An Ontological Approach to Analysing Developmental Processes in the Mouse Based on Co-Expression

Mei Sze Lam

Master of Science

School of Informatics

University of Edinburgh

2007

Abstract

Identifying the functions of uncharacterised genes is one of the challenges in bioinformatics. Numerous solutions have been proposed for the problem.

This work will outline one possible technique for identifying genes responsible for a developmental process in the mouse embryo. Our approach consists of analysing the co-expression of genes in different tissues which where we know a particular process is happening. We derive our data from GXD and annotate them with Biological Process terms using the Gene Ontology. Finally we analyse the significance of each term in order to create a functional view of our co-expressed genes.

We focus on a process that is little understood in the current literature, the mesenchymal to epithelial transformation of cells. Yet our approach is theoretically feasible for any other biological process that occurs in several tissues of an arbitrary organism during its developmental or gestation period.

Although our methods were proved unable to isolate the set of genes that are responsible for the mesenchyme-epithelial process, we obtained a number of interesting results. The principle validity of our approach and the underlying hypothesis remains questionable. However, the project succeeded in finding numerous candidate genes for our process of interest, which could be worth subjecting to more rigorous biological experimentation and analysis.

Acknowledgements

Aforemost, I would like to thank the two supervisors of my project, Dr. Stuart Aitken and Dr. Jonathan Bard, who have been so patient with me and provided me with assistance and reassurance over the past months.

Finally, my thanks go to my friends and family whom I have neglected so much recently and who have beared with me all this while.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Mei Sze Lam*)

Dedicated to snowdrops.

Table of Contents

Chapter 1

Introduction

Research on the mammalian genome has progressed rapidly over the past decade in conjunction with advancements in computing technology and the internet. Such a surge in our understanding of the the genome would have been unthinkable even a decade ago, before the advent of computer-aided and partially/fully-automated techniques that perform large-scale analyses of vast amounts of biological data.

And yet, despite the wealth of gene expression data that we have acquired in the post-genome age, biological research has only managed to understand the underlying processes and functions from a small portion of the gene and gene products we have sequenced. Consequently, we are still unaware of the functions of most genes, indeed even the functions themselves in even the most frequently studied mammalian genome - the laboratory mouse. Needless to say, our basic understanding of the human genome is equally or even more limited.

1.1 Objective

In this study, we will try to contribute yet another little part to the puzzle by employing some fairly straight-forward computational techniques to narrow down on the possible candidate genes for a functional process.

It is our declared objective to identify the set of genes that are responsible for regulating the transition from mesenchymal to epithelial cells, which happens at different

stages in various tissues of the gestating mouse embryo. We shall attempt to collate the genes that are expressed at each of these temporal and spatial points to see whether there is a common set that could be responsible for the process.

We hypothesize that there is a common set of genes for the mesenchyme-epithelial transformation that are differentially expressed in various tissues of the mouse embryo. We also postulate that it should be possible to isolate the set of genes responsible for this transformation by checking for their expression states in the relevant tissues at the appropriate embryonic stages of the mouse.

If a gene is differentially expressed across several developing tissues or organs, we can infer that the gene fulfils some developmental-related purpose at each site. Consequently, if we equip our technique with enough knowledge of when and where this takes place, given sufficient expression data, it should be possible to identify the set of mesenchymal -epithelial genes.

We implement a software package that allows for the simple definition of sets of tissues within the mouse embryo, the performance of all necessary computations on these sets, and the derivation of certain expression categories of genes. Lastly, we utilise an annotation analysis tool, Fatigo+ to return us the significant terms across these sets and review our results.

1.2 Mesenchyme and Epithelium

The process we are examining in our project is the change in cell differentiation which is transforming mesenchymal to epithelial cells. While the reverse project (epithelial to mesenchymal transition) has been well studied (e.g. [1, 2, 3]), the mesenchyme-epithelium transition has attracted less interest so far.

Mesenchymal cells typically pack in 3-dimensional clusters, while epithelial cells form 2-dimensional layers - often found in tubules. The mesenchyme-epithelium transition is therefore a crucial step in the development of many organs and associated systems. The process is, however, not restricted to the embryo, but can also occur in adult organisms [4, 5], where it (amongst others) is involved in wound healing and fibrosis. Disruptions in either of the two transformation processes is also commonly linked to cancer.

A good example is formation of the kidney (or metanephros) [6, 7, 8, 9]. Figure 1.1.(a) shows a rough outline of the development of the mouse kidney in the embryo. Between embryonic days 12.5 and 13.5 we can observe that a condensation of mesenchymal cells has transformed into a condensation of epithelial cells which will form the later nephron tubules. In a an confocal microscope picture taken at E13 (Fig. 1.1.(b), we can find both kinds of cells, mesenchymal and epithelial, as the transition is currently in progress at that time.

For the purposes of this project, the mesenchyme to epithelium transition is a adequate example process to test our general methodology on, since it occurs across different structures at different stages of embryonic development. We will consequently be able to search for genes which all these structures have in common to obtain candidate genes potentially responsible for this very process.

