these errors can be removed by using only SAGE tags derived from sequence of high quality as determined, for instance, by Phred (Ewing et al. 1998). Techniques have also been proposed (Akmaev and Wang 2004; Velculescu et al. 1999) to remove tags that are likely to represent sequence errors of more highly expressed tags as discussed in Section 1.5.4.2. These methods can help to limit the effect of sequence errors in SAGE analysis.

The general increase in proportion of mapped tags with increasing expression level may also be related to the quality of genome annotation. On average, highly expressed genes are easier to study and survey and so are more likely to be known and thus included in the predicted gene set, while rarely expressed genes may not be annotated. Quality of annotation may in part explain why fewer tags are assigned to genes in *Drosophila* than *C. elegans* in Figure 2.5, as the *C. elegans* genome has been available for a longer period and therefore may have more correct annotations. It is somewhat surprising that the proportion of non-singleton tags mapped to genomic sequence (61%) is even lower than the proportion mapped to genes (68% in *C. elegans*). This may be because SAGE tags that cross splice boundaries will not exist in genomic sequence.

Another reason that experimental SAGE tags may not be mapped to transcripts or genomic sequence is the presence of polymorphisms, especially single nucleotide

polymorphisms (SNPs). The *Drosophila* and *C. elegans* strains used to produce the SAGE data described here are slightly different or derived from those used for genomic sequencing, and so there are expected to be base pair changes in a fraction of SAGE tags that prevent exact matches to genomically derived sequence. The same polymorphism effect can be expected in humans. Large bodies of EST sequence from multiple strains or multiple individuals are likely to represent these polymorphisms; however, it is very difficult to separate true polymorphisms from EST sequence and clonal errors which can cause erroneous mappings as discussed above, and thus this issue is not simplified by the use of EST sequences for tag mapping. The polymorphism rate in most experimental situations is relatively low, thought to be below 1/1000 bp in human sequences (Lander et al. 2001; Venter et al. 2001). However, more recent work suggests a polymorphism rate as high as 1/100 to 1/300 bases, and an analysis using dbSNP to map polymorphisms to transcripts and SAGE tags indicates that 8.6% of human genes may produce tags with sequence variations due to SNPs (Silva 2004).

We find that even after comprehensive mapping to the known transcriptome, 15-30% of highly expressed (>50 copies) SAGE tags are not identified, which cannot be accounted for by sequence errors and polymorphisms. This emphasizes the current incompleteness of genomic annotation and underscores the role that SAGE can play in novel gene discovery.

2.4.4 Limitations of conceptual transcripts

As the conceptual transcript sets rely heavily on curated sets of known and predicted genes, errors and omissions in gene predictions result in inaccuracies and incompleteness in the tag mappings. For instance, 7% of tags from known *Drosophila* cDNAs in the test set were not mapped correctly, due to gene prediction errors. This is mostly likely due to the latency in the process of incorporating experimental data into the curated genomic databases. It is anticipated

that as gene predictions are updated, refined, and consolidated with more expression data, the constructed transcript sets will become a more accurate reflection of the true transcriptome and thus yield more accurate tag mappings. For this purpose, the full-length cDNA sequencing projects currently being carried out in multiple organisms have positive implications for SAGE analysis.

2.4.5 Conclusions

The conceptual transcript sets and tag mapping procedure described here provide tag-to-gene mappings for the *Drosophila* and *C. elegans* genomes, thus facilitating SAGE analysis in these organisms. They also permit analysis of the limitations of SAGE for transcript identification. Without complete transcript sets, incomplete sets of full-length sequences such as the RefSeq database of human transcripts can be used both for tag mapping and to estimate levels of ambiguity. This analysis allows researchers to make informed decisions as to the SAGE procedure most appropriate for the organism under study and for the applications of the research.

Since this research has been published, an equivalent method utilizing UTR length distributions, gene predictions, and genome sequence to create conceptual transcripts has been used for SAGE tag mapping and SAGE assessment in *Arabidopsis* (Robinson et al. 2004). This demonstrates our method can be usefully applied to other model organisms.

Cheaper and more accurate sequencing has already decreased significantly the cost and errors involved in SAGE library construction. Genome annotations are constantly being improved and, in combination with full-length cDNA sequencing projects, will result in more reliable tag-to-gene mappings. Thus, SAGE is likely to continue to be a significant tool for gene expression analysis and genome annotation.

Chapter 3: SAGE and EST analysis of PCD in the *Drosophila* salivary gland

A version of this chapter has been published.

Gorski, S.M., Chittaranjan, S., Pleasance, E.D., Freeman, J.D., Anderson, C.L., Varhol, R.J., Coughlin, S.M., Zuyderduyn, S.D., Jones, S.J., and Marra, M.A. 2003. A SAGE approach to discovery of genes involved in autophagic cell death. *Curr Biol* **13:** 358-363.

Co-authorship details: I was responsible for all of the computational analysis described in this chapter with the exception of raw SAGE sequence processing and implementation of statistical methods, and partially responsible for the data interpretation. Where others contributed to the work described, they are credited in the text.

3.1 Introduction

During *Drosophila* metamorphosis, the larval salivary glands undergo programmed cell death that is regulated by a transcriptional cascade induced by the steroid hormone ecdysone. Death occurs with features of both autophagy and apoptosis. Transcription of several known ecdysone-induced and cell death genes is upregulated. Subsequently, the entire larval salivary gland undergoes cell death in a rapid, stage-specific, and virtually synchronous manner. These features combine to make *Drosophila* salivary gland death an ideal system for analysis of gene expression associated with autophagic cell death. Large-scale analysis of the genes involved in this process would indicate which known PCD and hormone signaling genes are involved, as well as indicating which other genes are candidates for involvement and what other processes are important for PCD. This would contribute to the understanding of the regulation of cell death in this tissue and in other systems which undergo PCD, and provide insight into the relationship between the molecular mechanisms of apoptotic and autophagic cell death.

Gene expression analysis techniques permit simultaneous measurement of the majority of mRNA species expressed in a tissue, and as genes involved in *Drosophila* hormone-regulated cell death are transcriptionally up- and down-regulated prior to PCD, differential gene expression is an indicator of potential involvement in PCD. A combination of EST and SAGE analysis is an ideal way to examine gene expression, as no prior knowledge of the genes expressed or present in the genome is necessary. ESTs in particular can identify novel genes and transcripts, and SAGE provides quantitative measurements of gene expression. SAGE and EST libraries created by Suganthi Chittaranjan, Doug Freeman, and Sharon Gorski at the Genome Sciences Centre were the first libraries to comprehensively measure gene expression associated with autophagic cell death during normal metazoan development in vivo. SAGE libraries were created from

salivary glands dissected from developmental stages leading up to cell death, 16 hours, 20 hours, and 23 hours (at 18°C) after puparium formation (APF; see Section 1.4). An EST library was created, representing mixed stages between 16 h and 24 h. At 24 h APF, acridine orange staining is visible and salivary glands begin to be degraded.

This chapter describes the computational analysis of this EST and SAGE data, for the purpose of identifying genes and pathways potentially involved in cell death in the salivary gland. EST analysis involved removing poor quality sequences, grouping redundant ESTs, identifying the genes represented, and discovery of potentially novel genes. SAGE analysis involved removing poor quality SAGE tags, determining differential expression, and identification of the genes represented. EST sequences and all available *Drosophila* sequences including genes and genomic sequence were used for SAGE tag mapping using the method described in Chapter 2. This analysis identified known PCD genes, differentially expressed genes not previously known to be involved in PCD, and many genes of unknown function, some not previously predicted. These genes are of potential interest based on their expression patterns. Our data indicates that both apoptosis and autophagy genes are involved in salivary gland death, as are several other pathways, and identifies a number of genes of unknown function that are candidates for involvement in programmed cell death.

3.2 Methods

3.2.1 EST sequence processing and clustering

EST sequences were trimmed for vector using cross_match (P. Green, unpublished) and for poor quality sequence using Phred (Ewing and Green 1998). The minimum requirement for inclusion in analysis was at least 50 bp of non-vector sequence at Phred 20 quality, corresponding to average 99% accuracy at each base. Sequences were masked for low

complexity and repetitive regions using Dustn (NCBI toolkit release 6.1) and RepeatMasker (A. Smit, unpubl.; version available 04042000) using the *Drosophila* library from Repbase Update (Jurka 2000) vol. 7 no. 2, March 2002, option –dr. Clustering was based on cross_match alignments between all possible EST pairwise comparisons and required at least 95% identity over 80 bp with no greater than a 10 bp sequence end overhang.

3.2.2 BLAST analysis of ESTs

BLASTN (v. 2.0.14) comparison to genes revealed 3 chimeric sequences that were removed from further analyses. Representative EST sequences were chosen as (1) the longest sequence from a cluster or (2) any singleton EST sequence that aligned to genomic sequence (Release 2 of the *Drosophila* genome) by BLASTN with at least 95% identity over 80 bp. Comparisons were then conducted between representative EST sequences and all *Drosophila* annotated genes (14,350 including mitochondrial genes, Release 2 gene annotations), enhanced to include UTRs as described in Chapter 2, and all *Drosophila* expressed sequences (both ESTs and full-length cDNAs; total of 259620 as of May 14 2002) using BLASTN with a minimum requirement of 95% identity over 80 bp.

3.2.3 SAGE tag processing

SAGE tags were extracted from sequences and counted using a pipeline of Perl scripts written by Scott Zuyderduyn and Richard Varhol. Phred was used to call bases (Ewing et al. 1998) and assign quality scores (Ewing and Green 1998), and vector sequence was detected using cross_match (P. Green, unpublished). Ditags significantly shorter or longer than expected (due to errors creating or deleting CATG sites) were removed. Duplicate ditags, with identical sequences, were identified and all but one copy removed from analysis as described (Velculescu

et al. 1995). This can result in removal of true ditags, especially when certain mRNAs and therefore certain tags are highly represented in the library, but this only reduces the quantity of very highly expressed tags by a small percentage, and does not significantly affect differential expression. 14 bp tags were extracted from ditags, and tags were removed if (1) they were derived from linker sequences, (2) their overall quality, as calculated by multiplying the Phred error probability for each base, was less than 95%, i.e. >5% chance of at least one error, or (3) they were only observed once in all three libraries combined.

3.2.4 SAGE tag-to-gene mapping

Conceptual transcripts representing *Drosophila* genes were constructed and used for tag-to-gene mapping, as described in Chapter 2. The sequenced salivary gland ESTs were used in transcript construction to maximize the number of tags that could be mapped. Tags which did not match to genes were matched to *Drosophila* EST sequences (both 5' and reverse complemented 3'); the public ESTs used are described in Section 3.2.2. When such ESTs overlapped with a gene, or their clone partner overlapped with a gene, the SAGE tag was assigned to the gene as well. Multiple EST matches that resolved to the same gene or genomic location were considered unambiguous. Remaining SAGE tags were mapped to genomic sequence, and also to the reverse orientation of genes and ESTs to identify antisense tags. Gene annotations, including Gene Ontology functional categories and chromosomal locations, were obtained from the FlyBase database (FlyBase Consortium 2002).

3.2.5 SAGE statistics

Differential gene expression in SAGE libraries was determined in three pairwise comparisons. Statistical differences were determined using the formulas of Audic and Claverie

(1997), a SAGE-specific statistical method which accounts for sampling error. For each tag, the probability (p) that the observed difference in tag frequencies between two libraries would occur by chance is determined, given the respective library sizes. Tags with $p<0.05$ were considered differentially expressed; multiple testing was not taken into account in this analysis. This algorithm was implemented in Perl scripts and in the DiscoverySpace software (Scott Zuyderduyn, Richard Varhol et al, unpubl.) as well as in Java (Mehrdad Oveisi).

3.2.6 ACEDB database

The *Drosophila* ACEDB database as described in Section 2.2.1 was expanded to include and visualize the newly sequenced ESTs, SAGE tags, and other data such as protein and genome alignments. ESTs were aligned as described in Section 2.2.1. Mapped SAGE tags were positioned on the genome and associated with genes to facilitate searching for tags matching genes, tags matching ESTs, and tags matching genomic sequence only. All *Drosophila* proteins and all SWISSPROT-TREMBL proteins (obtained via SRS Sep 27 2002) were aligned against the genome with BLASTX ($E<0.001$) to facilitate identification of novel genes. In areas of particular interest, TBLASTX ($E<0.001$) alignments of *Drosophila pseudoobscura* (downloaded from the Baylor Human Genome Sequencing Centre June 12 2003) and *Anopheles gambiae* (downloaded from NCBI October 2 2002) partially sequenced genomes were also incorporated.

3.3 Results

3.3.1 Processing and clustering of EST sequences

A library of 3' ESTs was constructed from mixed-stage *Drosophila* OreR salivary glands (16 h to 24 h APF). A total of 7680 sequence reads were obtained, which were processed to select high-quality sequences and analyzed to group the ESTs based on sequence similarity

(Figure 3.1). Three ESTs were identified as chimeric and removed from analysis. Each of the 1696 representative EST sequences, 1323 of which were single-EST clusters, potentially represents a gene or transcript. The quality of this clustering was determined by comparing the ESTs to annotated *Drosophila* genes. Two genes were represented by two clusters each instead of one, and one cluster erroneously contained ESTs corresponding to two genes. In addition, 374 singleton ESTs matched genes also matched by other representative EST sequences; these singleton ESTs were not clustered due to small sequence differences, so that the correct cluster membership could not be assured. Overall, the clustering was very successful in grouping ESTs, as few clusters split or joined genes.

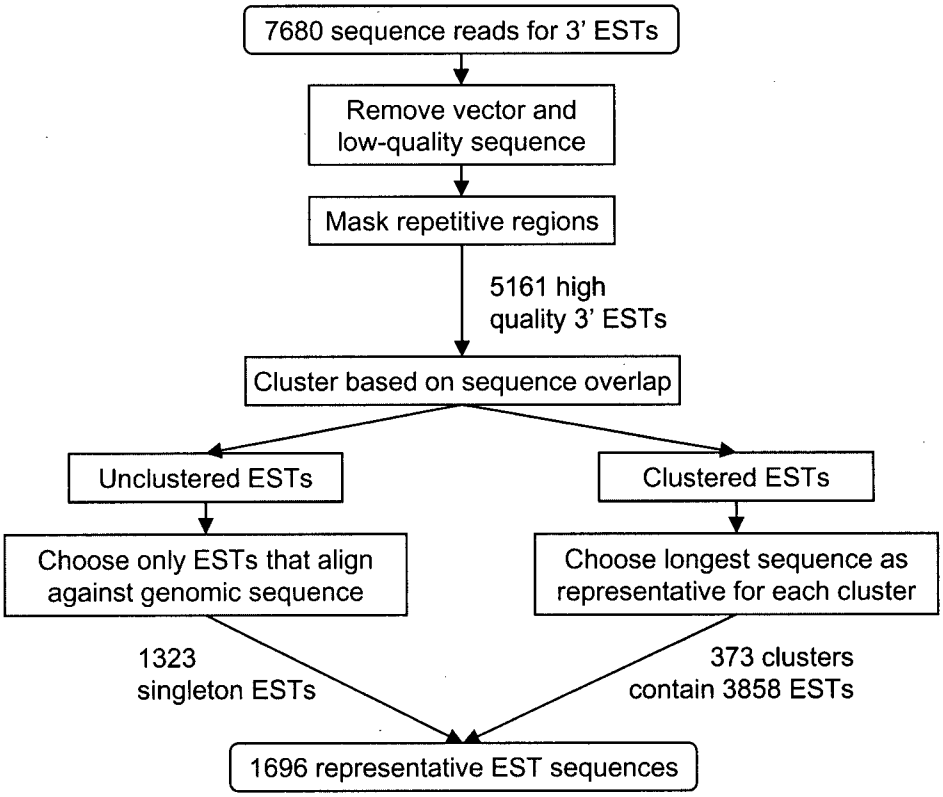

Figure 3.1 Processing and clustering of *Drosophila* 3' ESTs.

3.3.2 Genes identified by ESTs

Comparison of ESTs to predicted *Drosophila* genes and to publicly available ESTs (Table 3.1) identified subsets of ESTs that represent known genes, and potentially novel genes. The most abundant gene, a mitochondrially encoded ribosomal RNA, represented 14% of the ESTs. Other abundant genes were genes known to be ecdysone inducible, or genes of unknown function. 4 of the 10 most abundant genes were not already represented by any publicly available ESTs. Of the 1043 genes matched by the ESTs, 56 had no previous expressed sequence evidence; thus, our ESTs confirmed the expression of these predicted genes. Notably, 196 representative ESTs (16 clusters and 180 singletons, 38 of which are spliced) did not match any known or predicted genes, or any previously sequenced ESTs or cDNAs (Table 3.1). These ESTs represent potentially novel genes or transcripts. Examples of ESTs corresponding to known genes, novel transcripts, and potentially novel genes are shown in Figure 3.2. All representative ESTs, matching genes, and matching public ESTs are listed at http://sage.bcgsc.ca/tagmapping/SG_representative_ESTs.txt.

A number of known PCD genes were also identified by ESTs (Table 3.2), representing several parts of the salivary gland death pathway as well as several autophagy genes. All of these genes were identified by very few ESTs, as the EST library was not very deep, but further confirmation of the expression of PCD genes was found in the SAGE data (see below).

Table 3.1 Summary of sequences matched by representative *Drosophila* ESTs.

	Match predicted genes	Do not match predicted genes	Total
Match public ESTs	1280	145	1425
Do not match public ESTs	75	196	271
Total representative ESTs	1355	341	1696

Table 3.2 Cell death genes identified by ESTs.

Gene	Description	Number of ESTs
BR-C	Ecdysone PCD signaling	1
Rpr	Pro-death gene	1
Diap-1 / Th	Inhibitor of apoptosis	4
Dark / Apaf-1	Caspase activator	3
CG7188	Probable inhibitor of pro-apoptotic Bax	3
Dcp-1	Caspase	1
Dronc / Nc	Caspase	1
Drice	Caspase	3
Crq	Croquemort, engulfment receptor	1
Ced-6	Similar to *C. elegans* apoptosis engulfment gene	3
CG6194	Atg4-like, autophagy gene	2
CG1643	Atg5-like, autophagy gene	1

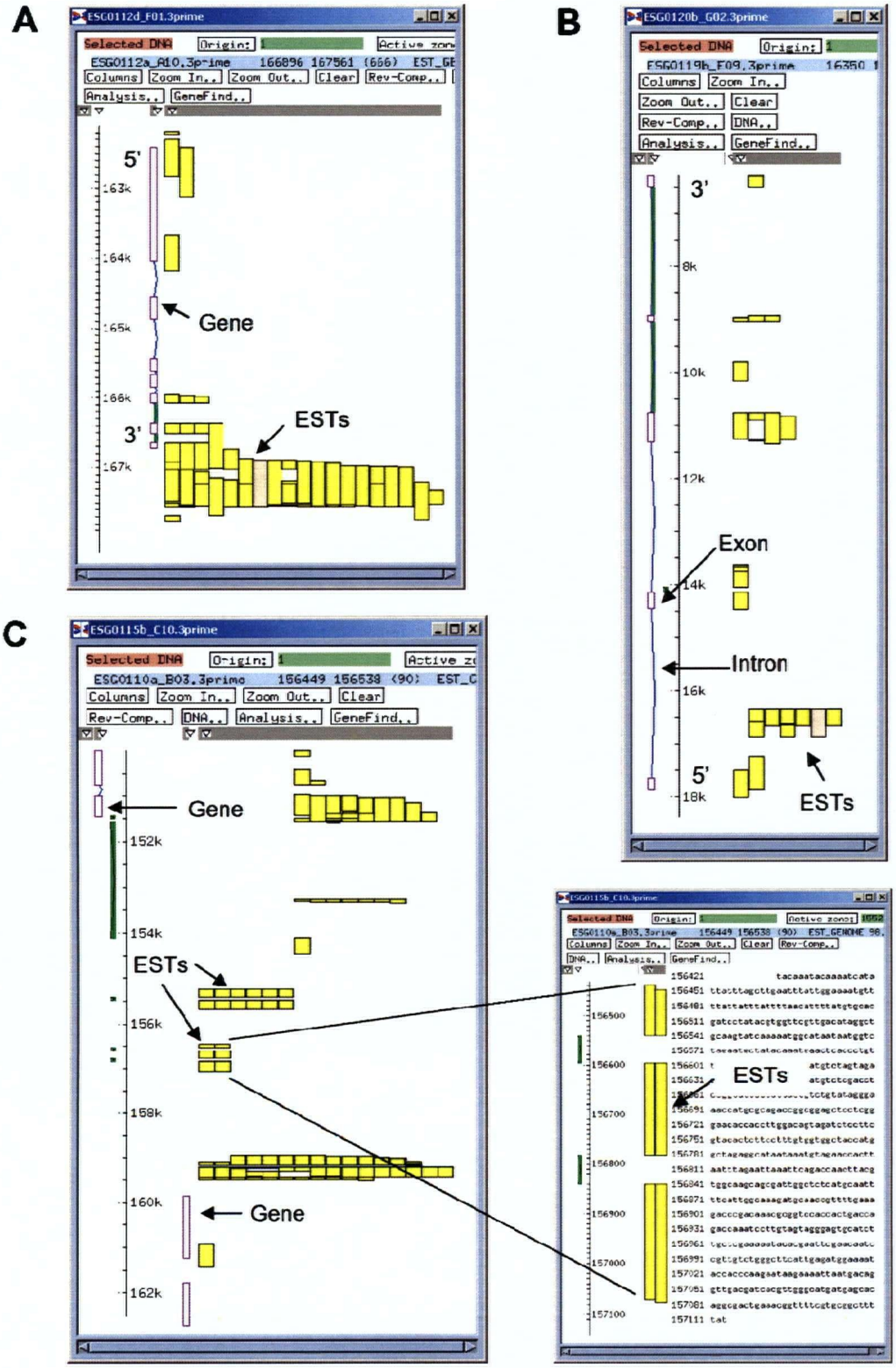

Figure 3.2 ESTs representing known and potentially novel genes.
Images generated from the *Drosophila* ACEDB database. (A) ESTs are in the 3' UTR of a gene. (B) ESTs are in the intron of a gene, possibly representing a novel exon or an entirely separate gene. (C) ESTs are not near any predicted genes and are spliced, and therefore are likely to represent novel genes.

3.3.3 Processing and differential expression of SAGE tags

Three SAGE libraries were constructed from 16 h, 20 h, and 23 h APF salivary glands (Suganthi Chittaranjan and Doug Freeman). Sequences containing SAGE tags were processed to extract tags (Scott Zuyderduyn and Richard Varhol); tags of poor sequence quality, tags potentially derived from amplification bias, and tags only seen once in all the libraries combined were removed. Each library was sequenced to a depth of over 30,000 high-quality tags (Table 3.3). In all libraries, a small number of tags were very highly abundant; at the 23 h timepoint, one tag species accounted for 9.1% of the tags in the library. The salivary gland is predominantly made up of a single cell type, and thus it is expected that this tissue will have reduced complexity compared to libraries constructed from mixed tissues or mixed timepoints. Nevertheless, a large number of tags are seen only a few times, suggesting that despite this reduced complexity there are many genes expressed at low levels, and the number of different genes expressed is most likely even higher than observed.

SAGE libraries were compared pairwise to identify genes up- or down-regulated in the three stages profiled. In the 16 h vs. 23 h comparison, 522 (12.1%) transcripts were upregulated significantly ($p<0.05$), and 331 (7.7%) transcripts were downregulated significantly ($p<0.05$) prior to cell death. Together, these transcripts account for almost 20% of all transcripts expressed in the salivary gland during these two stages. In the 16 h vs. 20 h and 20 h vs. 23 h comparisons, 288 (7.0%) and 459 (11.2%) transcripts were significantly upregulated, and 255 (6.2%) and 287 (7.0%) transcripts were significantly downregulated, respectively.