1.3 Overview

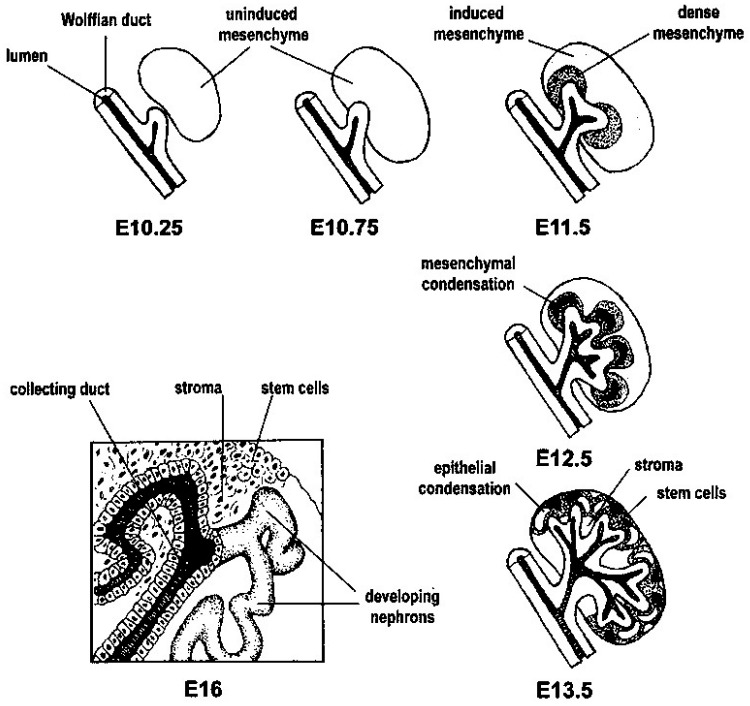
The remainder of this report is structured as follows:

In the next chapter, we will discuss some of the background behind what we are doing. We will first briefly review existing literature on co-expression of genes and how it is linked to gene function prediction and then proceed to a detailed presentation of the Gene Ontology, which is the basic resource that most gene annotation tools work with. These gene annotation tools are the topic of the last part of Chapter 2. We discuss several possible choice and narrow these down to a tool called *Fatigo+*, which will be used later on for the analysis of gene function.

Having introduced the basics, we continue with an extensive explanation of the methodology employed in this study. We clearly define the requirements of the system to be implemented , together with computations that will be necessary and the source of data used for discovering genes expressed in tissues, in the first part of the chapter. Thereafter, we present our solution to the problems outlined, at tool termed *Mouse Genome Intersector*. We describe the program's core components and functionality, before we conclude the chapter with a concise summary.

The third chapter focuses on the experiments we carried out and the results we obtained. We start by explaining which tissues in the mouse embryo have been studied

(a)



(b)

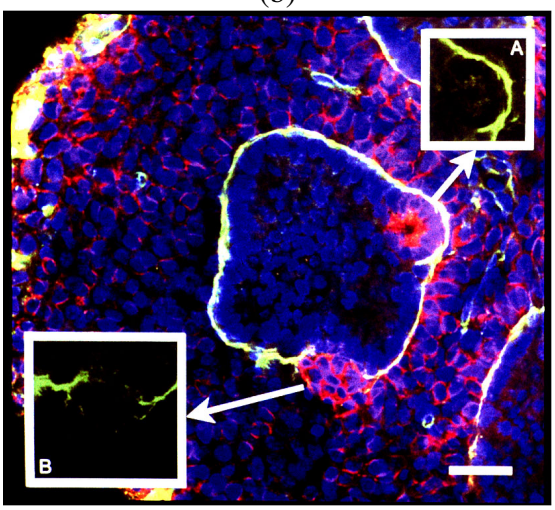


Figure 1.1: (a) Formation of the kidney across different stages in the development of the mouse embryo. Note how a condensation of mesenchymal cells transforms into epithelial cells between E12.5 and E13.5, (b) Confocal microscope picture of a developing mouse kidney (E13). Cell nuclei have been marked blue. Note the large tube with a green/yellow (laminin stain) membrane and the the red (an adhesion molecule stain) membranes of mesenchymal cells. [6]

and explain their relation to the mesenchyme-epithelium process. Afterwards, we describe the different experiments, i.e. the different kinds of intersections that have been carried out on the datasets and the results we obtained. Realising that the results do not satisfy our expectations, we then identify potential reason for this failure and further investigate the data with some promising results. The last part of the chapter comprises a biological interpretation of the results contributed by an expert in this field, Dr. Jonathan Bard.

We conclude the thesis in the fourth chapter with a more higher-level discussion of our finding and the implications of our results. We investigate the details of all significant enrichments of GO terms and discuss their meaning. We critically assess problems we encountered and outline how they connect to opportunities for future work in this field.

After the discussion, there are two appendices: In Appendix A we describe the implemented program in a little more detail by providing UML class diagrams of the code packages, listing database queries and providing a range of screenshots from the program in action, which may serve as a very basic tutorial to the tool itself. Appendix B is dedicated to the genes found in the experiments. Full lists of all genes expressed in the different datasets will be given to provide the possibility for other researchers to further study the mesenchyme-epithelium transition phenomenon on their basis.

Chapter 2

Background

In this chapter we shall give an overview of the literature relevant to our project. We start by reviewing previous approaches to analysing gene function from co-expression profiles, then go on to talk about the structure of the Gene Ontology and present some of its inherent weaknesses. The last section comprises a brief review of gene annotation analysis tools and, in particular, focus on one, *Fatigo+*, which we will employ later in this study.

2.1 Co-Expression of Genes

It is widely believed that genes sharing the same expression patterns are also more likely to be responsible for the same biological process. Many gene prediction studies are based on microarray analyses of gene co-expression across different tissues. Such types of co-expression can also be referred to as transcriptional co-expression. The general strategy behind most co-expression studies is to first form groups of genes that exhibit the same expression patterns. An unidentified gene within the the group is then assigned the annotations common to the known genes in the group.