Table 3.3 SAGE library tag counts and frequency distributions.

	Tags		% of tag species seen at frequency:			
Library	**Total**	**Unique**	**1**	**2-10**	**11-100**	**100+**
16 h	34,989	3,126	32.7%	55.7%	10.4%	1.2%
20 h	31,215	3,034	38.0%	50.9%	9.7%	1.4%
23 h	30,823	2,963	33.3%	54.0%	11.2%	1.4%
Total	97,027	4,628				

3.3.4 Genes identified by SAGE tags

The genes represented by the 4628 different SAGE tags were identified by a series of comparisons to available sequence databases (Figure 3.3). Tags were mapped to conceptual transcripts enhanced to include UTRs as described in Chapter 2, as well as to ESTs and genomic sequence. This comprehensive mapping method was successful in mapping all but 7% of the tags. Of the 4628 tags, 53% were matched to genes directly; half of these tags mapped to UTRs, demonstrating the necessity of using conceptual transcripts rather than simple gene predictions. An additional 7% were matched to genes through matches to ESTs. 217 tags (5%) unambiguously match ESTs not corresponding to predicted genes; 55 of these match ESTs from the salivary gland library only. 225 antisense tags were also found, where a tag matched both genomic DNA and the reverse strand of an EST, and the tag mapped to a single genomic location and all matching ESTs also corresponded to the same location. Notably, 294 SAGE tags (6%) matched only to genomic sequence, suggesting potential locations for genes which have not been predicted and do not have previously available expression evidence.

Genes identified by SAGE tags were associated with Gene Ontology (Ashburner et al. 2000) terms describing their function, as well as chromosome band location. A complete list of salivary gland SAGE tag sequences, frequencies, and mappings can be found at

http://sage.bcgsc.ca/tagmapping/SG_SAGE_tags.txt and summary tables at http://www.bcgsc.ca/lab/fg/dsage. As in the EST libraries, some of the most abundant genes identified by SAGE were ecdysone-inducible genes and mitochondrially-encoded rRNAs, as well as genes of unknown function. Several abundant SAGE tags corresponded only to ESTs. Examination of this gene set as described below, in particular determining the pathways of interest that were found to be differentially expressed, was done in conjunction with Sharon Gorski and Suganthi Chittaranjan based on the results of my analysis.

SAGE tags corresponding to known cell death and ecdysone-induced genes associated previously with salivary gland cell death (Jiang et al. 1997; Jiang et al. 2000; Lee et al. 2000) were detected in the SAGE libraries (Figure 3.4). BR-C, E74 and E75 are general ecdysone-induced primary response genes shown previously to regulate salivary gland expression of cell death genes. E93 is a stage and tissue-specific ecdysone-induced primary response gene required in prepupal salivary glands for maximal expression of both ecdysone-induced and cell death genes. Consistent with a previous report (Lee et al. 2000), we detected increased expression from 16 to 23 h APF of the pro-death genes Dark, Dronc and Crq. The genes E93, Rpr and Diap2 were detected but expressed at low levels.

Some of the genes showing differential expression that are involved in pathways other than those previously known to be active in the salivary gland are shown in Table 3.4. Multiple apoptosis genes not previously associated specifically with salivary gland cell death were identified both in the EST and SAGE libraries, as were a number of autophagy genes. The caspases Dronc, Dcp-1, Drice, and Dredd were present, as were Diap-1 and Diap-2. The Bcl-2 gene Debcl and the pro-death gene Sickle were upregulated. Nine putative homologs of yeast autophagy genes were detected, with marked increase in expression of Atg4/apg4/aut2. Multiple

cathepsins, proteins involved in lysosomal degradation which is part of the autophagic process, were also upregulated. These data identify specific autophagy genes which may be involved in salivary gland death, and indicate that this autophagic cell death may use some of the same molecular components as does apoptotic cell death.

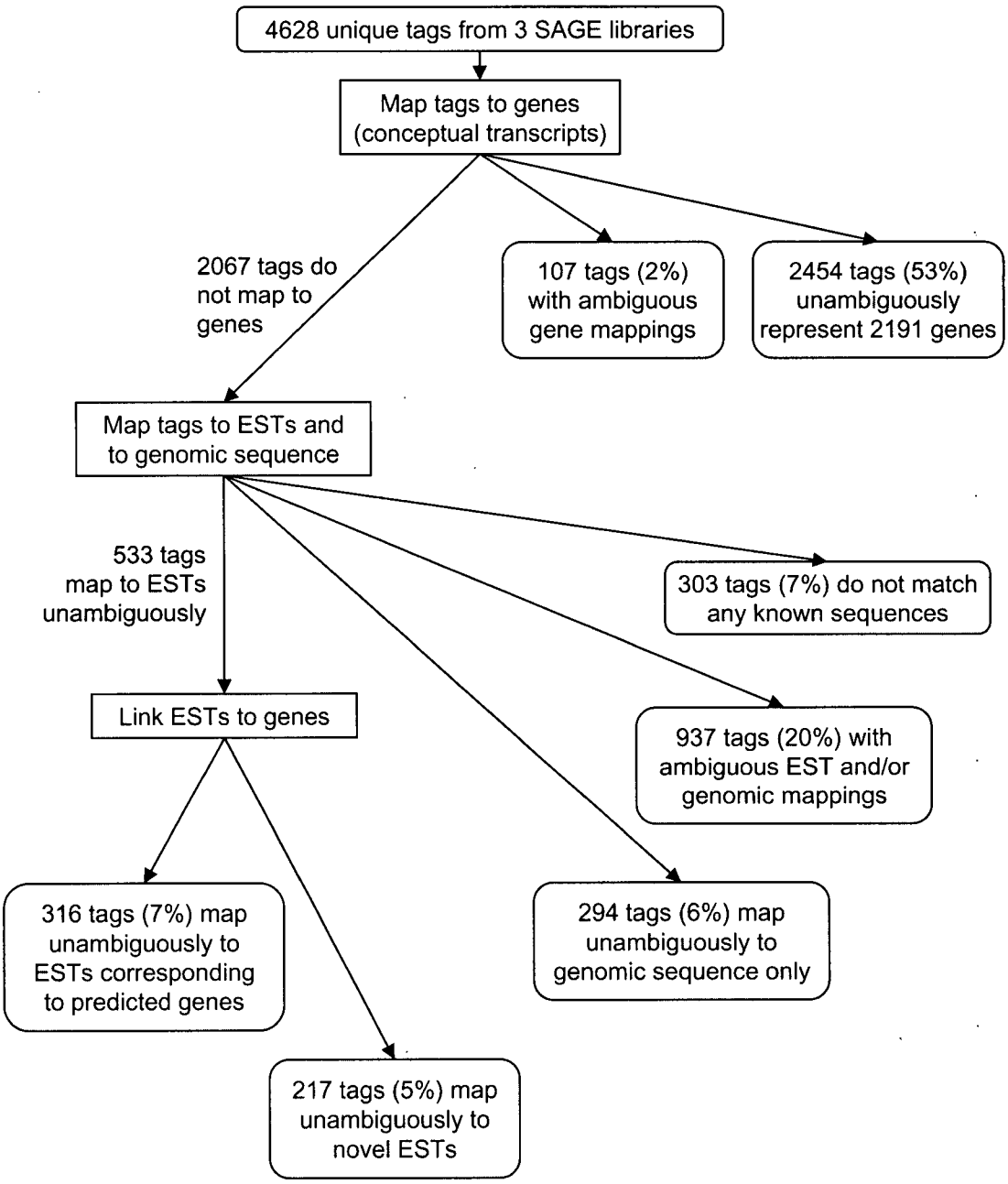

Figure 3.3 Mapping of *Drosophila* SAGE tags to genes, ESTs, and genomic sequence.
SAGE tags were mapped iteratively, first to genes, then to ESTs and genomic sequence. Mappings were to the 3'-most enzyme site. Ambiguous mappings refer to SAGE tags that exactly matched to two or more sequences.

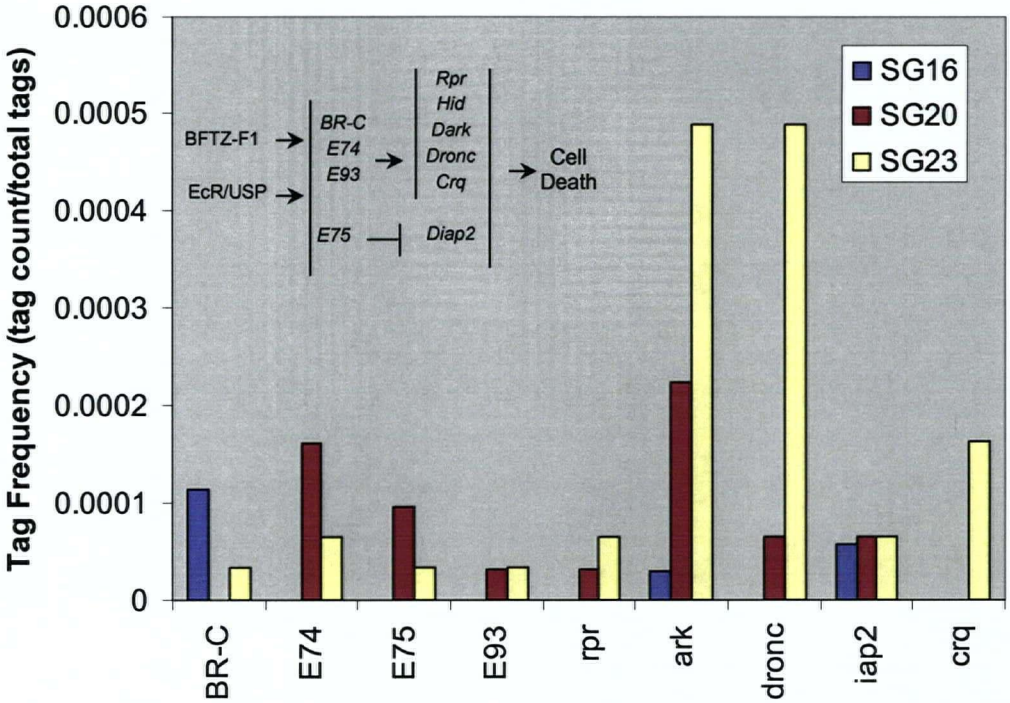

Figure 3.4 Expression of known salivary gland cell death related genes in SAGE libraries.
Observed number of SAGE tags corresponding to known salivary gland death related genes were converted to frequencies for purposes of comparison. SG16, SG20 and SG23 refer to the 16 h, 20 h, and 23 h SAGE libraries, respectively. The inset of a simplified cell death pathway indicates the relative timing of expression for the genes indicated (Jiang et al. 2000; Lee et al. 2000).

Table 3.4 Differentially expressed genes associated with salivary gland autophagic cell death in *Drosophila* SAGE libraries.

Tag Sequence	16 h	23 h	P-value	Gene	GO id	GO Molecular Function
Protein Synthesis						
ATATTGTCAA	11	24	8.34E-03	*Ef1gamma*	3746	translation elongation factor
AGCAGGGGGA	1	9	5.69E-03	*CG5605*	8079	translation termination factor
ATGAAAAACA	1	25	5.82E-08	*CG3845*	3743	translation initiation factor
TGGGAGGATG#	0	10	4.12E-04	*CG8277*	3743	translation initiation factor
TGGGAGGATG#	0	10	4.12E-04	*eIF-4E*	3743	translation initiation factor
ACCCACGAGC	4	15	4.35E-03	*CG9769*	3743	translation initiation factor
GGGTGTCTCT	0	5	1.95E-02	*CG10192*	3743	translation initiation factor
ATGAGCTATG	0	4	4.22E-02	*CG7439*	3743	translation initiation factor
TTTGAATAAC	60	81	7.70E-03	*eIF-5A*	3743	translation initiation factor
Ecdysone/hormone						
AACTGTAATG	65	0	3.28E-18	*Eig71Ed*		ecdysone-induced protein
AACGAGGGAT	1103	38	1.99E-239	*Eig71Ef*		ecdysone-induced protein
AGACGGATTC	1371	540	3.09E-58	*Eig71Ej*		ecdysone-induced protein
GGTTTATTGT	2	36	1.79E-10	*Hr78*	4879	ligand-dependent nuclear receptor
TAGCAACTAG	2	8	3.64E-02	*Hr78*	4879	ligand-dependent nuclear receptor
AGTCAAAAGG	32	532	2.32E-135	*CG15505*		
GATCCAGCCA	121	2339	0.00E+00	*CG7592*		
TGGATTCATA	2	*11	6.83E-03	*Eip63F-1*	5509	calcium binding
GCCGAATCTG	1	*7	2.50E-02	*Eip71CD*	8113	protein-met-S-oxide reductase
Transcription Factors						
TTAAGTTCGT	1	*6	4.79E-02	*Bun*	3702	RNA pol II transcription factor
TAGCTGGTGT	1	*8	1.30E-02	*EP2237*	16563	transcriptional activator
TCCAATTCCG	0	*5	2.16E-02	*CG9954*	3700	transcription factor
GAGCAGGAGT	0	*11	2.34E-04	*CG3350*	3700	transcription factor
Signal Transduction						
CGAATAATCC	3	67	3.03E-19	*Akap200*	5079	protein kinase A anchoring
AGAATCCAAC	0	5	1.95E-02	*Traf1*		
TGTACACTTC	0	33	8.10E-12	*Doa*	4674	protein serine/threonine kinase
CTGCGCTTGT	0	5	1.95E-02	*Doa*	4674	protein serine/threonine kinase
TAAATAAAGG	2	14	8.24E-04	*Sktl*	16308	1-PI-4-phosphate 5-kinase
AGAAGATAAA	0	4	4.22E-02	*Ptpmeg*	4725	protein tyrosine phosphatase
CAAGTAACCA	0	10	4.12E-04	*PR2*	4713	protein tyrosine kinase
TAGCTCTTAG	0	5	1.95E-02	*CG16708*	17050	D-erythro-sphingosine kinase
TGAACGAGGA	1	9	5.69E-03	*CG8655*	4702	receptor signaling S/T kinase
Cell Death						
TTCCGCATAT	4	13	1.32E-02	*Emp*	5044	scavenger receptor
GCTTTCGTGT	1	7	2.21E-02	*CG12789*	5044	scavenger receptor
CCCGTTCCAC	2	8	3.64E-02	*CG3829*	5044	scavenger receptor
GGCACCAGTC	4	*0	8.35E-02	*Debcl*	16506	apoptosis activator
TATTTTCTTT	1	38	4.24E-11	*Sickle*		
Autophagy						
TAGCGCTTAG	0	30	8.20E-11	*CG6194*		apg4/aut2-like; cysteine protease
TAAAATTGCT	7	12	1.44E-01	*Rab-7*	3928	RAB small monomeric GTPase
GATCCAGCCC	0	4	4.22E-02	*CG11159*	3796	lysozyme
CATCATCATC	19	566	3.86E-160	*CG3132*	4565	beta-galactosidase
GTTTCTTCCG	3	15	1.53E-03	*CG10992*	4213	cathepsin B
GGCAACGATC	8	43	2.23E-08	*cathD*	4192	cathepsin D
AAATAAATTG	66	240	1.04E-30	*CG17283*	4193	cathepsin E
TTCTTCAACC	0	4	4.22E-02	*CG12163*	16946	cathepsin F
ATGGCAGAGA	5	15	1.04E-02	*Cp1*	4217	cathepsin L
TATGATATAG	58	620	2.63E-139	*Cp1*	4217	cathepsin L

* tag number corresponds to the SG20 SAGE library
\# ambiguous mapping to two similar genes

3.4 Discussion

This study represents the first comprehensive analysis of genes associated with autophagic cell death in vivo. Our work was published alongside similar research by Lee et al (2003), who used oligonucleotide microarrays to examine gene expression in *Drosophila* salivary gland PCD and radiation-triggered cell death. Their findings were consistent with those described in this chapter; they similarly reported differential expression of apoptosis and autophagy genes, and changes in other cellular pathways. In addition to providing important clues to the molecules involved in PCD, however, our work also contributes significantly to knowledge of the *Drosophila* genome. With respect to the latter, this work provides evidence for the expression of more than 4,000 transcripts, including over 500 previously unpredicted, almost 300 previously undetected, and at least 225 overlapping and divergently transcribed. In total, 1244 different transcripts were expressed differentially prior to salivary gland cell death, and 377 of these did not correspond to predicted genes. Detection of these transcripts exemplifies the advantage of the SAGE and EST methods, ideally suited for the discovery of new genes.

3.4.1 Use of EST and SAGE to identify genes

The EST set, though small, comprised significant novelty. 341 ESTs (20%) did not match to predicted genes, and 196 of these (12%) did not correspond to any previously sequenced ESTs at the time of our analysis. Many of these ESTs not corresponding to genes likely represent previously unannotated genes, but some will represent novel splice variants or 3' ends of already predicted genes. This data therefore suggests that the number of genes in *Drosophila* may be substantially underestimated, as suggested by previous EST analysis studies (Andrews et al. 2000; Posey et al. 2001) and more recent expression analyses (Hild et al. 2003). Strikingly, despite the previous sequencing of over 200,000 *Drosophila* ESTs and cDNAs, over half of the

ESTs not matching genes were unrepresented in public databases. This demonstrates the uniqueness of this salivary gland-specific resource, and suggests that deeper sequencing from specific tissues will be necessary to identify all *Drosophila* genes and transcripts. The genes represented by such ESTs are likely to have a unique role in the salivary gland, possibly in PCD.

The SAGE tag-to-gene mapping results highlight one of the main advantages of the SAGE method compared to other large-scale profiling methods such as oligonucleotide- or cDNA array-based analyses. SAGE has the potential to reveal transcripts not previously identified, and indeed 45% of the SAGE tags did not correspond to known or predicted genes, consistent with more recent analyses of *Drosophila* SAGE libraries (Lee et al. 2005). It is possible that some of these SAGE tags do not correspond to predicted genes because they are derived from noncoding transcripts, as recent work in several species has demonstrated that such transcripts may account for a significant fraction of expressed mRNAs (Ota et al. 2004; Tupy et al. 2005) but only protein-coding transcripts were included in the *Drosophila* genome annotation at the time of my analysis. 217 tags, or 5% of the set, corresponded unambiguously to ESTs only. 55 of these mapped specifically and unambiguously to the novel salivary gland ESTs, thus confirming the expression of these putative salivary gland-specific novel genes and demonstrating the advantages of a complementary tissue-specific 3' EST and SAGE approach. 294 tags, or 6%, mapped uniquely and unambiguously to genomic DNA and may represent novel genes or novel 3' ends or splice forms of already predicted genes. Our SAGE data provides the first evidence of expression for these putative transcripts.

In at least 225 cases, SAGE tags represent apparent antisense transcription. Some fraction of these antisense tags could arise out of mispriming in cDNA library construction as discussed in Section 1.5.4.1, but such errors are likely to account for only a few percent of the observed

tags. These tags suggest the existence of previously unpredicted transcripts that may represent divergently transcribed overlapping gene sequences; this is not surprising as current gene finding programs are unable to readily detect overlapping genes (Rogic et al. 2001). More recent work analyzing SAGE and other expressed sequence data has indeed identified many such overlapping genes and antisense transcripts in mammals (Wahl et al. 2005; Yelin et al. 2003), and pairing of sense-antisense overlapping transcripts is conserved in evolution (Dahary et al. 2005). The significance of these transcripts is not known, but it is possible that they act as a form of antisense regulation at the transcriptional or post-transcriptional level.

Even after mapping to all *Drosophila* sequences available at the time of this analysis, 303 tags, 7% of the SAGE set, were left unassigned. The unmapped tags could be due to issues discussed in Section 2.4.3 such as sequence polymorphisms or common sequencing errors, or to lack of representation in the available sequence resources. The latter could occur if tags represented genes in heterochromatic regions, which at the time of my analysis were not sequenced but have since been shown to contain significant numbers of genes (Hoskins et al. 2002). It is also possible that unmapped tags span adjacent exons that are currently not represented in the EST or cDNA data set, and as such would not be identified in genomic sequence. Overall, the salivary gland EST and SAGE data confirms expression of predicted genes and can also aid in gene discovery, demonstrating the importance of expression data for genome annotation, especially data from specific tissues.

3.4.2 Expression of PCD genes

In general, the gene expression profiles of known salivary gland cell death genes as generated by SAGE are consistent with previous reports and can temporally distinguish known upstream transcriptional regulators from downstream death effector molecules. The genes E93,

Rpr and Diap2, which were detected only at very low levels, were analyzed further by quantitative RT-PCR by Suganthi Chittaranjan and Shaun Coughlin and this indicated expression profiles consistent with previous studies (Jiang et al. 1997; Lee et al. 2000). RT-PCR analysis of 96 genes chosen to have varying expression levels were concordant with respect to the direction of change in expression for 91/96 (95%) of the genes tested and the overall correlation in fold-difference of 0.48 was significantly positive. Thus, the SAGE data is confirmed by and consistent with lower-throughput, quantitative RT-PCR results. Genes known to be expressed during salivary gland PCD but not detected in our SAGE and EST libraries were EcR, USP, βFTZ-F1 and hid (Figure 3.4 inset). All of these genes possess putative *Nla*III recognition sites and thus theoretically can be associated with a SAGE tag. However, EcR, USP and βFTZ-F1 act upstream of the primary response genes BR-C, E74, E75, and E93, and thus may be expressed maximally prior to 16 h APF. This interpretation is consistent with Northern analysis of βFTZ-F1 (Jiang et al. 2000). Alternatively, these genes may be expressed at very low levels. Failure to detect Hid was not surprising because RT-PCR analysis indicates it is expressed at levels lower than Rpr which was detected only two times at the 23 h timepoint.

The degradation phase of autophagic cell death appears to utilize components of the machinery required for autophagy. While autophagic cell death was shown previously to share morphological features with autophagy, there had been no prior connection between the molecules involved in these two processes at the time of our study. Expression was detected in the EST library of a gene similar to Atg5, involved in one of the ubiquitin-like pathways required for autophagy in yeast (Ohsumi 2001). Particularly highly induced in the SAGE libraries was a gene similar to Atg4 which encodes a novel cysteine protease, whose yeast homolog processes and activates Atg8, an ubiquitin-like protein. Multiple lysosomal enzymes, including cathepsins,

were also upregulated prior to autophagic cell death. It is not expected that these genes would be specifically induced in apoptotic cells because the bulk of cellular degradation occurs within a macrophage or neighboring cell. Our analyses showed that multiple genes involved in apoptotic cell death are also expressed during autophagic cell death, supporting the view that these two processes occur simultaneously or utilize common pathway components (Baehrecke 2002; Lee and Baehrecke 2001; Lee et al. 2000). It is reasonable to expect, then, that some of the novel autophagic cell death associated genes identified in this study may also be associated with apoptotic cell death. Given the relationship of both autophagic and apoptotic cell death to disease, and the concomitant use of apoptotic genes as therapeutic targets (Reed 2002), it is essential to develop a detailed understanding of the molecules required for both processes. The genes discovered here provide a powerful starting point for protein function-based studies to determine the mechanisms essential for the execution of autophagic cell death and to understand how its unique components are integrated with those of known apoptotic cell death pathways.