A genome wide study on the mouse (*Mus Musculus*) using 40,000 mRNAs extracted from 55 tissues by Zhang, Morris, Chang established that transcriptional co-expression is a powerful tool for predicting gene function [10]. Not surprisingly, genes which are known to perform tissue specific functions are most prevalent in those tissues where those biological processes are primarily taking place. For example, 'synaptic

transmission' were found to have the highest expression levels within the neuronal tissues, while 'learning and memory' related genes were expressed strongly in the cortex and striatum. These findings are examples of tissue-specific co-expression. Different tissues or organs that require the same biological process can also share common expression patterns. For example, the lung, bladder and skin, three very different tissue structures that are constantly being exposed to the elements all exhibited the same expression pattern of 'immune-related' of genes.

The study also proved that the correlation of gene expression pattern to gene function can be observed independent of their source tissues. Genes are sorted into clusters based on their expression patterns across tissues without referring to any a priori knowledge of their origin or functions. They then found that each cluster tended be associated strongly with a handful of specific annotation categories. These clusters of genes were then used to train a SVM classifier to predict gene function based on its expression patterns. The classifier was capable of classifying more than half of 1092 unannotated genes to some known functional category, the rest of the genes did not have any similarity to any class of pattern. Of the ones that were assigned, manual reference to literature demonstrated that the classifications were likely to be correct.

Similarly, Lee et al. [11] investigated co-expression in the human genome by building a network of genes connected by 'co-expression links' that construct pairs of genes most similar to each other in terms of their expression patterns. Their research demonstrated that likewise in the mouse genome, patterns of correlated gene expression exist across multiple microarrays for humans, and these patterns correspond strongly to distinct functional categories. Clustered analysis of the gene network reveal functionally coherent groups of genes.

Nevertheless, the general stance of most researchers towards co-expression is that it provides only a weak means for the prediction of gene function (e.g., [12], where co-expression alone is said to provide insufficient basis for gene function prediction). Only comparably little work has been done to actually investigate this claim, the work by Allocco et al. [13] being one exception. They assessed the correlation between common mRNA expression patterns and gene function by comparing genome wide binding analysis and mRNA expression data with their functional annotations in the Gene Ontology (see next section). They found that the chances of genes actually sharing a common transcription factor binding even though co-expressing are relatively

low and one could therefore not necessarily conclude from co-expression to a common regulatory mechanism. Nevertheless, they also concluded that genes that actually share transcription factor are more likely to be functionally related (as defined by their annotations).

In a study on the essentiality (necessity for survival) of genes in yeast, Carlson et al. [14] formed weighted gene co-expression networks using data from a range of microarray datasets. Their results proved that the connectivity of genes within these networks was highly correlated with their essentiality and gene sequence preservation. In other words, co-expressing genes were found to be involved in the same functional process. The application of the same technique was said to be promising for a more detailed prediction of gene involvement in functional compartments.

Similarly, Stuart et al. [15] examined vast amounts of microarray data from different species (humans, flies, worms and yeast) with respect to co-expression and then investigated which of these had been conserved over evolution. They reason that a co-expressing set of genes is actually instantiating a common evolutionary selective advantage and consequently is likely to have a shared function. Further investigations on a number of examples within the found data confirmed these conclusions.

It should also be mentioned that there are numerous techniques to identify clusters of co-expressed genes, many of which are much more sophisticated than just picking genes on a boolean basis (i.e. 'on' or 'off') by hand. Existing and prospective future techniques in development include Self-Organizing Maps, various methods of hierarchical and qualitative clustering, mixture models and bi-clustering [16]. More importantly, the cited source also names a number of methods to assess the found cluster quality, one of which - what they call 'enrichment of functional categories' - can be used to link found clusters to functional processes.

The theory of co-expression has been tackled from many different angles and utilised in many approaches for numerous studies. Perhaps one of the studies most similar in approach to our project is described in a very recently published paper [17]. The authors used supervised learning to analyse genes expressed over a time period using multiple microarrays. Their aim was to predict new functions not yet documented in the Gene Ontology for uncharacterised genes using their model. First, the expression level over times of both characterised and uncharacterised genes were recorded as a temporal pattern. Next the authors annotated the characterised genes to all their

biological processes using the Gene Ontology. Using both these aspects of data to train their model, they hypothesized new biological processes for the unknown genes. The study discovered that a considerable number of their proposed biological processes were validated by a manual curation of the available literature and existing homology information. Finally, it should be said that this is the only study that is conceptually the most similar to our approach. It includes both the temporal and co-expression aspects for analysing genes, and it utilises the Biological Process terms in Gene Ontology as the basis of their predictions. The authors of this paper state that their method has been unprecedented in the field.

Countless studies have been performed on individual examples or small sets of co-expressing genes, clearly underlining their functional similarities. In summary, one can conclude, although linking co-expression with similar gene functions has been confronted with considerable doubts, previous work has often contradicted this disbelief and, in fact, provided promising results.

2.2 The Gene Ontology

The Gene Ontology (GO) is a structured, precisely defined and controlled vocabulary used to describe the gene roles in all organisms. Its aim is to standardise the descriptions of gene products in different databases, thus unifying all available information into a common biological language that can be shared by biologists world wide. GO is an ongoing project launched and maintained by the GO Consortium, a coalition of biological databases for various organisms worldwide. The Mouse Genome Informatics Database, which is the primary data resource for this project, is also a major participant and primary contributor of *Mus Musculus* terms to the Gene Ontology project.

2.2.1 Structure of the Ontology

GO dictates that all functions of genes and gene products are a part of one or more of three basic attributes Biological Process (BP), Molecular Function (MF) and Cellular Component (CC). Each attribute is an independent ontology on its own. One can browse within any one of these three ontologies for the related terms to a gene depending on the aspect of information that is required. Since all genes have the three

aforementioned attributes, it can be annotated within one or more of the separate ontologies, and may have multiple GO terms within each ontology to accommodate the fact that an individual gene may serve more than one function or participate in several processes.

The vocabulary of GO consists of entities known as 'GO terms'. These are biological terms that represent the function of a gene within an organism. All GO terms are represented by a brief textual description in scientific natural language and also a GO ID which has a numeric string representing it uniquely in the ontology. There are only two types of relations in GO, 'is-a' and 'part-of'. Each GO term may be related to child GO terms by one and only one of these relations. A GO term with an 'is-a' relation is understood to be a specialised instance of its parent in its own right. Otherwise 'part-of' denotes that the term is a sub-component, and does not constitute the entirety of the parent term on its own [18]. All GO term other than the three top nodes (BP, MF, CC) must be related to a more general term. Thus each GO term can be represented as a node in a multi-level parent-child node network. The entire ontology is commonly visualised as a hierarchical tree of GO terms that branch out into more GO terms. However, it is not technically a tree because individual terms may occur as children of more than one parent node. GO's architecture is accurately described as a directed acyclic graph because the nodes are always pointed from ancestor to descendants, no child is allowed to be a parent of any of its ancestors.

Our knowledge of the genome is constantly increasing and evolving in this era of genome sequencing. Bearing this in mind, the hierarchical structure of GO facilitates the organisation of biological knowledge at varying stages of completion [19]. Genes whose functions are poorly known can still be annotated in the ontology, albeit at a very high level, represented by only one or a handful of the GO terms it possesses, whereas genes whose functions are well understood can be represented by a richer supply of GO terms with progressively detailed associations within the lower levels of the tree. More importantly for the purposes of this project, the hierarchical and tree-like representation of GO allow users to specify the depth of queries, so that GO is able to provide a general overview of the gene, or provide progressively in-depth terms of the gene by traversing down the branches.

2.2.2 Potential Problems

Although GO is currently the *de facto* standard for gene annotation in the biological community, there are a number of issues which have been criticized about it. We shall now briefly recapitulate a few of these drawbacks that might affect our project.

2.2.2.1 Ontological and Structural Weaknesses

The structural design of GO has been a topic of debate amongst ontology experts and biologists alike. The GO Consortium's philosophy in setting up GO was so that it could provide a useful framework for the organisation of biological data. The GO Consortium's main priority in the design phase of GO was to provide a framework that would allow for the speedy population of the ontology. In comparison, not much focus was given to developing a framework capable of providing robust support to software applications [18]. Nevertheless, the GO is an important source of knowledge and reference for many biological analysis and predictive tools [20]. While the specialisations and functions of each tool differ, they are collectively termed as ontology-driven functional analysis programs. Such tools are knowledge-based software that exploit a knowledge source (the ontology) to infer its own assertions and answer queries. The accuracy of the software's inferences depend on the consistency of the ontology [18]. Consequently, the potential reasoning capacity of such tools could ultimately only be limited by representational adequacy and expressiveness of the ontology. The remainder of this section shall examine these factors in the GO in relation to our project, and also outline some anticipatory drawbacks of relying on GO as a primary reference for data evaluation.

Ontology purists argue that the GO does not qualify as an ontology in the strict sense [18]. The origin of ontology in Greek means 'the study of being or existence' [21]. A modern ontology seeks to represent all knowledge about the entities in a given domain using a set of terms (classes, objects). Relations between entities are described using a set of defined associations. There are formal axioms in place to dictate what constitutes each different term [22]. From a practical viewpoint, such a conceptualisation of the biological domain is difficult even now, because: (1) New discoveries are constantly changing our understanding of gene function and interaction, (2) Given the current situation, biologists cannot agree on a set of terms that can completely describe

everything to everyone's satisfaction. Instead of waiting for these problems to resolve themselves in the unforeseeable future, GO solves it now by employing three broad terms (molecular function, biological process, cellular component) and two relations (is-a, part-of). As a result, the Gene Ontology was set up and in use much earlier than it could have been otherwise. The two drawbacks of this approach are: (1) From a biologist's viewpoint, the complex nature of genes and subtler relations between different gene functions are subsequently lost in the generalisation. (2) From an information scientist's viewpoint, the lack of rigid logical formalisation provides little support for reasoning software.

2.2.2.2 Imperfect Knowledge Base

In the project at hand, we are interested in getting all BP annotations that describe the mesenchyme-epithelial cell transformation process. Since this transformation is not yet fully understood or thoroughly studied by biologists, we know beforehand that GO does not have any explicit GO term dedicated exclusively to the process. However there are many GO terms that describe the separate development of both types of cells, such as *Mesenchymal Cell Development* [GO:0048762], and *Epithelial Cell Development* [GO:0016021]. There are also other GO terms where the concept of either one type of cell development is an underlying theory, even though the terms themselves do not contain the keywords 'mesenchymal' or 'epithelial'. Such an example is the GO term *Tube Development* [GO:0002064]. A biology expert would know that a tube is constructed of epithelial cells. Even a layman can infer this by referring to the GO definition of Tube Development and discovering the keyword 'epithelial'. GO places Tube Development as a child node of *Anatomical Structure Development* [GO:0035295]. From a purely logical stance, one could argue that since a tube is constructed of epithelial cells, it would be informative if GO could indicate this by giving the former some sort of relation to *Epithelial Cell Development* [GO:0002066]. The two terms appear in separate branches of the BP ontology with no direct links to indicate that some form of epithelial cell building is actually taking place in the development of tubes. (Fig 1). However it would not be logical to classify the node for tube development as a child of epithelial cell development. The two biological processes, although related, are distinctly different. Tube development is neither a part of epithelial cell development nor is it a specialised process of epithelial cell development. On the other hand, epithelial cells by no means exist only for the formation of tubes, but form many