3.4.3 Novel putative PCD genes

Many genes involved in other pathways, not specifically related to cell death, also demonstrated changes in gene expression levels between the three SAGE libraries. Multiple other ecdysone- and hormone-inducible genes were differentially expressed. Upregulation of several translation initiation, elongation, and termination factors is consistent with the concept that autophagic cell death requires active protein synthesis. Multiple transcription factors, including several with no known function in *Drosophila*, were differentially expressed between timepoints. Components of multiple signal transduction pathways, including genes related to cytoskeletal remodelling, Ras, and defense response signaling, changed in expression level, indicating what is likely to be a complex interplay of pathways in autophagic death.

In addition to assigning a possible new role to genes already annotated functionally, we have implicated in the autophagic cell death process more than 732 differentially expressed genes with unknown function. As these genes change significantly in expression during stages leading up to cell death, their expression implicates them as having a role in salivary gland death, either directly or indirectly. 377 of these differentially expressed genes were unpredicted and 48 of these are represented solely by our salivary gland ESTs.

3.4.4 Conclusions

This work has demonstrated the value of large-scale expression data for examining gene expression patterns in *Drosophila*, and the utility of analysing SAGE and EST data for identification of novel genes. The genes observed and differentially expressed indicate that programmed cell death in the *Drosophila* salivary gland utilizes components of both the apoptosis and autophagy systems, and suggests a complex regulation of this process by multiple signalling pathways. A major challenge is to identify which genes, both previously described and newly discovered, are likely to play an important role in the autophagic cell death process. One computational approach to this analysis is to examine the human homologs of these genes, not only in terms of function but also in terms of expression, to identify which genes have expression patterns consistent with a role in programmed cell death across multiple species.

Chapter 4: Identification of PCD genes in *Drosophila* and cancer gene expression data

4.1 Introduction

Programmed cell death (PCD) is essential for cell homeostasis, and cells which lose the ability to initiate or execute this process have the potential to grow in an uncontrolled manner. Genes involved in the apoptosis pathway are commonly altered in sequence or expression in cancers, as inhibition of this pathway is one of the necessary steps in oncogenesis. However, the pathways regulating cell death are complex, and in many cases the genes that are most relevant to oncogenesis are unknown. There is great potential for the use of gene expression data to identify the relevant genes and pathways which are altered in cancers. In mammals, post-transcriptional regulation at the translation or protein activity level is widespread and thus changes in mRNA levels alone can be insufficient to identify active pathways. However, the widespread gene deregulation that occurs in cancers can cause genes normally regulated at the protein level, such as caspases, to change in expression at the mRNA level and thus be detectable by analysis of differential gene expression (Takita et al. 2000).

In this chapter, I aim to identify genes involved in regulation or execution of PCD, specifically those which are altered during oncogenesis either by activation or suppression. The larval salivary glands of *Drosophila melanogaster* undergo programmed cell death during development in a concerted, precisely timed, transcriptionally regulated manner, making this tissue an ideal system in which to identify genes involved in PCD by gene expression analysis. As PCD is altered in cancer and conserved from *Drosophila* to humans, genes which are differentially expressed during cancer development and have orthologs differentially expressed during *Drosophila* PCD are candidate PCD-related genes with involvement in oncogenesis. Using SAGE data as described in Chapter 3, genes differentially expressed during *Drosophila* PCD were identified. Human orthologs for these *Drosophila* genes were found, and the

Drosophila SAGE tags linked to their human SAGE tag counterparts. The expression of human SAGE tags was compared in cancerous and non-cancerous libraries representing a variety of tissues. Choosing appropriate statistical cutoffs, genes differentially expressed during cancer progression with orthologs differentially expressed during *Drosophila* PCD were identified as candidate PCD-related genes. Included in this set were genes with a previously defined role in PCD, or an association with PCD or cancer. The genes were classified by functional categories, and the functions overrepresented in the candidate gene set determined. The observed functions point to a role for several cellular processes related to autophagy in both cancer progression and PCD.

4.2 Methods

4.2.1 Drosophila PCD expression data

SAGE libraries were obtained and processed as described in Chapter 3, such that only tags of minimum quality 99% and minimum total count of 2 were considered. The ACEDB database and associated *Drosophila* gene annotations and genomic sequence were updated to Release 3.1 (Misra et al. 2002). SAGE tags were mapped to genes from this new release using methods described in Chapters 2 and 3; tags mapped to ESTs and genomic sequence were not considered in this analysis, as no gene predictions are available and so human orthologs of these putative novel genes could not be identified.

Three library vs. library statistical comparisons were performed: 16 h vs. 20 h, 16 h vs. 23 h, and 20 h vs. 23 h, to identify all genes which changed in timepoints leading up to cell death. P-values were determined for each SAGE tag for each comparison using the Audic-Claverie algorithm used in Chapter 3 (Audic and Claverie 1997). A mysql database was created to hold tags, counts, p-values, genes, and associated gene data such as gene functions for easy

querying across datasets. Data from a similar experiment on the salivary gland using Affymetrix oligonucleotide arrays (Lee et al. 2003), giving gene expression levels at 6 h and 12 h APF (at 25°C, equivalent to 12 h and 24 h APF at 18°C) and associated p-values for differential expression, were also obtained and included in the database.

4.2.2 Human cancer expression data

265 human normal and cancer SAGE libraries were downloaded from the Cancer Genome Anatomy Project website (http://cgap.nci.nih.gov/) on October 15 2004. From these, the libraries included in analysis were (1) not derived from cell lines or cultured cells, (2) short SAGE (14 bp tags), and (3) derived from tissues for which both normal and cancer data were available.

SAGE tags were mapped using SAGE Genie (Boon et al. 2002) "best gene for tag" mappings linking tags to UniGene entries, downloaded at the same time as the SAGE libraries. SAGE Genie was chosen as the mapping method as this database utilizes full-length cDNA sequences from the RefSeq and MGC databases as well as mRNAs from GenBank and ESTs to compile a set of mappings that takes into account the reliability of each data source. For each tissue, a comparison of the expression of each tag in normal libraries vs. cancer libraries was done and p-values computed with the Audic-Claverie algorithm, implemented as a command-line program by Mehrdad Oveisi. RefSeq, LocusLink, and UniGene database identifiers were also downloaded from NCBI (June 22 2004) so that SAGE tags could be linked to genes in LocusLink and RefSeq databases. SAGE tags, counts, p-values, mapping and library information, and all database mapping links were included in the same mysql database used for *Drosophila* data.

4.2.3 Drosophila-human orthology

The InParanoid program (Remm et al. 2001) was used to determine *Drosophila*-human orthologs; this program processes BLAST results to identify orthologs and paralogs. The publicly available InParanoid ortholog pairs for these species were only computed on SwissProt sequences, which do not represent the complete set of predicted proteins in *Drosophila* and cannot be completely mapped to NCBI databases used for SAGE tag mapping. Thus, the InParanoid software was run locally on compiled sequence sets. 18,489 *Drosophila* predicted proteins (Release 3.1) were extracted from ACEDB, 18691 human RefSeq proteins were downloaded from UCSC (ftp://genome.cse.ucsc.edu; April 2003 genome build), and 9478 predicted yeast proteins were downloaded from the Saccharomyces Genome Database (June 11 2003). RefSeq proteins were used because they are high quality and full-length. Reciprocal pairwise BLASTP comparisons were performed between each species pair, a score threshold of 50 was applied, the results were input into InParanoid, which identified *Drosophila*-human orthologs with yeast as an outgroup. The output was parsed using a script written by Keith Boroevich, and integrated into the mysql database to link *Drosophila* genes to RefSeq sequences.

4.2.4 False discovery rate

To determine a p-value cutoff for differential expression for each of the *Drosophila* and human SAGE library comparisons, a desired false discovery rate (FDR) was chosen and then a p-value approximating this FDR determined. A range of p-values cutoffs for *Drosophila* comparisons (0.001, 0.01, 0.05 and 0.1) and human comparisons (0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05 and 0.1) were applied and the number of genes differentially expressed in each comparison determined. The FDR was then calculated for each p-value as described in Storey (2002):

$$\text{Estimate of FDR} = \frac{(\text{\# observations not significant}) \times (\text{p-value cutoff})}{(\text{\# observations significant}) \times (1 - \text{p-value cutoff})}$$

The actual FDR is different for each of the three *Drosophila* comparisons and the nine human comparisons performed, and so an average FDR for *Drosophila* and an average FDR for human were calculated. An average FDR close to 30% was achieved for *Drosophila* p-value cutoff of 0.05 and human p-value cutoff of 0.01. For an average FDR close to 10%, as used in the analysis, the *Drosophila* p-value cutoff was 0.01 and the human p-value cutoff was 0.001.

4.2.5 Expected proportion of differentially expressed and PCD genes

To determine whether the overlap between the PCD and cancer sets was greater than expected given a random selection of genes, the expected number of human genes in the pcd-cancer set was determined to be:

Expected # human genes = (# human genes differentially expressed) x
(% of all human genes that have *Drosophila* orthologs) x
(% of all *Drosophila* orthologs that are differentially expressed)

The expected number of *Drosophila* genes in the pcd-cancer set was determined in an equivalent manner. Similarly, the number of human known PCD genes expected to be identified by chance in the pcd-cancer set was:

Expected # human PCD genes = (# human genes appearing in pcd-cancer set) x
(% of human genes with orthologs that are known PCD genes)

4.2.6 GO analysis

Gene Ontology (Ashburner et al. 2000) terms were obtained for *Drosophila* genes from FlyBase (FlyBase Consortium 2002). Human genes were linked to GO terms using DiscoverySpace (Varhol, Zuyderduyn et al, unpubl.). For the pie chart of GO terms (Figure 4.3),

genes associated with terms below level 4 of the GO hierarchy were mapped to parent terms at level 4.

The GoMiner software (Zeeberg et al. 2003) for identification of overrepresented functions in a gene set requires UniProt IDs, and thus Ensembl was used to associate the LocusLink IDs of the human pcd-cancer genes with UniProt IDs (as of Oct 26 2004). This set of genes was input into GoMiner, and associated GO categories from any of the hierarchies (Molecular Function, Biological Process, or Cellular Component) determined. The PCD-upregulated subset of the pcd-cancer set consisted of genes which (1) increased in expression between 16 h and 23 h, or (2) did not change in expression between 16 h and 23 h, and increased in expression between 16 h and 20 h. The remaining genes, which decrease in expression in the 16 h vs. 23 h or 16 h vs. 20 h comparisons, made up the PCD-downregulated subset. In GoMiner, each GO category is given a p-value representing the probability that the observed number of genes would be associated with that category by chance, based on a Fisher exact test. The GO categories with the lowest p-values are the most highly overrepresented.

4.3 Results

4.3.1 Drosophila PCD and human cancer expression

In total 97,027 high-quality SAGE tags were sequenced for the three *Drosophila* SAGE libraries constructed from salivary glands at developmental stages leading up to PCD, at 16 h, 20 h, and 23 h after puparium formation at 18°C (Chapter 3, Gorski et al. 2003). 2960 of the 4628 different tags in these libraries could be unambiguously mapped to predicted genes, giving a set of 2313 genes expressed prior to cell death in *Drosophila*.

SAGE libraries constructed from human cancerous and normal tissues were obtained from the Cancer Genome Anatomy Project (Boon et al. 2002). 9.1 million SAGE tags from nine

tissues were included in analysis, with the largest proportion of tags derived from brain and breast samples (Table 4.1). Of the 541325 different tag species, 233614 were unambiguously mapped to 40213 human UniGene identifiers using CGAP's SAGE Genie tool (Boon et al. 2002). These corresponded to 16655 LocusLink records, which approximates the number of genes.

Orthologous genes between the *Drosophila* and human genomes, those thought to be derived from a common ancestor and thus expected to have related function, were identified using the InParanoid algorithm (Remm et al. 2001) using yeast (*Saccharomyces cerevisiae*) as an outgroup. This identified 5792 *Drosophila* genes which were putative orthologs of 8723 human genes. Cases where more than one human gene was associated with a *Drosophila* gene could arise from gene duplication since the divergence of the human and *Drosophila* lineages.

Table 4.1 Tissues and SAGE libraries from CGAP used for analysis of cancer expression.

Tissue		Brain	Breast	Colon	Kidney	Lung	Pancreas	Peritoneum	Prostate	Stomach	Total
N*	Libraries	9	8	2	1	1	1	1	2	2	27
	Tags**	769K	446K	98K	41K	89K	22K	54K	123K	51K	1,693K
C*	Libraries	69	22	2	1	3	2	1	3	4	107
	Tags	5,364K	1,255K	97K	100K	159K	67K	33K	154K	249K	7,478
T*	Libraries	78	30	4	2	4	3	2	5	6	134
	Tags	6,133K	1,701K	195K	141K	248K	89K	87K	277K	300K	9,171K

* N=Normal, C=Cancer, T=Total
** "K" refers to thousands of tags

4.3.2 Differentially expressed genes

To identify genes which change in expression in both PCD in *Drosophila* and in cancer, the genes differentially expressed in each system need to be identified, orthologs linking them established, and the overlap between the two systems determined. Each *Drosophila* and human SAGE tag was assigned a p-value for each library comparison; three pairwise comparisons

between the three *Drosophila* libraries were performed, and nine normal vs. cancer comparisons were performed, one for each human tissue. Comparisons were not done across tissues, as substantial, non-cancer related differences in gene expression would be expected. Based on the distribution of p-values obtained, a p-value cutoff of 0.01 for *Drosophila* tags and 0.001 for human tags was chosen, to correspond with a false discovery rate of approximately 10% (see Methods). This indicates that for each species, the chance that a gene will be identified as differentially expressed when in fact there is no real difference and the apparent difference is due to chance is estimated to be 10% or 0.1. These p-value cutoffs resulted in 426 *Drosophila* genes and 4509 human genes identified as differentially expressed.

Using the orthologs as identified above, and linking between the UniGene and RefSeq databases, the genes that are differentially expressed in both *Drosophila* PCD and human cancer were identified. In total, 171 human genes, orthologous to 143 *Drosophila* genes, changed in expression in both systems, hereafter referred to as the "pcd-cancer set" (full listing of genes and expression at http://sage.bcgsc.ca/tagmapping/pcd_cancer_table.txt). This set contains 2-fold more genes than expected by chance, given the number of genes differentially expressed in each system and the number of orthologs (see Methods), indicating the two processes of PCD and cancer have significant overlap. The most common pattern in this set is opposite expression in PCD and cancer; 48 human genes upregulated in cancer have orthologs downregulated in PCD in *Drosophila*, and 18 human genes are downregulated in cancer with counterparts upregulated in PCD (Figure 4.1). This suggests that overall PCD is likely to be inhibited in cancer. The 44 genes which show the same direction of change in PCD and in cancer may have more complex roles, or may represent pathways within tumors that are actively attempting to prevent tumor

growth. 61 genes are both up- and down-regulated in cancer in different tissues, indicating the variety of responses to cancer and the variety of pathways that may be active in different tissues.

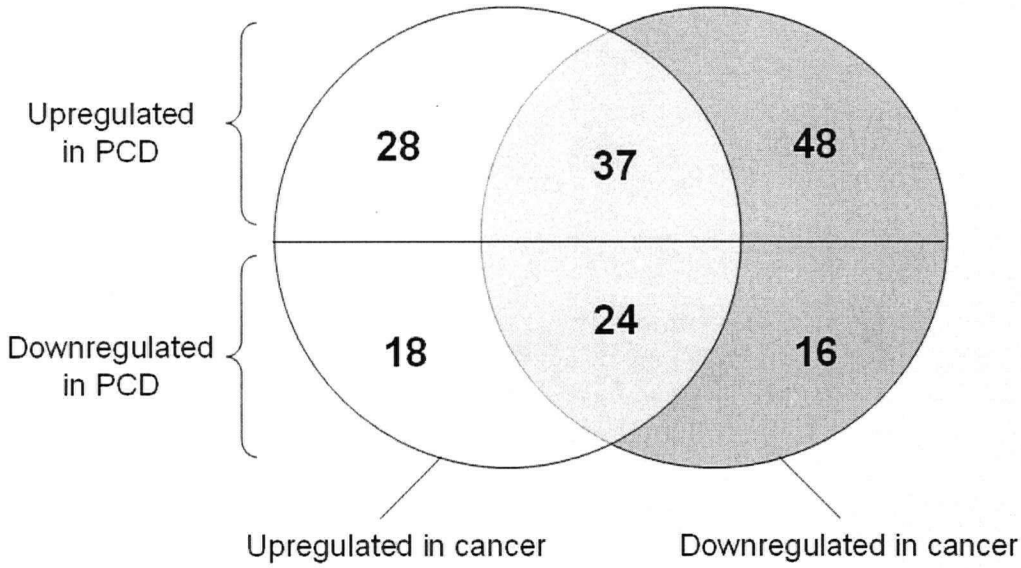

Figure 4.1 Expression categories of human pcd-cancer genes.
The 171 human genes in the pcd-cancer set were divided into categories based on their expression in cancer and the expression of their *Drosophila* orthologs in PCD. Genes shown in both circles were upregulated in cancer in one or more of the nine tissues examined and also downregulated in one or more tissues.

4.3.3 Roles of genes in PCD and cancer

Gene annotations and Gene Ontology terms were searched to identify 434 human and 50 *Drosophila* known programmed cell death genes in the respective genomes. Three of the human genes in the pcd-cancer set were in this list of known human PCD genes: PRKAG1, a subunit of the AMPK kinase involved in the regulation of autophagy (Samari and Seglen 1998), SGPL1, involved in ceramide-induced apoptosis (Reiss et al. 2004), and IkBα, a pro-apoptotic molecule (Castro-Alcaraz et al. 2002). Two of the *Drosophila* genes in the pcd-cancer set were in the list of known *Drosophila* PCD genes: Sec61a, involved in neuronal cell death (Kanuka et al. 2003), and BR-C, an early gene in ecdysone death signaling (Jiang et al. 2000) that was identified in my

analysis as orthologous to a human transcription factor that controls cell proliferation and death (Kamio et al. 2003). In both cases, the number of known PCD genes identified is not significantly greater than expected by chance given the size of the pcd-cancer set. This analysis was repeated based on Affymetrix microarray data from the *Drosophila* salivary gland (Lee et al. 2003) in place of the SAGE data, similarly identifying differentially expressed genes and human orthologs, and again no more known PCD genes were identified than expected by chance (data not shown).

Many genes, although not core components of PCD pathways, have been associated with cell death in literature. For instance, a gene may induce cell death when overexpressed or knocked out, but the mechanism may be unknown. Brief annotation and literature database searches were performed for each of the 171 human genes in the pcd-cancer set to identify which genes had been associated with cell death in this way. Similarly, implied roles in cancer, for example inhibition or promotion of tumor cell growth, involvement in translocations, or mutations in tumor cells were investigated. The majority of genes in the pcd-cancer set have some known function, and a significant portion of these were associated with cell death or cancer in the literature (Figure 4.2).

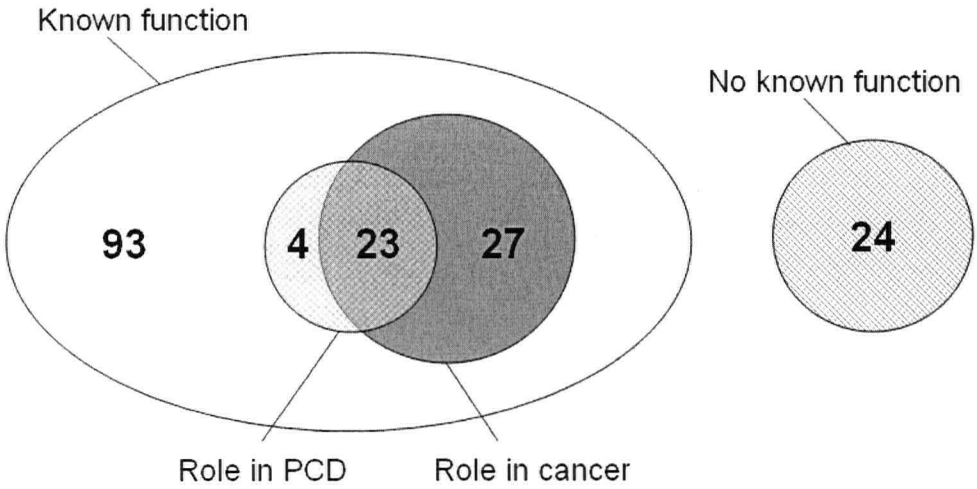

Figure 4.2 Categorized roles of human genes in pcd-cancer set as found in the literature. Searches of PubMed and other NCBI databases for each gene identified genes associated with cell death or cancer in previous experiments. Genes marked as no known function have not been studied at a molecular level.

4.3.4 Functional categorization and overrepresentation

To understand the cellular pathways and functions represented by the human genes in the pcd-cancer set, the Gene Ontology (GO) Biological Process terms associated with each gene were determined. 101 of the 171 genes had such annotated functions, representing a wide range of cellular pathways (Figure 4.3). Processes such as protein metabolism and biosynthesis appear highly represented, but this is partly due to the unequal nature of the GO hierarchical tree classification, where some processes are more highly represented due to a greater level of study or a difference in the branching pattern of the tree (e.g. some parts of the tree branch greatly near the root, while others branch at levels further down the hierarchy).

To gain an unbiased view of the functions in the pcd-cancer set, the functions most overrepresented in the set were identified using the GoMiner software (Zeeberg et al. 2003). This analysis takes into account the unequal distribution of genes and GO terms in the GO hierarchy,

and determines which GO terms are found associated with genes in a gene set more often than would be expected by chance. Terms and functions thus enriched in the pcd-cancer set are more likely to represent pathways altered in the systems under study. The pcd-cancer set, limited to 168 genes with UniProt IDs for use in GoMiner, was divided into two categories: a PCD-upregulated set of 110 human genes whose *Drosophila* homologs are upregulated in PCD, and a PCD-downregulated set of 58 human genes whose *Drosophila* homologs are downregulated in PCD. 121 genes were found to have at least one GO term in any GO category (Biological Process and others). A summary of the categories of genes found to be enriched in each set are shown in Table 4.2. Cathepsins, lysosomal proteins, and cytoskeletal proteins were primarily downregulated in cancers. Ribosomal proteins were found to be enriched both in the pcd-upregulated and the pcd-downregulated sets, as well as in several cancers. Metalloprotease inhibitors (TIMPs) changed in expression in cancers of multiple tissues: they are upregulated in brain and pancreatic cancer, and downregulated in lung and prostate cancer. Carbonic anhydrases and V-ATPases also showed varied expression in cancers of different tissues.