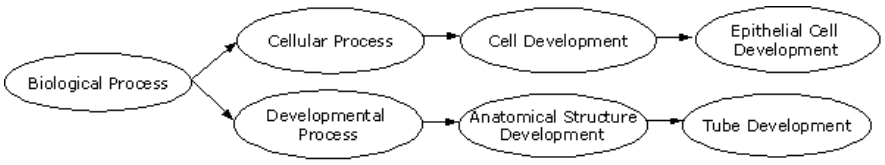


Figure 2.1: An extract of the Biological Process branch of the gene ontology. Note how *Tube Development* and *Epithelial Cell Development* appear in separate branches without any direct links, although one could argue, that they are clearly related, since tubes in organisms are formed when mesenchymal cells transform into epithelial cells.

other structures in the body, such as the skin for example. Therein lies the impossibility of linking the two terms together in GO using 'is-a' or 'part-of'. Indeed many studies have been conducted to suggest the revision of old terms and introduction of new relations in order to improve the expressibility of GO [23]. Although it is not within the scope of this project to suggest or explore new types of relationships to GO, we have to bear in mind that this situation is hypothetically applicable to other relevant terms in the ontology.

We have already demonstrated how 'is-a' and 'part-of' is insufficient for our purposes of discovering implicitly related terms. It is hoped that the impact this problem could have on our evaluation can be allayed by manually sifting through the definitions of interesting annotations. The disadvantages of such an approach are two-fold. Firstly, it partially defeats the purpose of ontology-driven functional analysis tools, which are meant to be an automated process. Secondly, the objectivity of the evaluation is called into question since deciding what constitutes as 'interesting' annotations is an arbitrary process prone to human bias.

2.2.2.3 Redundancy and Similarity

Another major factor that could ultimately affect the accurate annotation of genes is the fact that GO has no protocol for checking how many terms denote the same concept. The rationale behind the sub-classifications of existing terms have also not been documented sufficiently [18]. One implication of the lack of such of documentation this is of the lack of such documentation is that a contributor of a new term could find it difficult to ascertain whether their concept already exists in the tree under a differently named term. This is especially the case if it was a highly specialised concept and hap-

pened to be the deepest node in another branch of biology completely different from the contributor's area of expertise. Only a curator who happened to have an in depth knowledge of both areas would be able to detect the duplication of concepts.

This difficulty is further compounded by the fact that many GO terms are composites of other terms [22]. Using a program to measure the similarity of terms by counting the number of word edits required to transform one term into another term, Ogren observed that 68.8% of GO terms added within a one year period contain an existing term as a sub-string. This in itself would not prove a problem if terms sharing the same sub-strings still denoted different concepts. However by counting the number of times each GO term was cited in a publication and comparing this with the compositionality composition of terms, the study also discovered that the more composite a term, the less likely the annotation was to be cited in the corpus. In fact almost half of all terms in the GO have never been cited in a manual or automated context. One can infer from his study that there is a high degree of redundancy in the GO terminology [22].

To illustrate a potential problem this could cause in our project, imagine if *Gene A* has ten annotations containing the keyword 'mesenchymal'. Meanwhile *Gene B* has only one such annotation. How accurate is it to conclude that *Gene A* is more likely to be involved in the mesenchymal-epithelial process than *Gene B*? In such a situation, it would take a biological expert to determine whether the same concept repeats itself in the ten annotations for *Gene A*.

2.2.2.4 Levels of Abstraction

Earlier on, we discussed how different levels of GO terms within the hierarchy allow us to view functions at various degrees of specialisation. While it is intuitive to assume that the depth of a node would indicate the specificity of a term, in [24], the authors discovered that depth is not a good semantic gauge of function specialisation. For example, *High-Affinity Tryptophan Transporter* [GO:0005300] at Level 14 and *Anticoagulant* [GO:0008435] at Level 3 are both instances of equally specialised molecules. In short, just because some branches traverse longer distances than others, it does not necessarily connote that many in-depth studies have been performed in the related area of biology and therefore the discovery of many highly specialisation specialised terms. The implications of accepting this conclusion is effectively altering our strategy for

data evaluation so that we do not depend on GO term levels to measure good results. Yet eliminating the consideration of level depth entirely from our study is counter intuitive we expect that the lower a node, the more specific and detailed information it gives us about a gene.

2.2.2.5 Curation and Consistency

It has been proven that the inherent design of GO has considerably impeded the annotation and curation capabilities of GO [22]. The GO Consortium itself concedes that the expansion of new terms to the ontology will increase the difficulty of maintaining consistency and curation of the semantic relationships between terms [25]. The preceding section has demonstrated these problems in relation to our project. However the Gene Ontology is by far still the most reliable and complete source for gene annotation. Many developers of GO dependent software have accepted these inherent 'flaws' and attempted to implement counter-measures in their programs that work around these problems [18]. In the next section dedicated to the examination of such tools, we shall discuss some of these devises and their effectiveness.

2.3 Gene Annotation Analysis

In the previous section we briefly reviewed the capabilities of the Gene Ontology as a resource for annotations of gene functions. In order to make use of this rich pool of biological knowledge for the automated analysis of gene data, a tool will be required that allows us to find the GO terms associated with each gene in the list and to analyse the occurrence of these terms for the statistical significance in comparison to a list of randomly chosen genes.