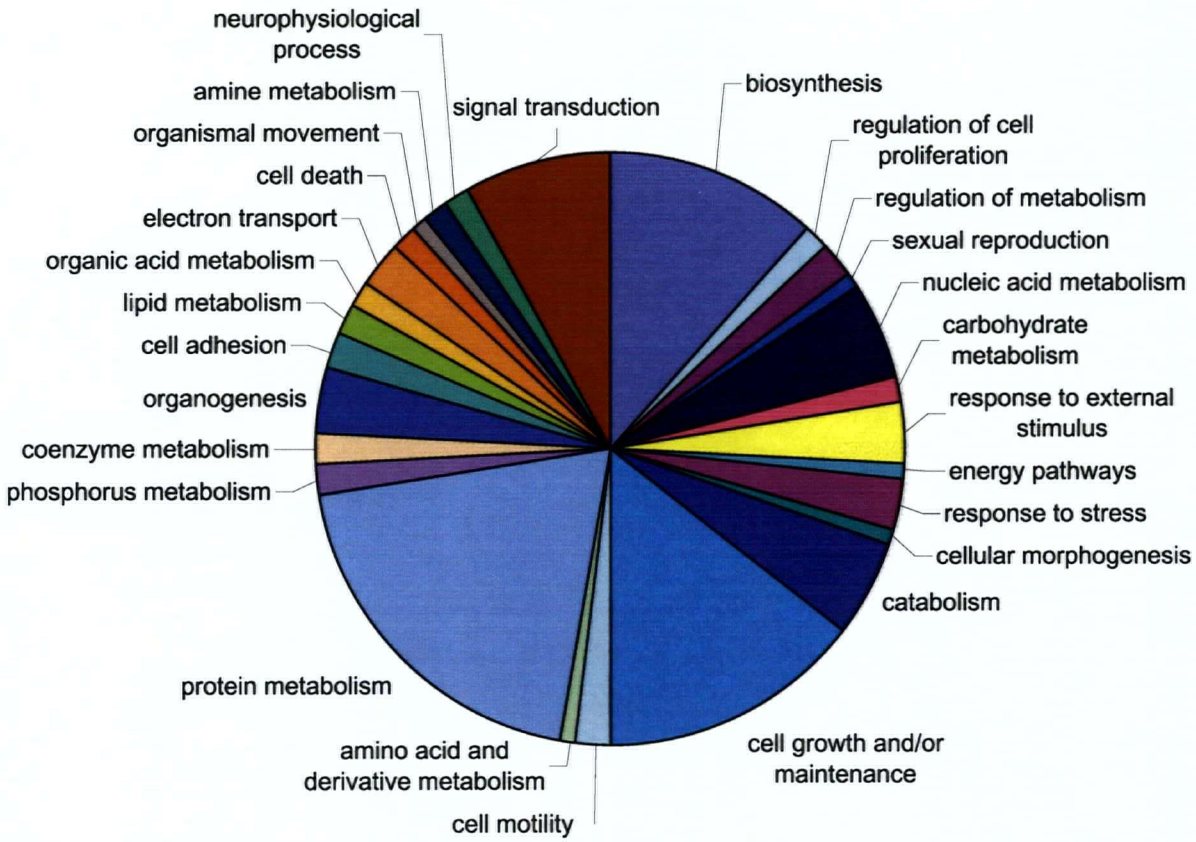

Figure 4.3 Functions associated with human genes in the pcd-cancer set.
The proportion of 101 genes associated with each GO Biological Process term were determined, and terms at level 4 of the GO hierarchy are shown. 70 genes did not have an associated GO Biological Process term.

Table 4.2 Functions overrepresented in the pcd-cancer set.

PCD-upregulated subset	PCD-downregulated subset
Cathepsins and lysosomal proteins	V-ATPases
Carbonic anhydrases	Metalloendopeptidase inhibitors
Cytoskeletal proteins	Protein carriers
Signaling and adaptor proteins	Ribosomal proteins
Ribosomal proteins	

4.4 Discussion

In this study I took a novel approach to understanding the cellular pathways important in cancer by combining expression analysis with cross-species similarity information to pinpoint genes altered in both programmed cell death and cancer. Human orthologs of *Drosophila* genes differentially expressed during salivary gland PCD were identified and the expression of these genes determined in cancers of nine tissues. 171 human genes and 143 *Drosophila* genes were found to be differentially expressed in both systems, and these are candidates for involvement in cancer through a function in PCD. Although few of these genes were found to be core components of apoptosis or autophagy pathways, a significant proportion are associated with cancer or cell death in the literature. The genes represent a wide range of functions, and the functions overrepresented in the set point to pathways regulated in both programmed cell death and cancer.

4.4.1 Cross-species integration

Model organisms are an essential aspect of molecular biology, allowing experiments that are inappropriate or impossible to perform in humans or mammals to be carried out in other, often simpler, organisms. Advent of genomic technologies has only increased the utility of such models, as comparative genomics allows identification of genes, regulatory sequences, and other genomic features (Ureta-Vidal et al. 2003). My approach expands the concept of model organism analysis to include analysis of expression data in multiple organisms concurrently. The *Drosophila* salivary gland has advantages as a model of programmed cell death that are not available in mammalian systems. The method used here directly applies gene expression studies in *Drosophila* to the much more complex system of cancer in humans, of which programmed cell death is one aspect, and thus permits the identification of genes involved in both processes. Some

previous work has been done using this concept, for instance, for identification of genes involved in breast cancer by concurrent analysis of mouse and human SAGE data (Hu et al. 2004), but overall cross-species analyses of genome-wide data have rarely been applied to gene expression.

Cross-species analysis presents several difficulties which must be considered. It is estimated that over half a billion years has elapsed since the divergence of human and *Drosophila* lineages (Hedges et al. 2004), during which time varying evolutionary requirements can result in multiple gene deletions, duplications, and mutations and thus make identification of true orthologs, rather than paralogs, difficult. The InParanoid algorithm is specifically designed to account for such complexities, as it takes into account all relevant gene-gene distances both between and within genomes (Remm et al. 2001). Indeed, several cases examined in detail showed that given the sequence alignments, the ortholog relationship found appeared most likely to be correct. However, as the full complement of human genes, and to a lesser extent *Drosophila* genes, is not yet determined, in some cases the correct ortholog relationship may not be discovered. In addition, even when genes are orthologous, there is no certainty that the genes perform the same function in both organisms; the pathway may serve a different purpose, or the gene's activity may have changed. Thus, absolute inferences cannot be made about the role of human genes in cases where their orthologs are involved in programmed cell death. However, given the similarities in the known pathways of apoptosis and autophagy in humans and *Drosophila*, in many cases the results are expected to be relevant to both systems.

4.4.2 Functions of genes in PCD and cancer

2-fold more genes were found in the pcd-cancer set than expected given the number of genes identified as differentially expressed in the PCD and cancer systems individually; this overlap indicates commonality between these processes. Similarly, the observation that a higher

proportion of genes have opposite expression tendencies in cell death and in cancer points to an inhibitory role for cell death processes in cancer, consistent with current understanding as discussed in Section 1.3. If the observed decrease in expression of genes in cancer is due to a mutation associated with cancer progression, those genes are potential tumor suppressors; similarly, genes that increase in expression have the potential to be oncogenes.

However, it is notable that the pcd-cancer gene set is not enriched for genes involved in programmed cell death, as might be expected given that their expression in *Drosophila* indicates a potential role in this process. There are several issues that may contribute to this. It is not expected that all programmed cell death genes would be regulated at the transcriptional level. In *Drosophila*, many are, and in cancer, many genes that are normally regulated post-transcriptionally are altered in expression due to drastic changes in gene regulation caused by mutations in cancer. However, genes not regulated at the mRNA level in both systems would be excluded from this analysis. Also, many of the genes known to be involved in programmed cell death are expressed at very low levels in the *Drosophila* SAGE libraries (see Chapter 3), and therefore do not have the opportunity to be recognized as differentially expressed. Analysis of Affymetrix oligonucleotide array data from the same tissue (Lee et al. 2003) gave similar results. This is a common problem in quantitation of gene and protein expression, as frequently molecules such as transcription factors which can have the greatest influence on cellular activities are expressed at levels that make them difficult to detect. Our analysis suggests that it is not the core PCD pathways that undergo the most dramatic changes in gene expression in cancer, but associated pathways which also have the potential to be important in controlling cancer cell survival and cancer progression.

Many of the genes identified in the pcd-cancer set have a previous association with cell death or cancer (Figure 4.2). Our analysis suggests that the genes with known roles in cancer may act through effects on PCD-related cellular processes. It also implicates a number of other genes as having a role in cancer, when no such role was previously described, and suggests that these genes may also function through PCD. In future, it would be beneficial to examine the pcd-cancer set for enrichment of known cancer genes as well as known programmed cell death genes, using a defined set of known cancer genes such as compiled in a recent cancer census (Futreal et al. 2004).

Several classes of genes in the PCD-upregulated subset (Table 4.2) are related to the process of autophagy. Cathepsins and other lysosomal proteins are responsible for the degradation of cellular proteins and organelles in the autolysosome (reviewed in Bursch 2001), and cytoskeletal proteins are necessary for the large amount of vesicular movement involved in autophagy (Bursch et al. 2000). Carbonate anhydrases and V-ATPases are responsible for cellular pH regulation, and proper acidification of vesicles is required for autophagy (Yamamoto et al. 1998). pH regulation is also particularly important in hypoxic cells, and V-ATPase expression can prevent apoptosis due to cellular acidosis (reviewed in Izumi et al. 2003).

Multiple signaling pathway components and adaptor proteins were identified in the pcd-cancer set, several with no known function; not surprisingly, this suggests that cell signaling is important both in PCD and in cancer. Interestingly, several metalloendopeptidase inhibitors (TIMPs) were identified as downregulated in PCD and changing in expression in cancer. TIMPs inhibit matrix metalloproteases which promote metastasis by breaking down the extracellular matrix. Although this activity could suppress tumorigenesis, the role and expression of TIMPs is complex (Noel et al. 2004), and they have also been described as anti-apoptotic (Hojilla et al.

2003). The downregulation of these genes during PCD in *Drosophila* and upregulation in cancers of the brain and pancreas suggests that their role in apoptosis may be important in cancer.

Functions associated with ribosomes and protein synthesis were found to be enriched both in the pcd-upregulated and the pcd-downregulated sets. This may reflect a potential bias in this analysis, as genes which are highly conserved will be more likely to appear as orthologs and thus become highly represented in the pcd-cancer set. Many genes involved in basic cellular processes, such as ribosomal proteins, are also often highly expressed and therefore more likely to be differentially expressed in any system. Their appearance in this set, as well as in many other studies of gene expression, may simply reflect these properties. However, changes in expression of genes associated with protein translation, including ribosomal proteins, may actually be related to changes in cellular growth pathways (Martin et al. 2004) and thus may truly be an important indicator of changes in pathway regulation.

4.4.3 Conclusions

My novel cross-species analysis of programmed cell death in cancer demonstrates the use of genome-scale data from multiple systems to better understand the roles of genes and pathways in complex processes. Such potentially powerful approaches are likely to become more common as the amount of genome, gene expression, and other large-scale molecular data increases and demands new analysis methodologies. The results described here suggest that pathways associated with autophagy may be regulated in cancer, and thus point to the importance of studying autophagy gene expression in cancer.

Chapter 5: The autophagy gene MAP1LC3B in cancer

A version of this chapter has been submitted for publication.

Pleasance, E.D., Jones, S.J.M, and Gorski, S.M.. 2005. The autophagy gene MAP1LC3B is downregulated in multiple human cancers. Manuscript submitted.

5.1 Introduction

Cancer is characterized by multiple molecular genetic alterations, including changes in growth regulation, cellular aging, and cell death pathways (reviewed in Hanahan and Weinberg 2000). Recently, the cellular mechanism of autophagy has been implicated in cancer, and represents an additional pathway whose modulation potentially impacts cancer progression (Edinger and Thompson 2003). However, the role of autophagy in cancer is enigmatic; there is evidence for its role in cancer inhibition through regulating cell growth and death, but also for contributing to cancer cell survival in conditions of nutrient- or oxygen-limitation or chemotherapy and radiation (reviewed in Ogier-Denis and Codogno 2003). The gene Beclin 1, which is essential for autophagy, has been demonstrated to be a tumor suppressor which can influence cell growth and tumorigenesis in cell lines and mice (Qu et al. 2003; Yue et al. 2003). To date, no other autophagy genes have been specifically studied in cancer.

To initiate further investigation into the role of autophagy in cancer and examine autophagy's possible dual role in oncogenesis, I describe in this chapter a comprehensive gene expression approach utilizing publicly available data. This approach was chosen and the analysis focused on the gene MAP1LC3B, or LC3, because previous studies indicated LC3 is the only autophagy gene demonstrated to be transcriptionally regulated in mammals. LC3 is indeed unique in this regard since regulation of other mammalian autophagy genes is not known to be at the transcriptional level. The localization of the LC3 protein to autophagosomes is used as a marker for autophagy (Mizushima et al. 2004), and importantly the level of expression of LC3 mRNA is correlated both with this localization and with onset of autophagy (Kanzawa et al. 2004). Additionally, transcriptional regulation of LC3 is evolutionarily conserved. The yeast LC3 homolog Atg8 is transcriptionally regulated in autophagy (Kirisako et al. 1999), multiple

Arabidopsis thaliana Atg8 homologs show increased expression during starvation when autophagy is observed (Rose et al. 2005), and expression of the *Drosophila melanogaster* LC3 homolog CG32675 correlates with autophagy in the midgut (E. Pleasance, based on data from Li and White 2003) and fat body (G. Juhasz and M. Sass, unpublished data), and autophagy stimuli including starvation and chemotherapeutic treatment increase LC3 expression in human systems (Kanzawa et al. 2004; Kanzawa et al. 2005; Nara et al. 2002). This conservation suggests that transcriptional control is an important mechanism of regulating LC3 function. Together, the above observations suggest that levels of LC3 gene expression correlate with autophagy levels. Thus, the expression patterns of LC3 not only may provide insight into the roles of LC3 itself, but also may provide insight into the possible roles of autophagy in different cancerous tissues, subtypes and stages. Results show that LC3 is indeed differentially expressed in multiple cancers, consistent with the tissues in which Beclin 1 effects are seen.

5.2 Methods

5.2.1 Processing of microarray expression data

Affymetrix oligonucleotide microarray data from Ramaswamy et al (2001) were downloaded from the authors' website. LC3 was represented by probes RC_AA283759_at and W28106_at on array Hu35KsubA, and Beclin 1 was represented by probe L38932_at on array Hu6800, as annotated by NetAffx (http://www.affymetrix.com/analysis/, March 30 2005). Only samples for which comparable normal and cancer tissues were available were considered in this analysis. Intensities were provided as absolute readings; values were converted to log(2) and, for the two LC3 probes, averaged. Affymetrix oligonucleotide microarray data from Bhattacharjee et al (2001) were downloaded from the authors' website. LC3 was represented by probe 39370_at on array U95A, as annotated by NetAffx (Feb 10 2005). Intensities were provided as absolute

readings, and log(2) was applied. For all cDNA and oligonucleotide microarray data, p-values for differences in group means were calculated with two-sample two-tailed Welch's t-tests.

Two-color cDNA microarray data from Chen et al (2002b) and Sorlie et al (2003) were obtained from the Stanford Microarray Database (http://genome-www5.stanford.edu/) as log(2) normalized ratios. Both studies used cDNA arrays (Perou et al. 2000), and for all samples standardized RNA was used as a reference in hybridization. LC3 was represented by two clones, IMAGE:796650 and IMAGE:795604, mapping to UniGene Hs.121849. It was confirmed by BLASTN that ESTs from these clones correctly align to LC3 RefSeq record NM_022818. The average of the log-ratio values for the two LC3 probes was computed. As full-length clones are spotted on these cDNA microarrays, there is potential for cross-hybridization of the MAP1LC3B probe sequence with other genes. BLASTN analysis of the LC3 RefSeq sequence compared to all RefSeq mRNA sequences identified a ~100 bp region of 80% identity between MAP1LC3B and LC3A gene sequences, as well as a region of >300 bp of >90% identity between MAP1LC3B and two hypothetical genes, only one of which has expression evidence. This indicates that a low-to-moderate level of cross-hybridization may be expected (Evertsz et al. 2001), which could alter the observed hybridization intensities.

5.2.2 Processing of SAGE expression data

The LC3 gene was mapped to sequences in the RefSeq, UniGene (via SAGE Genie, Boon et al. 2002), MGC, and Ensembl databases using DiscoverySpace (Varhol, Zuyderduyn et al, unpubl.). This identified the 14 bp tag (CATG)TGAGTGGTCA, at the 3'-most *Nla*III site in the 3' UTR, as unambiguously representing the LC3 gene. The tag (CATG)CTGAGGGGTG was identified as the next 3'-most tag corresponding to the second N*la*III site from the 3' end of the LC3 transcript. SAGE libraries were downloaded from the Cancer Genome Anatomy Project

website (http://cgap.nci.nih.gov/) as described in Chapter 4. Breast cancer libraries were classified based on descriptions in the associated publications (Lal et al. 1999; Porter et al. 2003; Porter et al. 2001); libraries that could not clearly be classified were excluded. P-values were determined both by summing tag counts in each group and using the Audic-Claverie method (Audic and Claverie 1997), which may produce overly optimistic p-values when used on summed libraries, and also using a two-sample two-tailed Welch's t-tests, which may be overly conservative for SAGE data (Baggerly et al. 2003).

5.2.3 Tag mapping and EST analysis of secondary LC3 tag

The 14 bp tag (CATG)CTGAGGGGTG, which corresponds to the *Nla*III site just upstream of the 3'-most *Nla*III site in LC3, also corresponds to an upstream tag in the gene FBXW11. This ambiguity was resolved by examining the 15th base pair in the ditags (Colinge and Feger 2001b) from which the observed SAGE tags were derived. In the LC3 gene, the sequence CATGCTGAGGGGTG is followed by "A"; in the FBXW11 gene, it is followed by "G". Kornelia Polyak, author of the breast cancer SAGE studies, kindly provided raw SAGE sequences. These were processed to extract ditags, and ditags containing the tag of interest that were of length 22-26 (30-34 including flanking CATG sites) were analyzed to determine the 15th base. In all cases where suitable length ditags were available (between 1 and 6 cases per library), the 15th base pair was "A", indicating unambiguous representation of the LC3 mRNA transcript.

The locations of the primary and secondary LC3 tags in the LC3 3' UTR were visualized in the UCSC Genome Browser (http://genome.ucsc.edu/), and it was noted that the tags reside ~100 bp from the 3' end of the transcript, and only 10 bp separates the two tags. For these two tags to be produced legitimately, there must be two forms of the LC3 transcript, one of which ends after the secondary tag but before the primary tag. To look for other evidence of such a

transcript, which would only differ from the canonical, full-length transcript by ~100 bp, EST alignments at UCSC were examined for 3' ESTs which ended between the two tags. 16 such ESTs were found, 10 of which had discernable poly-A tails and all of which ended in the short region between the two tags. Thus, ESTs and SAGE data indicate that a second, slightly shorter, transcript of LC3 is expressed.

5.3 Results and discussion

5.3.1 Overview of LC3 regulation in multiple cancer and normal tissues

To gain a comprehensive view of LC3 expression in normal and cancer tissues, I first examined oligonucleotide microarray data produced by Ramaswamy et al (2001) which profiles 90 normal and 190 cancer samples obtained from tissue biopsies prior to cancer treatment (Table 5.1). This dataset is particularly suitable for LC3 expression analysis as the Affymetrix oligonucleotide microarray platform used in this study produces consistent results (Nimgaonkar et al. 2003), and the LC3 gene can be uniquely identified by probes on this array. In addition, the number of cancer and normal samples available for each tissue combined with the variety of tissues allows for a comprehensive comparison and statistical assessment of LC3 expression patterns. Overall, a trend is observed toward decreased LC3 expression in cancerous tissues, but this varies substantially between tissues (Figure 5.1). Decreased LC3 expression ($p<0.05$, Welch's t-test), suggesting a decrease in levels of autophagy, is particularly pronounced in cancer of the bladder, colon, breast and ovary. These observations are consistent with reports of decreased Beclin 1 protein expression in cancer of the breast and ovary (Liang et al. 1999).

In contrast to the observation of decreased LC3 expression in some cancers, several other cancer types do not show differences in LC3 expression, including brain, kidney, and prostate cancers, and leukemias and lymphomas. Overall expression of LC3 in the brain is higher than in

many tissues, as has been observed previously (Mizushima et al. 2004; Tanida et al. 2004), but no difference in LC3 expression is seen in brain cancer compared to normal brain tissue, or in medulloblastoma or glioblastoma subtypes. This is confirmed by my analysis of alternate expression data (Boon et al. 2004; Lal et al. 1999; Siu et al. 2001) in medulloblastomas, glioblastomas, epyndymomas, astrocytomas, and meningiomas (data not shown). Although glioma cell lines have previously been observed to undergo autophagic cell death in response to therapy and thus have autophagic capacity (Kanzawa et al. 2004; Kanzawa et al. 2003), there is no evidence from LC3 expression for a change in autophagy in gliomas. Similarly, no difference in LC3 expression is seen in prostate cancer despite that autophagic death that has been observed in the LnCAP prostate cancer cell line in response to neuregulin (Tal-Or et al. 2003). This may reflect that, in the absence of certain chemotherapeutic agents, autophagic cell death does not occur and thus there is no survival advantage to the cancer cells in reducing autophagy. Additionally, there may be a difference between the behavior of cell lines and cancers in vivo; indeed, I have observed disparities in LC3 expression in some cell lines compared to bulk samples from the same tissue (data not shown).

In only one tissue, the pancreas, was observed an increase rather than a decrease in LC3 expression in cancer. The higher expression of LC3 in pancreatic adenocarcinomas compared to normal pancreatic tissues may simply reflect an unusually low expression of LC3 in normal pancreas compared to any other normal tissue examined. Alternatively, it may indicate that expression of LC3 has been elevated, potentially reflecting a different role for autophagy in this type of cancer. Although rat pancreatic carcinomas have decreased autophagic capacity, which is a different pattern than we observe in human pancreatic adenocarcinomas, early stage rat pancreatic cancers show increased levels of autophagy (Toth et al. 2002). This may imply that

the proliferation of this type of cancer, which also has a particularly high mortality rate (Lowenfels and Maisonneuve 2004), can be enhanced by increased autophagy levels.

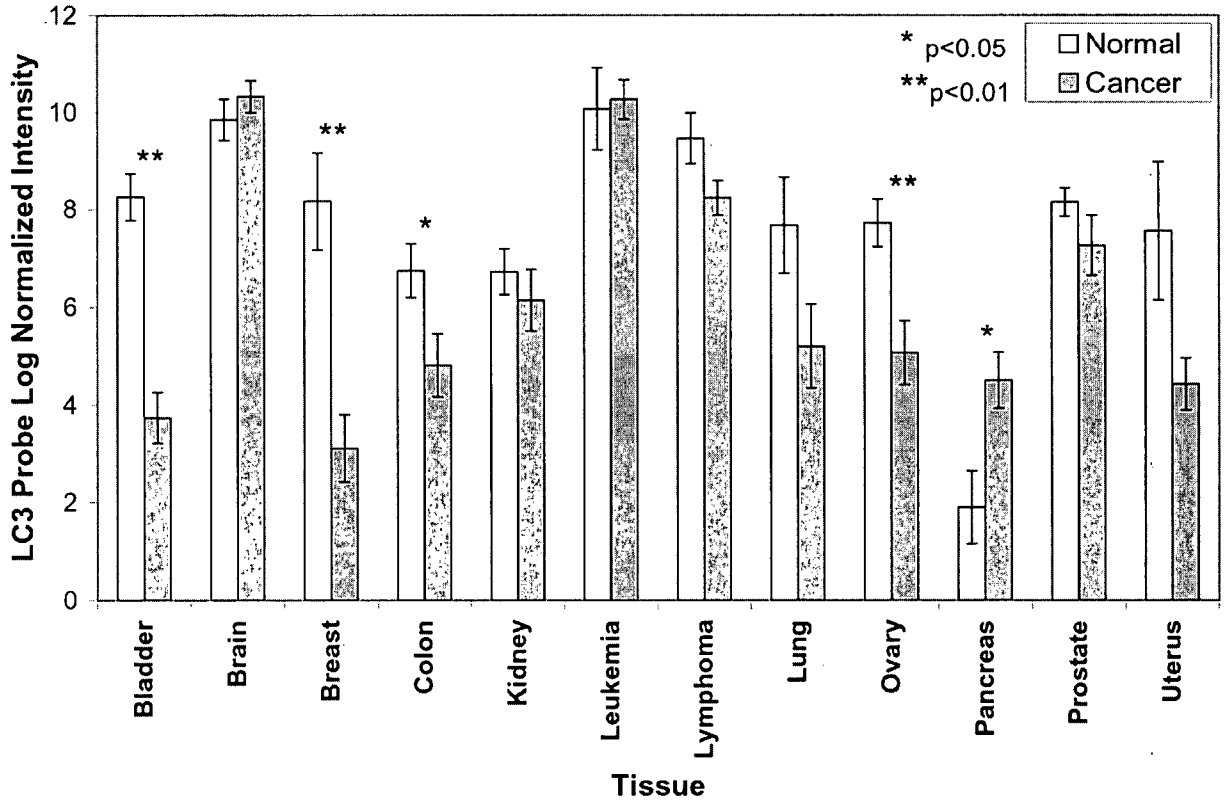

Figure 5.1 LC3 expression in multiple normal and cancer tissues.
Expression values are log transformed from Affymetrix oligonucleotide microarray data published by Ramaswamy et al (2001). P-values are calculated with a Welch's t-test for each cancer with respect to equivalent normal tissue; error bars represent standard error. Number and type of samples represented for each tissue are given in Table 5.1.