Over the past years, a wide range of software packages for the ontology-based analysis of gene expression data has been developed. Khatri et al. [26] provide a comprehensive overview about all tools available at the time the study was performed. A number of new tools have been published ever since then, yet the most popular once remain to be the ones considered in the paper and the criteria for choosing an appropriate one are still the same.

2.3.1 Criteria for Choosing an Annotation Tool

In the project at hand, we are to analyse a comparatively small set of expressed genes found from intersections of genes from different tissues. It is our main objective to assess the distinct biological processes these genes are involved in and to find out whether they exhibit any annotations that can not be explained away as a mere matter of chance. Hence, we require a tool that will accurately find annotations, analyse them and provide a useful overview about its findings. Along the line of the criteria suggested in [26], we consider the following aspects:

2.3.1.1 Input Values

While all tools will in general take a list of expressed genes as an input, they differ in the types of values they accept, e.g. Affymetrix probe IDs, GenBank accession numbers or UniProt IDs. The latter are the ones we will have as the outcome of our experiments, hence we need a tool which is able to read them.

2.3.1.2 Scope and Level of Abstraction

Many tools restrict their use to only one GO category and only one level within the tree at a time. While the restriction to one category is not a problem for us, since we are only interested in biological processes, an simultaneous analysis on all levels of the BP ontology would be preferable for the ease of analysis.

2.3.1.3 Statistical Test

Numerous statistical models and tests might be employed to judge the significance of the findings obtained. Amongst the most commonly used approaches are the hypergeometric [27] and binomial distribution, the χ^2-test and Fisher's exact test [28, 29, 30]. The hypergeometric distribution is reported to face problems with large arrays, while the binomial distribution and the χ^2-test, on the other hand, both need the expected number of records in the lists to be at least 5 [31]. Fisher's Exact Test, however, can also be applied for small data sets, which is why we will require a tool supporting this test, since many of the intersection results are expected to contain only a very small number of expressed genes.

In order to use Fisher's Exact Test the data will be arranged into a 2×2 contingency table (cp. Table 2.1), where the rows represent the presence of a certain GO term ('present' vs. 'absent') and the columns represent the two data sets [31, 28]. Having filled this table with the values corresponding to a specific GO term in the two data sets studied, we can calculate the probability of observing this table combination as

$$P = \frac{N_{1.}!N_{2.}!N_{.1}!N_{.2}!}{N_{..}!n_{11}!n_{12}!n_{21}!n_{22}!}.$$

In order to get a measure of the statistical significance of one finding, we sum up all P-values that are smaller or equal the P-value observed in the current table. With $\mathbf{P}$ the set of P-values for all possible tables, this amounts to:

$$p(P^*) = \sum_{P \in \mathbf{P}} \delta(P, P^*)P, \; where \; \delta(P, P^*) = \begin{cases} 1 & if \; P \leq P^* \\ 0 & otherwise \end{cases}.$$

The resulting value is known as the two-sided p-value and is generally believed to be a good indicator of statistical significance. The lower this value, the more striking is the enrichment of a GO term in a set in comparison to another set of genes. We would usually only consider genes with a $p \leq 0.05$ to be statistically significant (sometimes

	Gene Set A	**Gene Set** B	Row Total
GO Term ***present***	n_{11}	n_{12}	$N_{1.} = n_{11} + n_{12}$
GO Term ***absent***	n_{21}	n_{22}	$N_{2.} = n_{21} + n_{22}$
Column Total	$N_{.1} = n_{11} + n_{21}$	$N_{.2} = n_{12} + n_{22}$	$N_{..} = \sum_{i,j} n_{ij}$

Table 2.1: Fisher's Exact Test is calculated using an imaginary 2×2 contingency table, where the rows represent the presence of a certain GO term and the columns represent the two data sets. For further details, please refer to the text. (adopted from [31])

even more rigorous criteria like $p < 0.005$ are applied; a result confirming to this threshold of significance is said to be *highly significant*).

2.3.1.4 Reference Sets

As briefly mentioned before, for the purpose of the statistical test we require not only the set of (potentially) interesting genes, but also a reference set. This reference set will serve as a guideline of the frequency with which an arbitrary GO term will exist in a random data set. Only by comparing the set of interest to this baseline, it is possible to assess the actual relevance of the found GO terms.

For microarray experiments, the reference set would typically include either all the genes probed on the microarray or a randomly chosen subset of these [26]. By limiting the reference set to genes which are actually involved in the study, it can be prevented that genes with GO terms occur that would unpredictably alter the results of the analysis. Many tools provide pre-defined lists of reference sets for microarray experiments.

Unfortunately, these sets are not usable for our purposes, since they typically come from post-natal organisms and are consequently not a good reference point for the study of gene expression and their annotations in the developing mouse embryo. This is why we require a tool that allows for the input of custom reference sets, so we can compose a reference set of all genes expressed in all the structures of the mouse embryo investigated in the study ourselves. In the further study we will either use this full list of all genes in the mouse embryonic tissues involved, a subset of it or a random set of genes from the entire gene pool as a reference set.

2.3.1.5 Correction Method for Multiple Tests

A further crucial issue to consider is, which method is used to correct the results for multiple experiments. When the gene lists are analysed for their annotations and assessed for their significance, the tools will in fact test hundreds of hypotheses (one for each GO term occurring). The resulting p-values will not be reliable unless corrected for this fact [26, 32, 33], because they will contain a high number of false positives, i.e. hypotheses which have been approved although the observed differences in both sets were merely a matter of chance. Again, a wide range of approaches have been proposed in the literature. However, one has to be careful when deciding for one of them. Several methods, e.g. Bonferroni correction or Holm's step down adjustment, will only work properly for independent functional categories [34], all GO terms are, however, heavily interwoven and many categories depend upon each other.