Table 5.1 Composition of normal and cancer samples shown in Figures 5.1 and 5.5.

Tissue[a]	Cancer status[a]	Type[b]	# Samples
Bladder	Normal	Normal bladder tissue	7
	Cancer	Transitional cell carcinoma	11
Brain	Normal	Cerebellum	3
		Whole brain	5
	Cancer	Glioblastoma	10
		Medulloblastoma	10
Breast	Normal	Normal breast tissue	5
	Cancer	Breast carcinoma in situ	7
		Invasive breast carcinoma	4
Colon	Normal	Normal colon	11
	Cancer	Colorectal carcinoma	11
Kidney	Normal	Normal kidney	12
	Cancer	Renal cell carcinoma	11
Leukemia	Normal	Peripheral blood polymorphonuclear leukocytes	3
		Peripheral blood monocytes	2
	Cancer	Acute myelogenous leukemia	10
Lymphoma	Normal	Germinal centers	6
	Cancer	Follicular lymphoma	11
		Large B-cell lymphoma	11
Lung	Normal	Normal lung	7
	Cancer	Lung adenocarcinoma	11
Ovary	Normal	Normal ovary	4
	Cancer	Ovarian adenocarcinoma	11
Pancreas	Normal	Normal pancreas	10
	Cancer	Pancreatic adenocarcinoma	11
Prostate	Normal	Normal prostate	9
	Cancer	Prostate adenocarcinoma	10
Uterus	Normal	Normal uterus	6
	Cancer	Uterus adenocarcinoma	10

[a] As in Figures 5.1 and 5.5
[b] As described and analyzed in Ramaswamy et al (2001)

5.3.2 LC3 in cancer stages

To corroborate and expand on the observations above, I examined complementary expression data from different studies, using different expression technologies. In particular, we were interested to observe whether there were any stage-specific differences in autophagy gene expression in cancers, as one hypothesis is that increased autophagy enhances cancer cell survival in early-stage preangiogenic tumors, while later in cancer development autophagy is detrimental to cancer cells and may lead to cell death. Breast cancer was of particular interest, as

LC3 shows a difference in mRNA expression (Figure 5.1) and Beclin 1 differs in protein expression in this tissue (Liang et al. 1999), and Beclin 1's tumor suppressor activity has been demonstrated specifically in breast cancer cell lines. Liver and lung cancers were also of interest due to the observation of increased rates of these cancers in Beclin 1 knockout mice (Qu et al. 2003; Yue et al. 2003).

Data generated from liver cancers (Chen et al. 2002b) shows a decrease in expression of LC3 during cancer progression ($p<0.001$, Welch's t-test; Figure 5.2), consistent with a previously observed decrease in autophagy in liver cancer models (Schwarze and Seglen 1985), and the presence of hepatocellular carcinomas in Beclin 1 knockout mice (Qu et al. 2003; Yue et al. 2003). Unlike the trend observed in pancreatic cancer (Toth et al. 2002), no increase in autophagy in early stages was seen. This liver cancer dataset was derived from cDNA microarrays, and unlike oligonucleotide microarrays, the probe design of cDNA microarrays is such that the expression of LC3 observed may be influenced by the expression of other genes. Even so, the data are consistent with reduced LC3 expression levels in hepatocellular carcinoma.

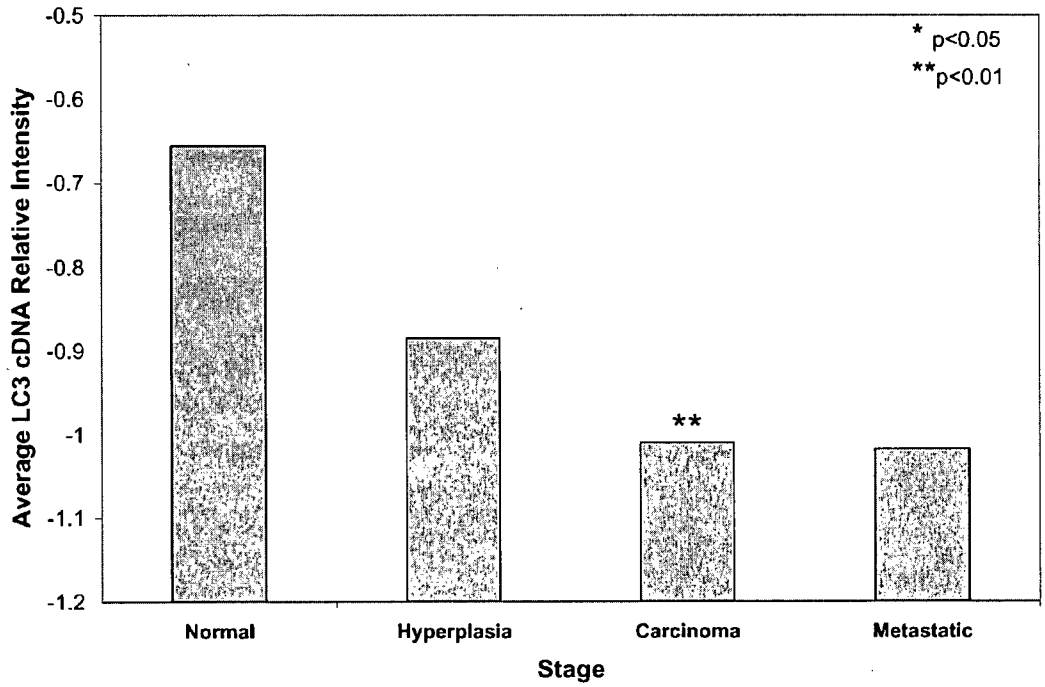

Figure 5.2 LC3 expression in liver cancer stages.
Expression values in normal, early, and late-stage liver cancers are averages of two cDNA microarray probes, with intensities given relative to reference RNA, from data published by Chen et al (2002b). P-values are calculated with a Welch's t-test for each stage relative to normal. Number of samples represented for each stage are: 76 (normal), 7 (hyperplasia), 104 (carcinoma) and 7 (metastatic).

5.3.3 LC3 in breast cancer progression

To investigate a possible stage-specific role for autophagy in cancer, I analyzed serial analysis of gene expression (SAGE) data from breast cancer samples representing cancer progression (Lal et al. 1999; Porter et al. 2003; Porter et al. 2001). Genes can have multiple SAGE tags that can represent alternative transcripts. For LC3, both the most common SAGE tag for LC3 and a secondary tag were sufficiently abundant in both normal and cancerous breast libraries to be statistically significant, and our analysis shows that these tags unambiguously represent expression of the LC3 gene. Consistent with the data shown in Figure 5.1, levels of LC3 were observed to decrease in ductal carcinoma in situ, as well as in invasive and metastatic breast carcinomas (Figure 5.3A; $p<0.05$ for Welch's t-test comparing each stage to normal).

Interestingly, however, the secondary SAGE tag for LC3 shows a similar decrease in DCIS and invasive stages, but increases again at the metastatic carcinoma stage (Figure 5.3B; $p<0.05$ for Welch's t-test comparing metastatic to DCIS). Additionally, invasive library INV5, which shows higher expression of the secondary LC3 tag (Figure 5.3B), is derived from the same patient as the metastatic library MET1. Analysis of Expressed Sequence Tag (EST) evidence suggests that this secondary SAGE tag represents a shorter version of the LC3 mRNA transcript which lacks ~100 bp of the 3' UTR (data not shown), possibly due to a difference in usage of polyadenylation recognition sequences. This is an example where subtle differences in transcript structure can be observed with SAGE, where such differences would be unlikely to be observed with microarray technologies. While we cannot rule out the possibility that the difference in expression level of the LC3 transcript variant is due to the effects of cancer treatments, there are numerous examples of genes with alternative splice forms which are differentially associated with cancer progression (Kirschbaum-Slager et al. 2004). 3' UTR sequences are known to contain regulatory elements that can control mRNA levels (Yan and Marr 2005), which may explain the difference in expression level. The differential regulation of this transcript variant in metastatic breast tissue suggests that LC3 may have a different role in different stages in breast cancer.

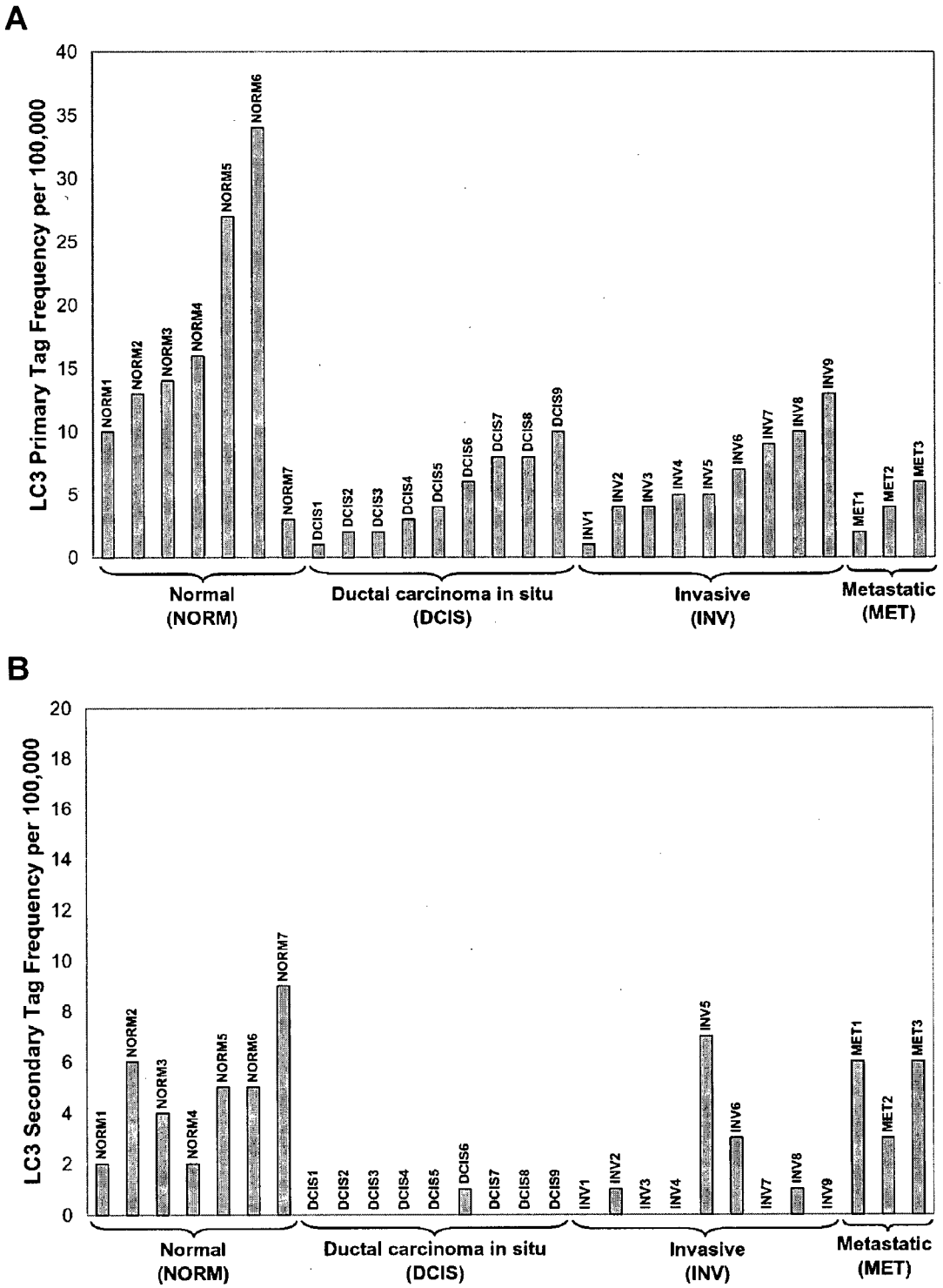

Figure 5.3 LC3 expression in breast cancer progression.
(A) LC3 primary SAGE tag frequency in breast cancer progression SAGE libraries from CGAP (Lal et al. 1999; Porter et al. 2003; Porter et al. 2001). Each bar represents a SAGE library; libraries are sorted from lowest to highest LC3 expression in each category. (B) LC3 secondary SAGE tag frequency in breast cancer progression SAGE libraries. Note that INV5 and MET1 are derived from the same patient, and NORM7 represents a breast hyperplasia.

5.3.4 LC3 in cancer subtypes

I also found evidence for a cancer subtype-specific pattern of LC3 expression. Analyzing oligonucleotide microarray data profiling specific histological subtypes of lung cancer (Bhattacharjee et al. 2001), a significant decrease in LC3 expression in lung adenocarcinoma was observed (Figure 5.4A), which was not significant in the dataset analyzed for Figure 5.1 due to a smaller number of samples. Interestingly, the expression of LC3 is not consistent between lung cancer subtypes. Small cell lung carcinomas and squamous carcinomas show the greatest decrease in expression. In contrast, LC3 expression in carcinoid tumors, a less common and lower grade subtype (Kufe et al. 2003), is in fact higher than in normal lung samples. I also examined breast cancer expression profiles from cDNA arrays (Sorlie et al. 2003) and found differences in LC3 expression in breast cancer subtypes (Figure 5.4B). In particular, significantly decreased expression of LC3 in the luminal A subtype is observed; this subtype also shows the lowest mortality rate. Results from both of these datasets suggest that autophagy may not only have different functions in different cancer stages, but may in fact be differentially regulated in different cancer subtypes.

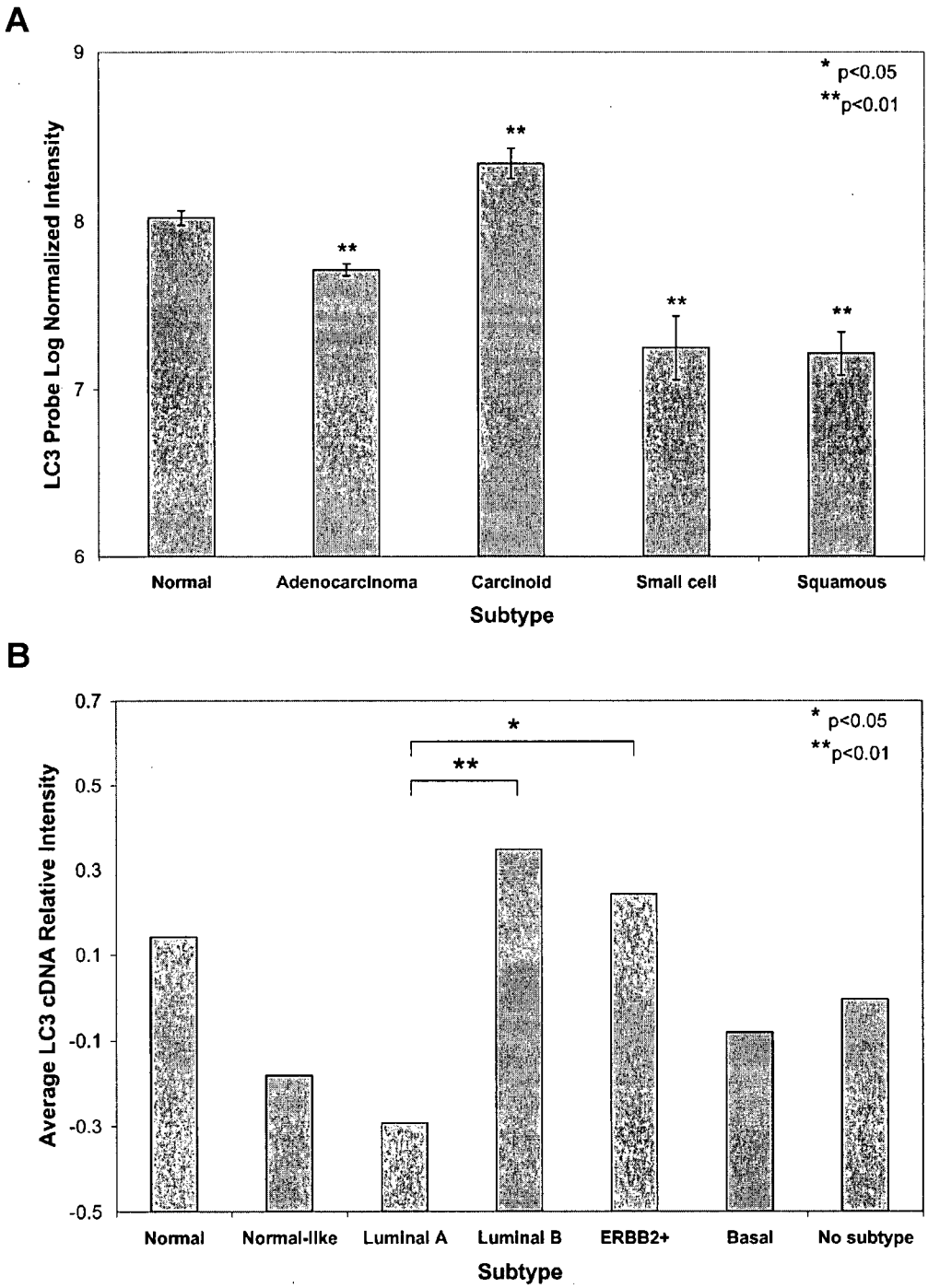

Figure 5.4 LC3 expression in lung and breast cancer subtypes.
(A) LC3 expression in lung cancer subtypes; values are log transformed from Affymetrix oligonucleotide microarray data published by Bhattacharjee et al (2001). P-values are calculated with a Welch's t-test with respect to normal lung tissue expression for each cancer type; error bars represent standard error. Number of samples represented: 17 (normal), 139 (adenocarcinoma), 20 (carcinoid), 6 (small cell), 21 (squamous). (B) LC3 expression in breast cancer subtypes; values are averages of two cDNA microarray probes, with intensities given relative to reference RNA, from data published by Sorlie et al (2003) where subtypes were also defined. P-values are calculated with a Welch's t-test comparing each subtype. Number of samples represented for each subtype are: 3 (normal), 6 (normal-like), 28 (luminal A), 11 (luminal B), 11 (ERBB2+), 19 (basal), 43 (no subtype given).

5.3.5 LC3 and Beclin 1 expression

My observations of LC3, the second autophagy gene to be investigated in cancer, are strikingly consistent with the previously reported behavior of the autophagy gene Beclin 1 in cancer, and provide strong support for a role of autophagy in this disease. Decreased expression of the autophagy gene LC3 in breast, ovarian, lung, and liver cancer was observed, while previous work showed that protein expression of the autophagy gene Beclin 1 is decreased in breast and ovarian cancer, and that mice lacking one copy of Beclin 1 develop lung and liver cancers as well as breast hyperplasias. My analysis thus lends support to the hypothesis that it is indeed Beclin 1's role in autophagy that is responsible for its tumor suppressor effects, as the major autophagy marker LC3 has an expression pattern suggesting decreased autophagy in cancers – and this is observed in the same tissues affected by decreased Beclin 1. Unlike the data available for LC3, there is no evidence for transcriptional regulation of mammalian Beclin 1 or its yeast counterpart. Analysis of Beclin 1 expression in multiple cancers (Figure 5.5) confirms that Beclin 1 mRNA levels are not consistent with protein levels (Aita et al. 1999), and when multiple testing is accounted for, the differences in Beclin 1 expression in cancers are insignificant. This emphasizes the importance of evaluating protein expression of Beclin 1 when studying the role of this gene in autophagy and cancer.

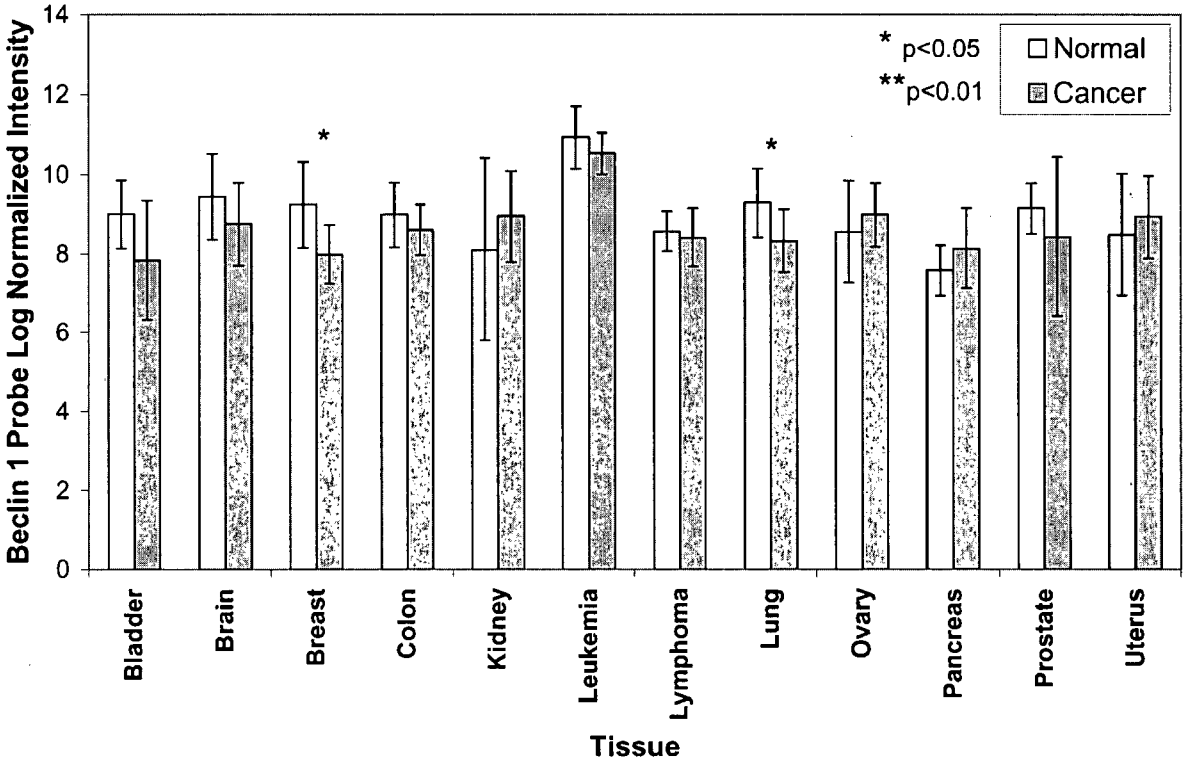

Figure 5.5 Beclin 1 expression in multiple normal and cancer tissues.
Expression values are log transformed from Affymetrix oligonucleotide microarray data published by Ramaswamy et al (2001). P-values are calculated with a Welch's t-test for each cancer with respect to equivalent normal tissue; error bars represent standard error. Number of samples represented for each tissue are given in Table 5.1.