The use of another method for correction is consequently preferable. Using the False Discovery Rate (FDR) as a criterion for removing false positives has been found to be the least conservative of the remaining methods [33], meaning it is the one removing the most of the supposedly irrelevant GO terms from our preliminary analysis and hence leaving us with the results most likely to be actual findings. However, this clearly comes at the cost of possibly missing out terms which might have shown up otherwise.

2.3.1.6 Visualization of Results and General Usability of the Interface

Since we will be dealing with a large amount of different gene lists each to be analysed for the significantly over-represented GO terms they contain, it is important that the tool allows for a speedy processing of the data (including the data input, e.g. via saved text files) and a concise presentation of all relevant results. Some tools present the results in the form of an ontology-based gradually expandable tree, which is – although useful for a initial analysis – less usable for a large-scale analysis of many experiments. A presentation of all important findings in one step would be advantageous.

Tool	UniProt Input?	All BP Levels?	Fisher's Test?	Custom Reference Sets?	FDR?	Interface Type	Visualization and Usability
Onto-Express	No	Yes	No	No	Yes	Java GUI	Unusable
GoMiner	Yes	Yes	Yes	Yes	No	Java GUI	Flat/tree view
GO TreeMachine	Yes	Yes	No	Yes	No	Web-based HTML	Flat/tree view
Fatigo/Fatigo+	Yes	Yes	Yes	Yes	Yes	Web-based HTML	Flat view
GoStat	No	Yes	Yes	Yes	Yes	Web-based HTML	Unusable
GoSurfer	No	No	No	Yes	No	C/C++ GUI	Unusable

Table 2.2: Comparison of various tools considered for the annotation and analysis of the results of our later gene expression results. We assessed all tools with respect to the criteria outlined in Section 2.3.1. [26]

2.3.2 Potential Candidates

On the basis of the aforementioned criteria, we considered six tools for our analysis: *Onto-Express* [31, 35], *GoMiner* [36], *GO TreeMachine* [37], *Fatigo/Fatigo+* [38, 39], *GoStat* [40] and *GoSurfer* [33, 41]. Table 2.2 gives an overview of their key features. Of all tools that we looked at, only *Fatigo+* satisfied all our requirements. Most remarkably, we found ourselves unable even to use a few programs – namely *Onto-Express*, *GoStat* and *GoSurfer* – without the need for further conversion of our inputs and extensive studying of tutorials.

2.3.3 Fatigo+

The tool we found most appropriate to our requirements was *Fatigo+* [38, 39]. Grounding on the well-established *Fatigo* tool [42], *Fatigo+* provides a web-based interface rich set of analysis features of annotations, not just GO terms actually, but also KEGG pathways, InterPro motifs, SwissProt keywords and text-mined entities related to diseases and chemical compounds. Moreover, it also allows for analysis with respect to regulatory and structural information, and many more functions whose use would go far beyond the limited scope of this study.

Fatigo/Fatigo+ has been used extensively by biologists over the past years, e.g. for the grouping of up-regulated proteins by biological function [43], the identification of significantly over-represented GO terms associated with microRNAs [44], the systematic functional analysis of gene-expression signatures in relation to cancer [45] or [46]. These are just a few arbitrary examples, an comprehensive list is impossible to give

due to the large amount of publications citing *Fatigo/Fatigo+*.

Fatigo+ employs Fisher's Exact Test for identifying relatively enriched categories of GO terms in a set of expressed genes in comparison to a set of arbitrary reference genes. Both sets can be provided in the form of text files. Alternatively, it is possible to choose a reference set from a number of pre-defined records, none of which, unfortunately, includes genes from a developing Mouse embryo. *Fatigo+* accepts a high number of different input formats, including most importantly the SwissProt/UniProt IDs used in this project. Internally, these are all translated to Ensembl identifiers to create universal cross-references. Apart from *Mus Musculus*, other species, such as *Home Sapiens*, *Rattus norvegicus*, *Drosophila Melanogaster* and others are covered by the functionality of the program.

While we are only interested in GO terms in the BP branch of the ontology, it is theoretically possible to investigate all three branches simultaneously. In a single overview page *Fatigo+* will present the results of all its test across all GO terms and all levels of abstraction. Bar graphs for each level of abstraction give a concise overview about all occurring annotations and the relative frequencies in both sets. Alongside *Fatigo+* shows the result of Fisher's Test (p-value) as well as the corrected value accounting for multiple tests adjusted using FDR (adjusted p-value). It is therefore extremely simple to discover interesting results.

Interestingly, *Fatigo+* – unlike all other tools investigated – also provides a brief summary of the input sets, stating clearly how many genes were in each of them and how many have been actually used for the analysis. This is important, since a considerable number of genes might have been left out, because either the provided UniProt ID was unknown or there simply did not exist any annotations for the given term.

Chapter 3

Methodology

It is our overall objective to identify the set of genes responsible for mesenchyme-epithelium transition process in the developing Mouse embryo. Unquestionably, a manual, step-wise assessment of individual genes is infeasible even for the most patient geneticist. Hence, it was our goal to implement a software solution that would allow researchers to find potentially interesting genes by looking at different biological structures involved in the processes studied and intersecting genes found within them to extract a tissue-independent set of genes supposedly responsible for the process (e.g. mesenchyme-epithelium transition).