5.3.6 Conclusions

In summary, we have analyzed for the first time the autophagy gene LC3 in the context of multiple human cancers. We have observed decreased expression of LC3 in several cancers, corroborating previous findings with Beclin 1 and indicating that downregulation of autophagy may indeed be an important step in oncogenesis. Not all cancers show this difference, suggesting that autophagy levels and the potential for autophagic cell death may be an important factor in some but not all tumors. We have also observed differences in LC3 expression during cancer progression, and in different cancer subtypes, lending support to the hypothesis that autophagy may play a dual role in oncogenesis and can be beneficial to cancer cell survival. Overall, these

results underline the importance of better understanding the role of autophagy in oncogenesis, and point to the modulation of the pathways of autophagy and autophagic cell death as a possible means of therapeutic intervention.

Chapter 6: Summary and Conclusions

6.1 Summary

This thesis describes the testing of the hypotheses that gene expression analysis can be used to identify genes potentially involved in programmed cell death, and that related apoptosis and autophagy genes play a role in *Drosophila* PCD and cancer; additionally, this work generates further hypotheses regarding the involvement of other genes in these processes. The aims of this work were therefore to develop and utilize methods for large-scale gene expression analysis, and identify genes and pathways regulated in programmed cell death and cancer. Methods for analyzing SAGE data and assessment of SAGE for gene identification were described and shown to be more effective than other available approaches. These methods were applied to expression data describing programmed cell death in the *Drosophila* salivary gland. The PCD-associated genes identified as differentially expressed in this analysis were used in conjunction with large-scale expression data from human cancers in an analysis of the role of programmed cell death in oncogenesis. Additionally, multiple datasets were examined for expression of the autophagy marker LC3. These analyses identified genes and pathways with potential roles in PCD and cancer.

6.2 Large-scale gene expression analysis with SAGE and other techniques

6.2.1 Measurement and comparison of gene expression

For large-scale gene expression analysis methods to be successful, correct identification of the genes under study is necessary. I found that the standard SAGE procedure has the ability to profile and unambiguously identify all but 7-20% of the transcriptome, depending on the organism under study. Use of different anchoring enzymes or use of longer SAGE tags reduces this intractable fraction to only 5-10% of genes. The ideal SAGE tag length and enzyme differs

for each organism; for instance, longer tags are more beneficial in analysis of human genes than in *Drosophila*. The ability of microarrays to correctly identify genes is not inherent to the procedure, as it is for SAGE, but is greatly dependent on the probes placed on the microarray. In the analysis of cDNA microarray data for the study of autophagy in cancer, it was noted that the probes corresponding to the gene of interest had the potential to match other genes as well, making unambiguous gene profiling difficult.

The vast repositories of gene expression data continuously being collected in databases such as the SAGE catalog at the Cancer Genome Anatomy Project or the array data in the Stanford Microarray Database provide an opportunity to study cellular pathways by data mining. Comparison of expression across experiments performed in different contexts is necessary to make full use of this data. This is difficult when analyzing microarray data, especially if the array probe design is not identical or different references are used for cDNA arrays. As such, in the examination of LC3 expression in cancer, the analysis looked at each microarray experiment individually rather than analyzing the results collectively; this can reduce the power of an analysis, but has advantages such as reduced variability due to differences in procedures. SAGE data, on the other hand, does not depend on array design and does not require a reference; SAGE tag counts are absolute and easily normalized to library size. These properties allowed direct comparison of breast cancer SAGE data from multiple publications, and the success of this comparison was demonstrated by the consistency of expression of the LC3 autophagy gene across different SAGE libraries representing the same stage of breast cancer and the significant differences between stages. Similarly, SAGE libraries from multiple cancer and normal tissues were mined for differential gene expression, and the genes found linked to genes differentially expressed in *Drosophila* PCD. In addition to variability in gene expression due to technical

issues or conditions, however, the contribution of biological variability due to differences between individuals or individual occurrences of cancer must be recognized, as many differences may be measured that may or may not be relevant to the system under study. Recognizing the most relevant expression changes in any system will require continued collection and further refined analysis of gene expression data.

6.2.2 Novel genes and alternative transcripts

Despite intense research focused on gene discovery and genome annotation, the entire gene complement of any organism is not yet established. SAGE and EST analyses have the potential for novel gene discovery to aid in genome annotation, while typical microarrays representing only known genes do not. Indeed, analysis of *Drosophila* salivary gland SAGE and EST data identified many potential novel genes. With an equivalent amount of sequencing, SAGE has a greater potential to identify novel genes than do ESTs, due to greater depth of sampling which reaches genes expressed at lower levels. Many genes of interest in the *Drosophila* salivary gland are observed at levels of less than 5 tags. A significant proportion of *Drosophila* salivary gland ESTs did not correspond to known or predicted genes and could be mapped directly to the genome. *Drosophila* SAGE tags could be used to confirm expression of potential novel genes by mapping directly to ESTs, and in a limited fashion could be mapped directly to the genome, identifying potential novel expressed transcripts. In smaller genomes such as that of *Drosophila*, mapping 14 bp SAGE tags directly and uniquely to the genome is possible in approximately 60% of cases, while in larger vertebrate genomes this is not practical unless longer SAGE tags are used. The finding that only 55-65% of *Drosophila* SAGE tags could be mapped to genes, and less than half of human SAGE tags in CGAP could be mapped

unambiguously, underscores the possibility that there are a large number of expressed transcripts which are not yet annotated.

In addition to the potential number of unannotated gene loci, many alternative transcripts for known genes may not yet have been described. Alternative splicing and alternative 5' and 3' ends increase the number of different transcripts and proteins that can be derived from coding sequences in the genome. cDNA microarrays have little potential for identifying alternative transcripts, as different transcripts are likely to cross-hybridize. Oligonucleotide microarrays, depending on probe design, can potentially target specific transcripts. SAGE has the potential to distinguish alternative transcripts if the variation occurs near the 3' end of the gene. In mapping of *Drosophila* SAGE tags, the ratio of mapped tags to genes was approximately 1.1:1. The situation is even more extreme in human tag mapping, where hundreds of thousands of tags mapped to tens of thousands of genes. This indicates that multiple SAGE tags are extracted from each gene, which may be accounted for in part by alternative transcripts. Indeed, SAGE data from breast cancers not only identified an alternative, shorter LC3 transcript, but indicated a difference in expression of this transcript in metastatic breast cancer. Continued research into identifying both novel genes and novel transcript forms through analysis of large-scale expression data will be necessary for the understanding of cellular regulatory networks.

6.3 Genes and pathways associated with programmed cell death

6.3.1 In Drosophila

Changes in gene expression measured by SAGE in the *Drosophila* salivary gland indicated that significant transcriptional changes are associated with programmed cell death, and that genes known to be involved in autophagic PCD and apoptosis are expressed and transcriptionally regulated. Transcriptional profiling using both ESTs and SAGE was successful

in detecting expression of the ecdysone-triggered transcriptional cascade known to regulate PCD in the salivary gland. Several of these genes were expressed at very low levels, however, indicating that much of the control of PCD may be exercised by genes whose differential expression is difficult to measure; accordingly, some of the many genes that were not found to be differentially expressed due to low expression levels may also play important roles in PCD. Interestingly, in addition to expression of autophagy genes (discussed in Section 6.4.1), many apoptosis genes including multiple caspases, that were not previously known to be involved in salivary gland cell death were expressed prior to autophagic PCD. Death of the salivary gland, although it is autophagic as demonstrated by the accumulation of autophagic vacuoles, cytoskeletal rearrangements, and degradation with little phagocyte involvement, shows marked apoptotic features such as cytoplasm and DNA fragmentation (Jiang et al. 1997; Lee and Baehrecke 2001; Martin and Baehrecke 2004). Interestingly, caspase inhibitors that prevent apoptosis also prevent complete salivary gland destruction, but do not avert autophagic vacuole formation and rearrangements of the actin cytoskeleton (Martin and Baehrecke 2004). The nature of the relationship between apoptosis and autophagy is not known. The two pathways may be independent but triggered by the same upstream signal; for instance, the E93 transcription factor is required for both autophagic and apoptotic morphologies (Lee and Baehrecke 2001). Alternatively or additionally, there may be direct crosstalk between the pathways, as is suggested to be the case in mammals (Yu et al. 2004). The changes in expression of ecdysone cascade genes, autophagy genes, and apoptosis genes suggest an intriguing picture whereby both autophagy and apoptosis are required to correctly and completely remove a tissue by PCD.

Nearly one quarter of the transcripts expressed in the *Drosophila* salivary gland in stages prior to cell death show changes in expression, pointing to a vast reorganization of the

transcriptome preceding PCD. Most likely transcriptional control is mediated in part by the activity of many transcription factors which are observed at low levels and peak several hours before death is initiated. There may be additional upstream factors that peak at even earlier time points and so were not observed in the time frame examined, as is probably the case with the ecdysone cascade initiators EcR/USP and βFTZF-1. These various regulators act on downstream pathways such as those controlling protein synthesis and defense responses, and possibly trigger signaling through kinase cascades including the protein kinase A and Ras pathways. Future research may determine the role of each of these transcription factors and interaction of these signaling pathways in cell death. Not surprisingly, it seems the death of an entire tissue is complex and requires significant cellular changes and regulation.

6.3.2 In cancer

Transcriptional profiling of cancers has become increasingly common, and is carried out with goals of identifying pathways which are altered and understanding the underlying causes of these alterations for the purposes of diagnosis and therapy. My analysis of SAGE data indicates thousands of human genes are altered in expression in cancers of various tissues, representing a broad spectrum of cellular functions. Given the number of changes observed, recognizing the pathways that are related to a specific cellular process such as PCD can be difficult. Using the better defined system of the *Drosophila* salivary gland as a filter to pick out genes more likely to be involved in PCD, my approach identified genes known or suspected of being involved in cell death that are changed in cancer. Additionally, pathways related to pH regulation, the cytoskeleton, protein synthesis, and metalloprotease regulation were implicated in cell death.

Although some changes are consistent between cancers of different tissues, in other cases a gene may show a significant increase in expression in some cancers, and a significant decrease

in others. This may reflect tissue-specific differences that are not directly relevant to the cancer, for instance, a higher or lower basal level of activity of a pathway in that tissue. Alternatively, such differences could be related to variations in cellular environments and growth requirements in different tissues. They may also be specific to the individual cancer or the individual patient, especially in cases where only one or two SAGE libraries, derived only from one or two patients, were available to represent a tissue. The large number of changes observed which are common to multiple cancers represent common mechanisms in cancer which are the most promising targets for generally applicable therapies, whereas genes that behave significantly differently between individuals suggest targets that would require individualized therapies. In future, application of this filtering principle with other gene expression datasets, possibly focused on mammalian systems, may be a powerful tool for dissecting pathway deregulation in cancer.

6.4 Roles of autophagy in PCD and cancer

6.4.1 In Drosophila PCD

Large-scale expression data, both from our *Drosophila* EST and SAGE libraries and from co-published microarray data (Lee et al. 2003) provided the first direct evidence for expression of multiple *Drosophila* homologs of yeast autophagy genes in salivary gland PCD. Many genes in the autophagy pathway were observed and several were differentially expressed, indicating that the majority of the putative autophagy pathway is present and active in the *Drosophila* salivary gland. Though this pathway has not until recently been examined in detail in *Drosophila* (Scott et al. 2004), it is thought to function similarly to the yeast pathway in regulating the formation of autophagic vacuoles, which then engulf and degrade cytoplasmic contents. Not surprisingly, significant changes in the expression of cytoskeletal genes are observed in the SAGE libraries, as the cytoskeleton is required for the increased vesicular movement in

autophagy. Autophagy is observed in situations where, as in the death of the salivary gland, an entire tissue must be removed simultaneously. Death through apoptosis produces apoptotic bodies which must be removed by phagocytes, whereas death by autophagy internalizes much of the required cellular degradation, leading to more efficient destruction and removal of the tissue. To what extent the activity of the autophagic pathway itself is responsible for cell death is as yet an unanswered question.

6.4.2 In cancer

Evidence for the important role of autophagy in cancer is growing. My analysis of the LC3 gene has supported the suggestion that autophagy has a dual role in cancer, as the expression of this autophagy gene differs between tissues, stages, and subtypes of cancer. Marked decreases in LC3 expression are consistently observed in breast tumors and several other cancers. The role of autophagy may be dependent on the cellular environment or the pathways naturally active in a tissue, on other cellular changes such as inhibition of apoptosis pathways, or on external conditions such as lack of oxygen or treatment by chemotherapy. My analysis of genes regulated in both *Drosophila* PCD and cancer identified changes in other cellular machinery such as lysosomal activity, pH regulation and cytoskeletal rearrangements that could relate to activity of the autophagy pathway. Increases in expression of autophagy genes and associated pathways could reflect positive selection within the tumor, where autophagy is advantageous due to its protective effects. Alternatively, increases may reflect the triggering of natural cellular mechanisms which are in place to prevent unwanted cell growth, which the cancer may need to counter to survive. Which of these circumstances may be the case in different cancers will be better understood as the genes and pathways responsible for initiating autophagy are examined further.

6.5 Conclusions

Large-scale gene expression analysis has the potential to discover genes and reveal molecular components of complex cellular processes. SAGE is a powerful technique, facilitating comparative analyses of multiple systems, but has important limitations which must be understood. Analyses of expression data in *Drosophila* PCD and human cancer reveal that pathways of autophagy and apoptosis may act together to execute programmed cell death. Additionally, autophagy is clearly important in oncogenesis, as both core autophagy genes and potentially associated processes are regulated in cancer progression in multiple tissues. Autophagy and programmed cell death represent some of the many aspects of the cellular machine that are essential to homeostasis and must be subverted in tumor cells to permit oncogenesis.

References

Adams, M.D., S.E. Celniker, R.A. Holt, C.A. Evans, J.D. Gocayne, P.G. Amanatides, S.E. Scherer, P.W. Li, R.A. Hoskins, R.F. Galle et al. 2000. The genome sequence of Drosophila melanogaster. *Science* **287:** 2185-2195.

Aita, V.M., X.H. Liang, V.V. Murty, D.L. Pincus, W. Yu, E. Cayanis, S. Kalachikov, T.C. Gilliam, and B. Levine. 1999. Cloning and genomic organization of beclin 1, a candidate tumor suppressor gene on chromosome 17q21. *Genomics* **59:** 59-65.

Akker, S.A., P.J. Smith, and S.L. Chew. 2001. Nuclear post-transcriptional control of gene expression. *J Mol Endocrinol* **27:** 123-131.

Akmaev, V.R. and C.J. Wang. 2004. Correction of sequence-based artifacts in serial analysis of gene expression. *Bioinformatics* **20:** 1254-1263.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403-410.

Ambrosini, G., C. Adida, and D.C. Altieri. 1997. A novel anti-apoptosis gene, survivin, expressed in cancer and lymphoma. *Nat Med* **3:** 917-921.

Andrews, J., G.G. Bouffard, C. Cheadle, J. Lu, K.G. Becker, and B. Oliver. 2000. Gene discovery using computational and microarray analysis of transcription in the Drosophila melanogaster testis. *Genome Res* **10:** 2030-2043.

Anisimov, S.V. and A.A. Sharov. 2004. Incidence of "quasi-ditags" in catalogs generated by Serial Analysis of Gene Expression (SAGE). *BMC Bioinformatics* **5:** 152.

Anisimov, S.V., K.V. Tarasov, D. Riordon, A.M. Wobus, and K.R. Boheler. 2002a. SAGE identification of differentiation responsive genes in P19 embryonic cells induced to form cardiomyocytes in vitro. *Mech Dev* **117:** 25-74.

Anisimov, S.V., K.V. Tarasov, D. Tweedie, M.D. Stern, A.M. Wobus, and K.R. Boheler. 2002b. SAGE identification of gene transcripts with profiles unique to pluripotent mouse R1 embryonic stem cells. *Genomics* **79:** 169-176.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408:** 796-815.

Arbeitman, M.N., E.E. Furlong, F. Imam, E. Johnson, B.H. Null, B.S. Baker, M.A. Krasnow, M.P. Scott, R.W. Davis, and K.P. White. 2002. Gene expression during the life cycle of Drosophila melanogaster. *Science* **297:** 2270-2275.

Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25:** 25-29.

Asker, C., K.G. Wiman, and G. Selivanova. 1999. p53-induced apoptosis as a safeguard against cancer. *Biochem Biophys Res Commun* **265:** 1-6.

Audic, S. and J.M. Claverie. 1997. The significance of digital gene expression profiles. *Genome Res* **7:** 986-995.

Baehrecke, E.H. 2000. Steroid regulation of programmed cell death during Drosophila development. *Cell Death Differ* **7:** 1057-1062.

Baehrecke, E.H. 2002. How death shapes life during development. *Nat Rev Mol Cell Biol* **3:** 779-787.

Baehrecke, E.H. 2003. Autophagic programmed cell death in Drosophila. *Cell Death Differ* **10:** 940-945.

Baehrecke, E.H. 2005. Autophagy: dual roles in life and death? *Nat Rev Mol Cell Biol* **6:** 505-510.

Baggerly, K.A., L. Deng, J.S. Morris, and C.M. Aldaz. 2003. Differential expression in SAGE: accounting for normal between-library variation. *Bioinformatics* **19:** 1477-1483.

Bala, P., R.W. Georgantas, 3rd, D. Sudhir, M. Suresh, K. Shanker, B.M. Vrushabendra, C.I. Civin, and A. Pandey. 2005. TAGmapper: A web-based tool for mapping SAGE tags. *Gene.*

Barrett, T., T.O. Suzek, D.B. Troup, S.E. Wilhite, W.C. Ngau, P. Ledoux, D. Rudnev, A.E. Lash, W. Fujibuchi, and R. Edgar. 2005. NCBI GEO: mining millions of expression profiles--database and tools. *Nucleic Acids Res* **33:** D562-566.

Beaudoing, E., S. Freier, J.R. Wyatt, J.M. Claverie, and D. Gautheret. 2000. Patterns of variant polyadenylation signal usage in human genes. *Genome Res* **10:** 1001-1010.

Beissbarth, T., L. Hyde, G.K. Smyth, C. Job, W.M. Boon, S.S. Tan, H.S. Scott, and T.P. Speed. 2004. Statistical modeling of sequencing errors in SAGE libraries. *Bioinformatics* **20 Suppl 1:** I31-I39.

Bergamini, E., G. Cavallini, A. Donati, and Z. Gori. 2003. The anti-ageing effects of caloric restriction may involve stimulation of macroautophagy and lysosomal degradation, and can be intensified pharmacologically. *Biomed Pharmacother* **57:** 203-208.

Bhattacharjee, A., W.G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette et al. 2001. Classification of human lung carcinomas by mRNA expression

profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci U S A* **98:** 13790-13795.

Blackshaw, S., W.P. Kuo, P.J. Park, M. Tsujikawa, J.M. Gunnersen, H.S. Scott, W.M. Boon, S.S. Tan, and C.L. Cepko. 2003. MicroSAGE is highly representative and reproducible but reveals major differences in gene expression among samples obtained from similar tissues. *Genome Biol* **4:** R17.

Boguski, M.S., T.M. Lowe, and C.M. Tolstoshev. 1993. dbEST--database for "expressed sequence tags". *Nat Genet* **4:** 332-333.

Boheler, K.R. and M.D. Stern. 2003. The new role of SAGE in gene discovery. *Trends Biotechnol* **21:** 55-57; discussion 57-58.

Boon, K., J.B. Edwards, C.G. Eberhart, and G.J. Riggins. 2004. Identification of astrocytoma associated genes including cell surface markers. *BMC Cancer* **4:** 39.

Boon, K., E.C. Osorio, S.F. Greenhut, C.F. Schaefer, J. Shoemaker, K. Polyak, P.J. Morin, K.H. Buetow, R.L. Strausberg, S.J. De Souza et al. 2002. An anatomy of normal and malignant gene expression. *Proc Natl Acad Sci U S A* **99:** 11287-11292.

Brachmann, C.B. and R.L. Cagan. 2003. Patterning the fly eye: the role of apoptosis. *Trends Genet* **19:** 91-96.

Bursch, W. 2001. The autophagosomal-lysosomal compartment in programmed cell death. *Cell Death Differ* **8:** 569-581.

Bursch, W., A. Ellinger, H. Kienzl, L. Torok, S. Pandey, M. Sikorska, R. Walker, and R.S. Hermann. 1996. Active cell death induced by the anti-estrogens tamoxifen and ICI 164 384 in human mammary carcinoma cells (MCF-7) in culture: the role of autophagy. *Carcinogenesis* **17:** 1595-1607.

Bursch, W., K. Hochegger, L. Torok, B. Marian, A. Ellinger, and R.S. Hermann. 2000. Autophagic and apoptotic types of programmed cell death exhibit different fates of cytoskeletal filaments. *J Cell Sci* **113 (Pt 7):** 1189-1198.

Butte, A. 2002. The use and analysis of microarray data. *Nat Rev Drug Discov* **1:** 951-960.

C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282:** 2012-2018.

Castro-Alcaraz, S., V. Miskolci, B. Kalasapudi, D. Davidson, and I. Vancurova. 2002. NF-kappa B regulation in human neutrophils by nuclear I kappa B alpha: correlation to apoptosis. *J Immunol* **169:** 3947-3953.

Chen, G., T.G. Gharib, C.C. Huang, J.M. Taylor, D.E. Misek, S.L. Kardia, T.J. Giordano, M.D. Iannettoni, M.B. Orringer, S.M. Hanash et al. 2002a. Discordant protein and mRNA expression in lung adenocarcinomas. *Mol Cell Proteomics* **1:** 304-313.

Chen, J.J., J.D. Rowley, and S.M. Wang. 2000. Generation of longer cDNA fragments from serial analysis of gene expression tags for gene identification. *Proc Natl Acad Sci U S A* **97:** 349-353.

Chen, X., S.T. Cheung, S. So, S.T. Fan, C. Barry, J. Higgins, K.M. Lai, J. Ji, S. Dudoit, I.O. Ng et al. 2002b. Gene expression patterns in human liver cancers. *Mol Biol Cell* **13:** 1929-1939.

Cheng, G. and J.D. Porter. 2002. Transcriptional profile of rat extraocular muscle by serial analysis of gene expression. *Invest Ophthalmol Vis Sci* **43:** 1048-1058.

Chi, S., C. Kitanaka, K. Noguchi, T. Mochizuki, Y. Nagashima, M. Shirouzu, H. Fujita, M. Yoshida, W. Chen, A. Asai et al. 1999. Oncogenic Ras triggers cell suicide through the activation of a caspase-independent cell death program in human cancer cells. *Oncogene* **18:** 2281-2290.

Clark, T., S. Lee, L. Ridgway Scott, and S.M. Wang. 2002. Computational Analysis of Gene Identification with SAGE. *J Comput Biol* **9:** 513-526.

Clarke, P.G. 1990. Developmental cell death: morphological diversity and multiple mechanisms. *Anat Embryol (Berl)* **181:** 195-213.

Colinge, J. and G. Feger. 2001a. Detecting the impact of sequencing errors on SAGE data. *Bioinformatics* **17:** 840-842.

Colinge, J. and G. Feger. 2001b. Improving SAGE di-tag processing. *Genome Biol* **http://genomebiology.com/2001/2/3/preprint/0002.1**.

Colussi, P.A., L.M. Quinn, D.C. Huang, M. Coombe, S.H. Read, H. Richardson, and S. Kumar. 2000. Debcl, a proapoptotic Bcl-2 homologue, is a component of the Drosophila melanogaster cell death machinery. *J Cell Biol* **148:** 703-714.

Cuervo, A.M. 2004a. Autophagy: in sickness and in health. *Trends Cell Biol* **14:** 70-77.

Cuervo, A.M. 2004b. Autophagy: many paths to the same end. *Mol Cell Biochem* **263:** 55-72.

Curwen, V., E. Eyras, T.D. Andrews, L. Clarke, E. Mongin, S.M. Searle, and M. Clamp. 2004. The Ensembl automatic gene annotation system. *Genome Res* **14:** 942-950.

Dahary, D., O. Elroy-Stein, and R. Sorek. 2005. Naturally occurring antisense: transcriptional leakage or real overlap? *Genome Res* **15:** 364-368.

Day, D.A. and M.F. Tuite. 1998. Post-transcriptional gene regulatory mechanisms in eukaryotes: an overview. *J Endocrinol* **157:** 361-371.

Dinel, S., C. Bolduc, P. Belleau, A. Boivin, M. Yoshioka, E. Calvo, B. Piedboeuf, E.E. Snyder, F. Labrie, and J. St-Amand. 2005. Reproducibility, bioinformatic analysis and power of the SAGE method to evaluate changes in transcriptome. *Nucleic Acids Res* **33:** e26.

Dlamini, Z., Z. Mbita, and M. Zungu. 2004. Genealogy, expression, and molecular mechanisms in apoptosis. *Pharmacol Ther* **101:** 1-15.

Doumanis, J., L. Quinn, H. Richardson, and S. Kumar. 2001. STRICA, a novel Drosophila melanogaster caspase with an unusual serine/threonine-rich prodomain, interacts with DIAP1 and DIAP2. *Cell Death Differ* **8:** 387-394.

Du, C., M. Fang, Y. Li, L. Li, and X. Wang. 2000. Smac, a mitochondrial protein that promotes cytochrome c-dependent caspase activation by eliminating IAP inhibition. *Cell* **102:** 33-42.

Durbin, R. and J. Thierry-Mieg. 1991. A *C. elegans* Database.

Edinger, A.L. and C.B. Thompson. 2003. Defective autophagy leads to cancer. *Cancer Cell* **4:** 422-424.

Evan, G.I. and K.H. Vousden. 2001. Proliferation, cell cycle and apoptosis in cancer. *Nature* **411:** 342-348.

Evans, S.J., N.A. Datson, M. Kabbaj, R.C. Thompson, E. Vreugdenhil, E.R. De Kloet, S.J. Watson, and H. Akil. 2002. Evaluation of Affymetrix Gene Chip sensitivity in rat hippocampal tissue using SAGE analysis. Serial Analysis of Gene Expression. *Eur J Neurosci* **16:** 409-413.

Evertsz, E.M., J. Au-Young, M.V. Ruvolo, A.C. Lim, and M.A. Reynolds. 2001. Hybridization cross-reactivity within homologous gene families on glass cDNA microarrays. *Biotechniques* **31:** 1182, 1184, 1186 passim.

Ewing, B. and P. Green. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8:** 186-194.

Ewing, B., L. Hillier, M.C. Wendl, and P. Green. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8:** 175-185.

FlyBase Consortium. 2002. The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Res* **30:** 106-108.

Friedlander, R.M. 2003. Apoptosis and caspases in neurodegenerative diseases. *N Engl J Med* **348:** 1365-1375.

Friedman, R. and A.L. Hughes. 2001. Gene duplication and the structure of eukaryotic genomes. *Genome Res* **11:** 373-381.

Frisch, S.M. and R.A. Screaton. 2001. Anoikis mechanisms. *Curr Opin Cell Biol* **13:** 555-562.

Fujii, S. and H. Amrein. 2002. Genes expressed in the Drosophila head reveal a role for fat cells in sex-specific physiology. *Embo J* **21:** 5353-5363.

Furuya, D., N. Tsuji, A. Yagihashi, and N. Watanabe. 2005. Beclin 1 augmented cis-diamminedichloroplatinum induced apoptosis via enhancing caspase-9 activity. *Exp Cell Res* **307:** 26-40.

Futreal, P.A., L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M.R. Stratton. 2004. A census of human cancer genes. *Nat Rev Cancer* **4:** 177-183.

Ghobrial, I.M., T.E. Witzig, and A.A. Adjei. 2005. Targeting apoptosis pathways in cancer therapy. *CA Cancer J Clin* **55:** 178-194.

Gorski, S.M., S. Chittaranjan, E.D. Pleasance, J.D. Freeman, C.L. Anderson, R.J. Varhol, S.M. Coughlin, S.D. Zuyderduyn, S.J. Jones, and M.A. Marra. 2003. A SAGE approach to discovery of genes involved in autophagic cell death. *Curr Biol* **13:** 358-363.

Goyal, L., K. McCall, J. Agapite, E. Hartwieg, and H. Steller. 2000. Induction of apoptosis by Drosophila reaper, hid and grim through inhibition of IAP function. *Embo J* **19:** 589-597.

Gozuacik, D. and A. Kimchi. 2004. Autophagy as a cell death and tumor suppressor mechanism. *Oncogene* **23:** 2891-2906.

Graber, J.H., C.R. Cantor, S.C. Mohr, and T.F. Smith. 1999. In silico detection of control signals: mRNA 3'-end-processing sequences in diverse species. *Proc Natl Acad Sci U S A* **96:** 14055-14060.

Greenbaum, D., R. Jansen, and M. Gerstein. 2002. Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts. *Bioinformatics* **18:** 585-596.

Griffith, O.L., E.D. Pleasance, D.L. Fulton, M. Oveisi, M. Ester, A.S. Siddiqui, and S.J. Jones. 2005. Assessment and integration of publicly available SAGE, cDNA microarray, and oligonucleotide microarray expression data for global coexpression analyses. *Genomics* **86:** 476-488.

Griffiths, A.J.F., J.H. Miller, D.T. Suzuki, R.C. Lewontin, and W.M. Gelbart. 1999. *Introduction to Genetic Analysis*. W.H. Freeman & Co., New York.

Gronbaek, K., P.T. Straten, E. Ralfkiaer, V. Ahrenkiel, M.K. Andersen, N.E. Hansen, J. Zeuthen, K. Hou-Jensen, and P. Guldberg. 1998. Somatic Fas mutations in non-Hodgkin's lymphoma: association with extranodal disease and autoimmunity. *Blood* **92:** 3018-3024.

Gygi, S.P., Y. Rochon, B.R. Franza, and R. Aebersold. 1999. Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol* **19:** 1720-1730.

Halaschek-Wiener, J., J.S. Khattra, S. McKay, A. Pouzyrev, J.M. Stott, G.S. Yang, R.A. Holt, S.J. Jones, M.A. Marra, A.R. Brooks-Wilson et al. 2005. Analysis of long-lived C. elegans daf-2 mutants using serial analysis of gene expression. *Genome Res* **15:** 603-615.

Hanahan, D. and R.A. Weinberg. 2000. The hallmarks of cancer. *Cell* **100:** 57-70.

Harding, T.M., K.A. Morano, S.V. Scott, and D.J. Klionsky. 1995. Isolation and characterization of yeast mutants in the cytoplasm to vacuole protein targeting pathway. *J Cell Biol* **131:** 591-602.

Harvey, N.L., T. Daish, K. Mills, L. Dorstyn, L.M. Quinn, S.H. Read, H. Richardson, and S. Kumar. 2001. Characterization of the Drosophila caspase, DAMM. *J Biol Chem* **276:** 25342-25350.

Haverty, P.M., L.L. Hsiao, S.R. Gullans, U. Hansen, and Z. Weng. 2004. Limited agreement among three global gene expression methods highlights the requirement for non-global validation. *Bioinformatics* **20:** 3431-3441.

Hay, B.A., J.R. Huh, and M. Guo. 2004. The genetics of cell death: approaches, insights and opportunities in Drosophila. *Nat Rev Genet* **5:** 911-922.

Hay, B.A., D.A. Wassarman, and G.M. Rubin. 1995. Drosophila homologs of baculovirus inhibitor of apoptosis proteins function to block cell death. *Cell* **83:** 1253-1262.

Hedges, S.B., J.E. Blair, M.L. Venturi, and J.L. Shoe. 2004. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC Evol Biol* **4:** 2.

Hegde, R., S.M. Srinivasula, Z. Zhang, R. Wassell, R. Mukattash, L. Cilenti, G. DuBois, Y. Lazebnik, A.S. Zervos, T. Fernandes-Alnemri et al. 2002. Identification of Omi/HtrA2 as a mitochondrial apoptotic serine protease that disrupts inhibitor of apoptosis protein-caspase interaction. *J Biol Chem* **277:** 432-438.

Hengartner, M.O. 2000. The biochemistry of apoptosis. *Nature* **407:** 770-776.

Hild, M., B. Beckmann, S.A. Haas, B. Koch, V. Solovyev, C. Busold, K. Fellenberg, M. Boutros, M. Vingron, F. Sauer et al. 2003. An integrated gene annotation and transcriptional profiling approach towards the full gene content of the Drosophila genome. *Genome Biol* **5:** R3.

Hipfner, D.R. and S.M. Cohen. 2004. Connecting proliferation and apoptosis in development and disease. *Nat Rev Mol Cell Biol* **5:** 805-815.

Hojilla, C.V., F.F. Mohammed, and R. Khokha. 2003. Matrix metalloproteinases and their tissue inhibitors direct cell fate during cancer development. *Br J Cancer* **89:** 1817-1821.

Hoskins, R.A., C.D. Smith, J.W. Carlson, A.B. Carvalho, A. Halpern, J.S. Kaminker, C. Kennedy, C.J. Mungall, B.A. Sullivan, G.G. Sutton et al. 2002. Heterochromatic sequences in a Drosophila whole-genome shotgun assembly. *Genome Biol* **3:** RESEARCH0085.

Hu, S. and X. Yang. 2000. dFADD, a novel death domain-containing adapter protein for the Drosophila caspase DREDD. *J Biol Chem* **275:** 30761-30764.

Hu, Y., H. Sun, J. Drake, F. Kittrell, M.C. Abba, L. Deng, S. Gaddis, A. Sahin, K. Baggerly, D. Medina et al. 2004. From mice to humans: identification of commonly deregulated genes in mammary cancer via comparative SAGE studies. *Cancer Res* **64:** 7748-7755.

Hui, L., X. Zhang, X. Wu, Z. Lin, Q. Wang, Y. Li, and G. Hu. 2004. Identification of alternatively spliced mRNA variants related to cancers by genome-wide ESTs alignment. *Oncogene* **23:** 3013-3023.

Huminiecki, L., A.T. Lloyd, and K.H. Wolfe. 2003. Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases. *BMC Genomics* **4:** 31.

Iseli, C., B.J. Stevenson, S.J. de Souza, H.B. Samaia, A.A. Camargo, K.H. Buetow, R.L. Strausberg, A.J. Simpson, P. Bucher, and C.V. Jongeneel. 2002. Long-range heterogeneity at the 3' ends of human mRNAs. *Genome Res* **12:** 1068-1074.

Ishii, M., S. Hashimoto, S. Tsutsumi, Y. Wada, K. Matsushima, T. Kodama, and H. Aburatani. 2000. Direct comparison of GeneChip and SAGE on the quantitative accuracy in transcript profiling analysis. *Genomics* **68:** 136-143.

Izumi, H., T. Torigoe, H. Ishiguchi, H. Uramoto, Y. Yoshida, M. Tanabe, T. Ise, T. Murakami, T. Yoshida, M. Nomoto et al. 2003. Cellular pH regulators: potentially promising molecular targets for cancer chemotherapy. *Cancer Treat Rev* **29:** 541-549.

Jasper, H., V. Benes, C. Schwager, S. Sauer, S. Clauder-Munster, W. Ansorge, and D. Bohmann. 2001. The genomic response of the Drosophila embryo to JNK signaling. *Dev Cell* **1:** 579-586.

Jiang, C., E.H. Baehrecke, and C.S. Thummel. 1997. Steroid regulated programmed cell death during Drosophila metamorphosis. *Development* **124:** 4673-4683.

Jiang, C., A.F. Lamblin, H. Steller, and C.S. Thummel. 2000. A steroid-triggered transcriptional hierarchy controls salivary gland cell death during Drosophila metamorphosis. *Mol Cell* **5:** 445-455.

Jochova, J., Z. Zakeri, and R.A. Lockshin. 1997. Rearrangement of the tubulin and actin cytoskeleton during programmed cell death in Drosophila salivary glands. *Cell Death Differ* **4:** 140-149.

Johnson, J.M., J. Castle, P. Garrett-Engele, Z. Kan, P.M. Loerch, C.D. Armour, R. Santos, E.E. Schadt, R. Stoughton, and D.D. Shoemaker. 2003. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **302:** 2141-2144.

Jones, S.J., D.L. Riddle, A.T. Pouzyrev, V.E. Velculescu, L. Hillier, S.R. Eddy, S.L. Stricklin, D.L. Baillie, R. Waterston, and M.A. Marra. 2001. Changes in gene expression associated with developmental arrest and longevity in Caenorhabditis elegans. *Genome Res* **11:** 1346-1352.

Juhasz, G., G. Csikos, R. Sinka, M. Erdelyi, and M. Sass. 2003. The Drosophila homolog of Aut1 is essential for autophagy and development. *FEBS Lett* **543:** 154-158.

Jurka, J. 2000. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet* **16:** 418-420.

Kamio, T., T. Toki, R. Kanezaki, S. Sasaki, S. Tandai, K. Terui, D. Ikebe, K. Igarashi, and E. Ito. 2003. B-cell-specific transcription factor BACH2 modifies the cytotoxic effects of anticancer drugs. *Blood* **102:** 3317-3322.

Kamradt, M.C., M. Lu, M.E. Werner, T. Kwan, F. Chen, A. Strohecker, S. Oshita, J.C. Wilkinson, C. Yu, P.G. Oliver et al. 2005. The small heat shock protein alpha B-crystallin is a novel inhibitor of TRAIL-induced apoptosis that suppresses the activation of caspase-3. *J Biol Chem* **280:** 11059-11066.

Kanuka, H., E. Kuranaga, T. Hiratou, T. Igaki, B. Nelson, H. Okano, and M. Miura. 2003. Cytosol-endoplasmic reticulum interplay by Sec61alpha translocon in polyglutamine-mediated neurotoxicity in Drosophila. *Proc Natl Acad Sci U S A* **100:** 11723-11728.

Kanzawa, T., I.M. Germano, T. Komata, H. Ito, Y. Kondo, and S. Kondo. 2004. Role of autophagy in temozolomide-induced cytotoxicity for malignant glioma cells. *Cell Death Differ* **11:** 448-457.

Kanzawa, T., Y. Kondo, H. Ito, S. Kondo, and I. Germano. 2003. Induction of autophagic cell death in malignant glioma cells by arsenic trioxide. *Cancer Res* **63:** 2103-2108.

Kanzawa, T., L. Zhang, L. Xiao, I.M. Germano, Y. Kondo, and S. Kondo. 2005. Arsenic trioxide induces autophagic cell death in malignant glioma cells by upregulation of mitochondrial cell death protein BNIP3. *Oncogene* **24:** 980-991.

Kent, W.J. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12:** 656-664.

Kern, W., A. Kohlmann, C. Wuchter, S. Schnittger, C. Schoch, S. Mergenthaler, R. Ratei, W.D. Ludwig, W. Hiddemann, and T. Haferlach. 2003. Correlation of protein expression and gene expression in acute leukemia. *Cytometry B Clin Cytom* **55:** 29-36.

Kerr, J.F., A.H. Wyllie, and A.R. Currie. 1972. Apoptosis: a basic biological phenomenon with wide-ranging implications in tissue kinetics. *Br J Cancer* **26:** 239-257.

Kim, H.L. 2003. Comparison of oligonucleotide-microarray and serial analysis of gene expression (SAGE) in transcript profiling analysis of megakaryocytes derived from CD34+ cells. *Exp Mol Med* **35:** 460-466.

Kirisako, T., M. Baba, N. Ishihara, K. Miyazawa, M. Ohsumi, T. Yoshimori, T. Noda, and Y. Ohsumi. 1999. Formation process of autophagosome is traced with Apg8/Aut7p in yeast. *J Cell Biol* **147:** 435-446.

Kirschbaum-Slager, N., G.M. Lopes, P.A. Galante, G.J. Riggins, and S.J. de Souza. 2004. Splicing factors are differentially expressed in tumors. *Genet Mol Res* **3:** 512-520.

Kitanaka, C., K. Kato, R. Ijiri, K. Sakurada, A. Tomiyama, K. Noguchi, Y. Nagashima, A. Nakagawara, T. Momoi, Y. Toyoda et al. 2002. Increased Ras expression and caspase-independent neuroblastoma cell death: possible mechanism of spontaneous neuroblastoma regression. *J Natl Cancer Inst* **94:** 358-368.

Klionsky, D.J., J.M. Cregg, W.A. Dunn, Jr., S.D. Emr, Y. Sakai, I.V. Sandoval, A. Sibirny, S. Subramani, M. Thumm, M. Veenhuis et al. 2003. A unified nomenclature for yeast autophagy-related genes. *Dev Cell* **5:** 539-545.

Kornbluth, S. and K. White. 2005. Apoptosis in Drosophila: neither fish nor fowl (nor man, nor worm). *J Cell Sci* **118:** 1779-1787.

Kufe, D.W., R.E. Pollock, R.R. Weichselbaum, R.C. Bast, T.S. Gansler, J.F. Holland, and E. Frei. 2003. *Cancer Medicine*. BC Decker Inc., Hamilton.

Kuma, A., M. Hatano, M. Matsui, A. Yamamoto, H. Nakaya, T. Yoshimori, Y. Ohsumi, T. Tokuhisa, and N. Mizushima. 2004. The role of autophagy during the early neonatal starvation period. *Nature* **432:** 1032-1036.

Kumar, S. and J. Doumanis. 2000. The fly caspases. *Cell Death Differ* **7:** 1039-1044.

Lal, A., A.E. Lash, S.F. Altschul, V. Velculescu, L. Zhang, R.E. McLendon, M.A. Marra, C. Prange, P.J. Morin, K. Polyak et al. 1999. A public database for gene expression in human cancers. *Cancer Res* **59:** 5403-5407.

Lander, E.S., L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Lash, A.E., C.M. Tolstoshev, L. Wagner, G.D. Schuler, R.L. Strausberg, G.J. Riggins, and S.F. Altschul. 2000. SAGEmap: a public gene expression resource. *Genome Res* **10:** 1051-1060.

Lee, C.Y. and E.H. Baehrecke. 2001. Steroid regulation of autophagic programmed cell death during development. *Development* **128:** 1443-1455.

Lee, C.Y., E.A. Clough, P. Yellon, T.M. Teslovich, D.A. Stephan, and E.H. Baehrecke. 2003. Genome-wide analyses of steroid- and radiation-triggered programmed cell death in Drosophila. *Curr Biol* **13:** 350-357.

Lee, C.Y., C.R. Simon, C.T. Woodard, and E.H. Baehrecke. 2002a. Genetic mechanism for the stage- and tissue-specific regulation of steroid triggered programmed cell death in Drosophila. *Dev Biol* **252:** 138-148.

Lee, C.Y., D.P. Wendel, P. Reid, G. Lam, C.S. Thummel, and E.H. Baehrecke. 2000. E93 directs steroid-triggered programmed cell death in Drosophila. *Mol Cell* **6:** 433-443.

Lee, S., J. Bao, G. Zhou, J. Shapiro, J. Xu, R.Z. Shi, X. Lu, T. Clark, D. Johnson, Y.C. Kim et al. 2005. Detecting novel low-abundant transcripts in Drosophila. *Rna* **11:** 939-946.

Lee, S., T. Clark, J. Chen, G. Zhou, L.R. Scott, J.D. Rowley, and S.M. Wang. 2002b. Correct identification of genes from serial analysis of gene expression tag sequences. *Genomics* **79:** 598-602.

Levine, B. and D.J. Klionsky. 2004. Development by self-digestion: molecular mechanisms and biological functions of autophagy. *Dev Cell* **6:** 463-477.

Li, T.R. and K.P. White. 2003. Tissue-specific gene expression and ecdysone-regulated genomic networks in Drosophila. *Dev Cell* **5:** 59-72.

Liang, X.H., S. Jackson, M. Seaman, K. Brown, B. Kempkes, H. Hibshoosh, and B. Levine. 1999. Induction of autophagy and inhibition of tumorigenesis by beclin 1. *Nature* **402:** 672-676.

Liang, X.H., L.K. Kleeman, H.H. Jiang, G. Gordon, J.E. Goldman, G. Berry, B. Herman, and B. Levine. 1998. Protection against fatal Sindbis virus encephalitis by beclin, a novel Bcl-2-interacting protein. *J Virol* **72:** 8586-8596.

Lipshutz, R.J., S.P. Fodor, T.R. Gingeras, and D.J. Lockhart. 1999. High density synthetic oligonucleotide arrays. *Nat Genet* **21:** 20-24.

Lowe, S.W. and A.W. Lin. 2000. Apoptosis in cancer. *Carcinogenesis* **21:** 485-495.

Lowenfels, A.B. and P. Maisonneuve. 2004. Epidemiology and prevention of pancreatic cancer. *Jpn J Clin Oncol* **34:** 238-244.

Lum, J.J., R.J. DeBerardinis, and C.B. Thompson. 2005. Autophagy in metazoans: cell survival in the land of plenty. *Nat Rev Mol Cell Biol* **6:** 439-448.

Maglott, D.R., K.S. Katz, H. Sicotte, and K.D. Pruitt. 2000. NCBI's LocusLink and RefSeq. *Nucleic Acids Res* **28:** 126-128.

Man, M.Z., X. Wang, and Y. Wang. 2000. POWER_SAGE: comparing statistical tests for SAGE experiments. *Bioinformatics* **16:** 953-959.

Margulies, E.H., S.L. Kardia, and J.W. Innis. 2001. Identification and prevention of a GC content bias in SAGE libraries. *Nucleic Acids Res* **29:** E60-60.

Marino, G. and C. Lopez-Otin. 2004. Autophagy: molecular mechanisms, physiological functions and relevance in human pathology. *Cell Mol Life Sci* **61:** 1439-1454.

Martin, D.E., A. Soulard, and M.N. Hall. 2004. TOR regulates ribosomal protein gene expression via PKA and the Forkhead transcription factor FHL1. *Cell* **119:** 969-979.

Martin, D.N. and E.H. Baehrecke. 2004. Caspases function in autophagic programmed cell death in Drosophila. *Development* **131:** 275-284.

Matsumura, H., S. Reich, A. Ito, H. Saitoh, S. Kamoun, P. Winter, G. Kahl, M. Reuter, D.H. Kruger, and R. Terauchi. 2003. Gene expression analysis of plant host-pathogen interactions by SuperSAGE. *Proc Natl Acad Sci U S A* **100:** 15718-15723.

Melendez, A., Z. Talloczy, M. Seaman, E.L. Eskelinen, D.H. Hall, and B. Levine. 2003. Autophagy genes are essential for dauer development and life-span extension in C. elegans. *Science* **301:** 1387-1391.

Misra, S., M.A. Crosby, C.J. Mungall, B.B. Matthews, K.S. Campbell, P. Hradecky, Y. Huang, J.S. Kaminker, G.H. Millburn, S.E. Prochnik et al. 2002. Annotation of the Drosophila melanogaster euchromatic genome: a systematic review. *Genome Biol* **3:** RESEARCH0083.

Mizushima, N., A. Yamamoto, M. Matsui, T. Yoshimori, and Y. Ohsumi. 2004. In vivo analysis of autophagy in response to nutrient starvation using transgenic mice expressing a fluorescent autophagosome marker. *Mol Biol Cell* **15:** 1101-1111.

Mott, R. 1997. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comput Appl Biosci* **13:** 477-478.

Nara, A., N. Mizushima, A. Yamamoto, Y. Kabeya, Y. Ohsumi, and T. Yoshimori. 2002. SKD1 AAA ATPase-dependent endosomal transport is involved in autolysosome formation. *Cell Struct Funct* **27:** 29-37.

Ng, G. and J. Huang. 2005. The significance of autophagy in cancer. *Mol Carcinog* **43:** 183-187.

Nimgaonkar, A., D. Sanoudou, A.J. Butte, J.N. Haslett, L.M. Kunkel, A.H. Beggs, and I.S. Kohane. 2003. Reproducibility of gene expression across generations of Affymetrix microarrays. *BMC Bioinformatics* **4:** 27.

Noel, A., C. Maillard, N. Rocks, M. Jost, V. Chabottaux, N.E. Sounni, E. Maquoi, D. Cataldo, and J.M. Foidart. 2004. Membrane associated proteases and their inhibitors in tumour angiogenesis. *J Clin Pathol* **57:** 577-584.

Ogier-Denis, E. and P. Codogno. 2003. Autophagy: a barrier or an adaptive response to cancer. *Biochim Biophys Acta* **1603:** 113-128.

Ohsumi, Y. 2001. Molecular dissection of autophagy: two ubiquitin-like systems. *Nat Rev Mol Cell Biol* **2:** 211-216.

Ota, T., Y. Suzuki, T. Nishikawa, T. Otsuki, T. Sugiyama, R. Irie, A. Wakamatsu, K. Hayashi, H. Sato, K. Nagai et al. 2004. Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat Genet* **36:** 40-45.

Paglin, S., T. Hollister, T. Delohery, N. Hackett, M. McMahill, E. Sphicas, D. Domingo, and J. Yahalom. 2001. A novel response of cancer cells to radiation involves autophagy and formation of acidic vesicles. *Cancer Res* **61:** 439-444.

Park, S. and C.D. James. 2005. ECop (EGFR-coamplified and overexpressed protein), a novel protein, regulates NF-kappaB transcriptional activity and associated apoptotic response in an IkappaBalpha-dependent manner. *Oncogene* **24:** 2495-2502.

Park, W.S., R.R. Oh, Y.S. Kim, J.Y. Park, S.H. Lee, M.S. Shin, S.Y. Kim, P.J. Kim, H.K. Lee, N.Y. Yoo et al. 2001. Somatic mutations in the death domain of the Fas (Apo-1/CD95) gene in gastric cancer. *J Pathol* **193:** 162-168.

Pattingre, S., A. Tassa, X. Qu, R. Garuti, X.H. Liang, N. Mizushima, M. Packer, M.D. Schneider, and B. Levine. 2005. Bcl-2 antiapoptotic proteins inhibit beclin 1-dependent autophagy. *Cell* **122:** 927-939.

Pauws, E., A.H. van Kampen, S.A. van de Graaf, J.J. de Vijlder, and C. Ris-Stalpers. 2001. Heterogeneity in polyadenylation cleavage sites in mammalian mRNA sequences: implications for SAGE analysis. *Nucleic Acids Res* **29:** 1690-1694.

Perou, C.M., T. Sorlie, M.B. Eisen, M. van de Rijn, S.S. Jeffrey, C.A. Rees, J.R. Pollack, D.T. Ross, H. Johnsen, L.A. Akslen et al. 2000. Molecular portraits of human breast tumours. *Nature* **406:** 747-752.

Pitti, R.M., S.A. Marsters, D.A. Lawrence, M. Roy, F.C. Kischkel, P. Dowd, A. Huang, C.J. Donahue, S.W. Sherwood, D.T. Baldwin et al. 1998. Genomic amplification of a decoy receptor for Fas ligand in lung and colon cancer. *Nature* **396:** 699-703.

Pleasance, E.D. and S.J.M. Jones. 2005. Evaluation of SAGE tags for transcriptome study. In *SAGE: Current Technologies and Applications* (ed. S.M. Wang). Horizon Bioscience, Wymondham, UK.

Pleasance, E.D., M.A. Marra, and S.J. Jones. 2003. Assessment of SAGE in transcript identification. *Genome Res* **13:** 1203-1215.

Porter, D., J. Lahti-Domenici, A. Keshaviah, Y.K. Bae, P. Argani, J. Marks, A. Richardson, A. Cooper, R. Strausberg, G.J. Riggins et al. 2003. Molecular markers in ductal carcinoma in situ of the breast. *Mol Cancer Res* **1:** 362-375.

Porter, D.A., I.E. Krop, S. Nasser, D. Sgroi, C.M. Kaelin, J.R. Marks, G. Riggins, and K. Polyak. 2001. A SAGE (serial analysis of gene expression) view of breast tumor progression. *Cancer Res* **61:** 5697-5702.

Posey, K.L., L.B. Jones, R. Cerda, M. Bajaj, T. Huynh, P.E. Hardin, and S.H. Hardin. 2001. Survey of transcripts in the adult Drosophila brain. *Genome Biol* **2:** RESEARCH0008.

Pylouster, J., C. Senamaud-Beaufort, and T.E. Saison-Behmoaras. 2005. WEBSAGE: a web tool for visual analysis of differentially expressed human SAGE tags. *Nucleic Acids Res* **33:** W693-695.

Qu, X., J. Yu, G. Bhagat, N. Furuya, H. Hibshoosh, A. Troxel, J. Rosen, E.L. Eskelinen, N. Mizushima, Y. Ohsumi et al. 2003. Promotion of tumorigenesis by heterozygous disruption of the beclin 1 autophagy gene. *J Clin Invest* **112:** 1809-1820.

Quackenbush, J. 2001. Computational analysis of microarray data. *Nat Rev Genet* **2:** 418-427.

Quere, R., L. Manchon, M. Lejeune, O. Clement, F. Pierrat, B. Bonafoux, T. Commes, D. Piquemal, and J. Marti. 2004. Mining SAGE data allows large-scale, sensitive screening of antisense transcript expression. *Nucleic Acids Res* **32:** e163.

Ramaswamy, S., P. Tamayo, R. Rifkin, S. Mukherjee, C.H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J.P. Mesirov et al. 2001. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci U S A* **98:** 15149-15154.

Rampino, N., H. Yamamoto, Y. Ionov, Y. Li, H. Sawai, J.C. Reed, and M. Perucho. 1997. Somatic frameshift mutations in the BAX gene in colon cancers of the microsatellite mutator phenotype. *Science* **275:** 967-969.

Reed, J.C. 2002. Apoptosis-based therapies. *Nat Rev Drug Discov* **1:** 111-121.

Reiss, U., B. Oskouian, J. Zhou, V. Gupta, P. Sooriyakumaran, S. Kelly, E. Wang, A.H. Merrill, Jr., and J.D. Saba. 2004. Sphingosine-phosphate lyase enhances stress-induced ceramide generation and apoptosis. *J Biol Chem* **279:** 1281-1290.

Remm, M., C.E. Storm, and E.L. Sonnhammer. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* **314:** 1041-1052.

Rhodes, D.R., J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey, and A.M. Chinnaiyan. 2004. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc Natl Acad Sci U S A* **101:** 9309-9314.

Richardson, H. and S. Kumar. 2002. Death to flies: Drosophila as a model system to study programmed cell death. *J Immunol Methods* **265:** 21-38.

Riddle, D.L., T. Blumenthal, B.J. Meyer, and J.R. Preiss. 1997. *C. Elegans II.* Cold Spring Harbor Laboratory Press, New York.

Robinson, S.J., D.J. Cram, C.T. Lewis, and I.A. Parkin. 2004. Maximizing the efficacy of SAGE analysis identifies novel transcripts in Arabidopsis. *Plant Physiol* **136:** 3223-3233.

Rogic, S., A.K. Mackworth, and F.B. Ouellette. 2001. Evaluation of gene-finding programs on mammalian sequences. *Genome Res* **11:** 817-832.

Rose, T.L., L. Bonneau, C. Der, D. Marty-Mazars, and F. Marty. 2005. Starvation-induced expression of autophagy-related genes in Arabidopsis. *Biol Cell, Advance online publication Apr 6.*

Ruijter, J.M., A.H. Van Kampen, and F. Baas. 2002. Statistical evaluation of SAGE libraries: consequences for experimental design. *Physiol Genomics* **11:** 37-44.

Rusconi, J.C., R. Hays, and R.L. Cagan. 2000. Programmed cell death and patterning in Drosophila. *Cell Death Differ* **7:** 1063-1070.

Ryo, A., N. Kondoh, T. Wakatsuki, A. Hada, N. Yamamoto, and M. Yamamoto. 2000. A modified serial analysis of gene expression that generates longer sequence tags by nonpalindromic cohesive linker ligation. *Anal Biochem* **277:** 160-162.

Saha, S., A.B. Sparks, C. Rago, V. Akmaev, C.J. Wang, B. Vogelstein, K.W. Kinzler, and V.E. Velculescu. 2002. Using the transcriptome to annotate the genome. *Nat Biotechnol* **20:** 508-512.

Samari, H.R. and P.O. Seglen. 1998. Inhibition of hepatocytic autophagy by adenosine, aminoimidazole-4-carboxamide riboside, and N6-mercaptopurine riboside. Evidence for involvement of amp-activated protein kinase. *J Biol Chem* **273:** 23758-23763.

Schena, M., D. Shalon, R.W. Davis, and P.O. Brown. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270:** 467-470.

Schuler, G.D. 1997. Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J Mol Med* **75:** 694-698.

Schwarze, P.E. and P.O. Seglen. 1985. Reduced autophagic activity, improved protein balance and enhanced in vitro survival of hepatocytes isolated from carcinogen-treated rats. *Exp Cell Res* **157:** 15-28.

Schweichel, J.U. and H.J. Merker. 1973. The morphology of various types of cell death in prenatal tissues. *Teratology* **7:** 253-266.

Scott, R.C., O. Schuldiner, and T.P. Neufeld. 2004. Role and regulation of starvation-induced autophagy in the Drosophila fat body. *Dev Cell* **7:** 167-178.

Shimizu, S., T. Kanaseki, N. Mizushima, T. Mizuta, S. Arakawa-Kobayashi, C.B. Thompson, and Y. Tsujimoto. 2004. Role of Bcl-2 family proteins in a non-apoptotic programmed cell death dependent on autophagy genes. *Nat Cell Biol* **6:** 1221-1228.

Shintani, T. and D.J. Klionsky. 2004. Autophagy in health and disease: a double-edged sword. *Science* **306:** 990-995.

Shoemaker, D.D., E.E. Schadt, C.D. Armour, Y.D. He, P. Garrett-Engele, P.D. McDonagh, P.M. Loerch, A. Leonardson, P.Y. Lum, G. Cavet et al. 2001. Experimental annotation of the human genome using microarray technology. *Nature* **409:** 922-927.

Silva, A.P., J.E. De Souza, P.A. Galante, G.J. Riggins, S.J. De Souza, and A.A. Camargo. 2004. The impact of SNPs on the interpretation of SAGE and MPSS experimental data. *Nucleic Acids Res* **32:** 6104-6110.

Siu, I.M., A. Lal, and G.J. Riggins. 2001. A database for regional gene expression in the human brain. *Brain Res Gene Expr Patterns* **1:** 33-38.

Sonnhammer, E.L. and R. Durbin. 1994. A workbench for large-scale sequence homology analysis. *Comput Appl Biosci* **10:** 301-307.

Sorlie, T., R. Tibshirani, J. Parker, T. Hastie, J.S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler et al. 2003. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A* **100:** 8418-8423.

Soung, Y.H., J.W. Lee, S.Y. Kim, J. Jang, Y.G. Park, W.S. Park, S.W. Nam, J.Y. Lee, N.J. Yoo, and S.H. Lee. 2005. CASPASE-8 gene is inactivated by somatic mutations in gastric carcinomas. *Cancer Res* **65:** 815-821.

Spinella, D.G., A.K. Bernardino, A.C. Redding, P. Koutz, Y. Wei, E.K. Pratt, K.K. Myers, G. Chappell, S. Gerken, and S.J. McConnell. 1999. Tandem arrayed ligation of expressed sequence tags (TALEST): a new method for generating global gene expression profiles. *Nucleic Acids Res* **27:** e22.

Srinivasula, S.M., P. Datta, M. Kobayashi, J.W. Wu, M. Fujioka, R. Hegde, Z. Zhang, R. Mukattash, T. Fernandes-Alnemri, Y. Shi et al. 2002. sickle, a novel Drosophila death gene in the reaper/hid/grim region, encodes an IAP-inhibitory protein. *Curr Biol* **12:** 125-130.

Stapleton, M., G. Liao, P. Brokstein, L. Hong, P. Carninci, T. Shiraki, Y. Hayashizaki, M. Champe, J. Pacleb, K. Wan et al. 2002. The Drosophila gene collection: identification of putative full-length cDNAs for 70% of D. melanogaster genes. *Genome Res* **12:** 1294-1300.

Steen, B.R., T. Lian, S. Zuyderduyn, W.K. MacDonald, M. Marra, S.J. Jones, and J.W. Kronstad. 2002. Temperature-Regulated Transcription in the Pathogenic Fungus Cryptococcus neoformans. *Genome Res* **12:** 1386-1400.

Stein, L., P. Sternberg, R. Durbin, J. Thierry-Mieg, and J. Spieth. 2001. WormBase: network access to the genome and biology of Caenorhabditis elegans. *Nucleic Acids Res* **29:** 82-86.

Stein, L.D. and J. Thierry-Mieg. 1998. Scriptable access to the Caenorhabditis elegans genome sequence and other ACEDB databases. *Genome Res* **8:** 1308-1315.

Stollberg, J., J. Urschitz, Z. Urban, and C.D. Boyd. 2000. A quantitative evaluation of SAGE. *Genome Res* **10:** 1241-1248.

Storey, J.D. 2002. A direct approach to false discovery rates. *J.R. Statist. Soc. B* **64:** 479-498.

Strausberg, R.L., E.A. Feingold, R.D. Klausner, and F.S. Collins. 1999. The mammalian gene collection. *Science* **286:** 455-457.

Sun, M., G. Zhou, S. Lee, J. Chen, R.Z. Shi, and S.M. Wang. 2004. SAGE is far more sensitive than EST for detecting low-abundance transcripts. *BMC Genomics* **5:** 1.

Takita, J., H.W. Yang, F. Bessho, R. Hanada, K. Yamamoto, V. Kidd, T. Teitz, T. Wei, and Y. Hayashi. 2000. Absent or reduced expression of the caspase 8 gene occurs frequently in neuroblastoma, but not commonly in Ewing sarcoma or rhabdomyosarcoma. *Med Pediatr Oncol* **35:** 541-543.

Tal-Or, P., A. Di-Segni, Z. Lupowitz, and R. Pinkas-Kramarski. 2003. Neuregulin promotes autophagic cell death of prostate cancer cells. *Prostate* **55:** 147-157.

Tanida, I., T. Ueno, and E. Kominami. 2004. LC3 conjugation system in mammalian autophagy. *Int J Biochem Cell Biol* **36:** 2503-2518.

Teitz, T., T. Wei, M.B. Valentine, E.F. Vanin, J. Grenet, V.A. Valentine, F.G. Behm, A.T. Look, J.M. Lahti, and V.J. Kidd. 2000. Caspase 8 is deleted or silenced preferentially in childhood neuroblastomas with amplification of MYCN. *Nat Med* **6:** 529-535.

Tenev, T., A. Zachariou, R. Wilson, A. Paul, and P. Meier. 2002. Jafrac2 is an IAP antagonist that promotes cell death by liberating Dronc from DIAP1. *Embo J* **21:** 5118-5129.

Thumm, M., R. Egner, B. Koch, M. Schlumpberger, M. Straub, M. Veenhuis, and D.H. Wolf. 1994. Isolation of autophagocytosis mutants of Saccharomyces cerevisiae. *FEBS Lett* **349:** 275-280.

Thumm, M. and T. Kadowaki. 2001. The loss of Drosophila APG4/AUT2 function modifies the phenotypes of cut and Notch signaling pathway mutants. *Mol Genet Genomics* **266:** 657-663.

Toth, S., K. Nagy, Z. Palfia, and G. Rez. 2002. Cellular autophagic capacity changes during azaserine-induced tumour progression in the rat pancreas. Up-regulation in all premalignant stages and down-regulation with loss of cycloheximide sensitivity of segregation along with malignant transformation. *Cell Tissue Res* **309:** 409-416.

Tsukada, M. and Y. Ohsumi. 1993. Isolation and characterization of autophagy-defective mutants of Saccharomyces cerevisiae. *FEBS Lett* **333:** 169-174.

Tupy, J.L., A.M. Bailey, G. Dailey, M. Evans-Holm, C.W. Siebel, S. Misra, S.E. Celniker, and G.M. Rubin. 2005. Identification of putative noncoding polyadenylated transcripts in Drosophila melanogaster. *Proc Natl Acad Sci U S A* **102:** 5495-5500.

Unneberg, P., A. Wennborg, and M. Larsson. 2003. Transcript identification by analysis of short sequence tags--influence of tag length, restriction site and transcript database. *Nucleic Acids Res* **31:** 2217-2226.

Ureta-Vidal, A., L. Ettwiller, and E. Birney. 2003. Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat Rev Genet* **4:** 251-262.

van 't Veer, L.J., H. Dai, M.J. van de Vijver, Y.D. He, A.A. Hart, M. Mao, H.L. Peterse, K. van der Kooy, M.J. Marton, A.T. Witteveen et al. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415:** 530-536.

Varghese, J., H. Sade, P. Vandenabeele, and A. Sarin. 2002. Head involution defective (Hid)-triggered apoptosis requires caspase-8 but not FADD (Fas-associated death domain) and is regulated by Erk in mammalian cells. *J Biol Chem* **277:** 35097-35104.

Velculescu, V.E., S.L. Madden, L. Zhang, A.E. Lash, J. Yu, C. Rago, A. Lal, C.J. Wang, G.A. Beaudry, K.M. Ciriello et al. 1999. Analysis of human transcriptomes. *Nat Genet* **23:** 387-388.

Velculescu, V.E., L. Zhang, B. Vogelstein, and K.W. Kinzler. 1995. Serial analysis of gene expression. *Science* **270:** 484-487.

Velculescu, V.E., L. Zhang, W. Zhou, J. Vogelstein, M.A. Basrai, D.E. Bassett, Jr., P. Hieter, B. Vogelstein, and K.W. Kinzler. 1997. Characterization of the yeast transcriptome. *Cell* **88:** 243-251.

Venter, J.C., M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, C.A. Evans, R.A. Holt et al. 2001. The sequence of the human genome. *Science* **291:** 1304-1351.

Vilain, C., F. Libert, D. Venet, S. Costagliola, and G. Vassart. 2003. Small amplified RNA-SAGE: an alternative approach to study transcriptome from limiting amount of mRNA. *Nucleic Acids Res* **31:** e24.

Virlon, B., L. Cheval, J.M. Buhler, E. Billon, A. Doucet, and J.M. Elalouf. 1999. Serial microanalysis of renal transcriptomes. *Proc Natl Acad Sci U S A* **96:** 15286-15291.

Wahl, M.B., U. Heinzmann, and K. Imai. 2005. LongSAGE analysis revealed the presence of a large number of novel antisense genes in the mouse genome. *Bioinformatics* **21:** 1389-1392.

Warnecke, P.M., C. Stirzaker, J.R. Melki, D.S. Millar, C.L. Paul, and S.J. Clark. 1997. Detection and measurement of PCR bias in quantitative methylation analysis of bisulphite-treated DNA. *Nucleic Acids Res* **25:** 4422-4426.

Waterston, R.H., K. Lindblad-Toh, E. Birney, J. Rogers, J.F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520-562.

Wei, C.L., P. Ng, K.P. Chiu, C.H. Wong, C.C. Ang, L. Lipovich, E.T. Liu, and Y. Ruan. 2004. 5' Long serial analysis of gene expression (LongSAGE) and 3' LongSAGE for transcriptome characterization and genome annotation. *Proc Natl Acad Sci U S A* **101:** 11701-11706.

Xie, H., W.Y. Zhu, A. Wasserman, V. Grebinskiy, A. Olson, and L. Mintz. 2002. Computational analysis of alternative splicing using EST tissue information. *Genomics* **80:** 326-330.

Xu, Q., B. Modrek, and C. Lee. 2002. Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res* **30:** 3754-3766.

Yamamoto, A., Y. Tagawa, T. Yoshimori, Y. Moriyama, R. Masaki, and Y. Tashiro. 1998. Bafilomycin A1 prevents maturation of autophagic vacuoles by inhibiting fusion between autophagosomes and lysosomes in rat hepatoma cell line, H-4-II-E cells. *Cell Struct Funct* **23:** 33-42.

Yamamoto, M., T. Wakatsuki, A. Hada, and A. Ryo. 2001. Use of serial analysis of gene expression (SAGE) technology. *J Immunol Methods* **250:** 45-66.

Yan, J. and T.G. Marr. 2005. Computational analysis of 3'-ends of ESTs shows four classes of alternative polyadenylation in human, mouse, and rat. *Genome Res* **15:** 369-375.

Yelin, R., D. Dahary, R. Sorek, E.Y. Levanon, O. Goldstein, A. Shoshan, A. Diber, S. Biton, Y. Tamir, R. Khosravi et al. 2003. Widespread occurrence of antisense transcription in the human genome. *Nat Biotechnol* **21:** 379-386.

Yu, L., A. Alva, H. Su, P. Dutt, E. Freundt, S. Welsh, E.H. Baehrecke, and M.J. Lenardo. 2004. Regulation of an ATG7-beclin 1 program of autophagic cell death by caspase-8. *Science* **304:** 1500-1502.

Yue, Z., S. Jin, C. Yang, A.J. Levine, and N. Heintz. 2003. Beclin 1, an autophagy gene essential for early embryonic development, is a haploinsufficient tumor suppressor. *Proc Natl Acad Sci U S A* **100:** 15077-15082.

Zeeberg, B.R., W. Feng, G. Wang, M.D. Wang, A.T. Fojo, M. Sunshine, S. Narasimhan, D.W. Kane, W.C. Reinhold, S. Lababidi et al. 2003. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol* **4:** R28.

Zhang, J., R.P. Finney, R.J. Clifford, L.K. Derr, and K.H. Buetow. 2005. Detecting false expression signals in high-density oligonucleotide arrays by an in silico approach. *Genomics* **85:** 297-308.

Zhang, J.H., Y. Zhang, and B. Herman. 2003. Caspases, apoptosis and aging. *Ageing Res Rev* **2:** 357-366.

Zhao, X., R.E. Ayer, S.L. Davis, S.J. Ames, B. Florence, C. Torchinsky, J.S. Liou, L. Shen, and R.A. Spanjaard. 2005. Apoptosis factor EI24/PIG8 is a novel endoplasmic reticulum-localized Bcl-2-binding protein which is associated with suppression of breast cancer invasiveness. *Cancer Res* **65:** 2125-2129.

Zornig, M., A. Hueber, W. Baum, and G. Evan. 2001. Apoptosis regulators and their role in tumorigenesis. *Biochim Biophys Acta* **1551:** F1-37.