In the following chapter we will first assess the specific requirements that a software system needed to satisfy in order to be useful as a tool for researchers and afterwards outline some important points of our solution to the problem. In doing so, we will focus on the core aspects of the program, rather than explaining every little detail of the implementation.

3.1 System Requirements

The principal idea is, that we can find genes responsible for a specific biological process by finding genes common to different parts of the body that undergo this process at some stage. Hence, it will first be necessary to retrieve lists of relevant genes from a reliable source of gene expression information. We chose the GXD database for this purpose. Having extracted these gene lists, different types of computations can

be performed on them to find subsets of interesting genes, which might thereafter be analysed for their biological relevance using a tool like *Fatigo+* (cp. Sec. 2.3.3). In the following sections we shall explain these requirements in more detail.

3.1.1 Clarification of Terms

Before we can proceed any further, it will be necessary to clarify the use of a few ambiguous terms in order to avoid misunderstandings:

Tissues, Structures and Sets of Tissues: A *structure* (or alternatively *tissue*) is a location in the body, where a developing process is taking place so that the structure is being transformed. A set of multiple different, but related structures, when taken as a whole, comprises a functional and distinct component in an organ. This latter component will be termed *set of tissues* or *set of structures*. It is necessary to be aware of the difference of *structures* and *sets of structures* to avoid confusion in the remainder of this paper. In this project, we will investigate the structures from several different sets of structures.

Stages: *Stage* refers to a *Theiler stage* (TS), which is a developmental stage in the mouse embryo [47]. While it is helpful and not entirely wrong to think of a stage as a temporal concept, it is also important to note that each stage is constrained by a distinct biological phase. There are 26 such distinct phases in the mouse embryo. The first stage TS1 is the state of the one-cell egg at fertilisation and the final stage TS27 is a new born mouse. The remaining 25 Theiler stages each refer to the state of the mouse embryo at progressively advanced gestation periods. While each stage corresponds roughly to a period of time, it is not time itself that defines a Theiler stage. Individual embryos develop at slightly differing rates that make time itself an unsuitable gauge of the state of development [48]. Rather each stage has a set of biological criteria that must be fulfilled by the state of the mouse embryo in order to belong to a given stage. For example, a mouse embryo with the earliest signs of fingers falls under TS20. Theiler stages are an essential concept in our basic methodology since we are not collating genes based on different points in time, but from formed and unformed structures, which are in turn categorised under different Theiler stages.

Developmental Processes: Structures change over time to form more specialised components. There are many different types of structural changes in embryo development. The tissues we are analysing here each undergo a different type of developmental process, and it is important to note that this is a distinct process not to be confused with the mesenchymal-epithelial transition process. The latter is one of the many sub-processes that are participants contributing to the former. When we talk about a developmental process in the context of this project, we are always referring to a process specific to the development of a certain tissue or organ. Conversely, the mesenchymal-epithelial process is common to all our tissues of interest. During a developmental process, old structures are replaced by new ones. The former are referred to as the set of *Before* structures, and the latter as the set of *After* structures.

3.1.2 Example: Angiogenesis

In order to further clarify the definitions outlined above, we will have a brief look at one particular process and how it was to be processed here. *Angiogenesis* is a developmental process that involves the growth of new blood vessels from pre-existing vessels. The mesenchymal-epithelium process is responsible for constructing the epithelial cell layer that form the walls of these new vessels [49]. Moreover, Angiogenesis is the process that is responsible for building the system of arteries and veins in the developing heart. Since a new type of structure is being developed, the same tissues where we look for the set of *Before* genes will be replaced by new tissues over time. For the Angiogenesis process in the heart, the structures that will transform over time into these new system of blood vessels are the mesoderm, lateral trunk mesenchyme and the yolk sac of the embryo. Hence we term the set of genes that are expressed in these earlier structures the *Before* set. Once Angiogenesis - and thus the mesenchymal-epithelial transformation as its sub-process - is completed, we look for the set of genes that are expressed in the later tissues, i.e. the arterial and venous system of the heart and we label these genes as the *After* set. By repeatedly asking this question for the different tissues in the embryo where we know the process to be happening, we can obtain a set of genes that are common to each of these tissues.

3.1.3 The GXD as a Data Source

The Mouse Genome Informatics (MGI) project provides integrated access on all aspects of biological information about the laboratory mouse. Of the four participant projects in MGI, the Gene Expression Database (GXD) is the branch responsible for collecting and integrating different types of expression data [50].

We choose to use GXD for two reasons:

- GXD is a community resource. It is accessible to the public for expression queries via its website but more importantly, it also provides a public access SQL interface.
- The GXD standardises the expression data from a wide variety of assays [50, 51]. Part of this standardisation includes a collaboration with the Edinburgh Mouse Atlas Project (EMAP) to develop an anatomy-based system for classifying gene expression data from the mouse embryo [48, 52, 53, 54, 55]. The result of this collaboration is the classification of gene expression data so that information from assays with different spatial, temporal and expression resolution can be queried in a standardised manner.

For our software package, it will be necessary to interact with the GXD taking the way it organizes the data into account.

3.1.3.1 Data Structure

As mentioned earlier, the anatomical dictionary for the mouse is divided into 26 Theiler stages. There is a distinct set of anatomical terms for each stage to describe the anatomical domains existent at a given stage. These terms correspond exactly to the structures in our data computations. Each term (or anatomical domain) represents part of a distinct organ/embryonic material at the given TS of development. In the dictionary, they are organised in a hierarchical tree not unlike the format used in the Gene Ontology. Each stage has its own independent tree, where the major components, for example *embryo* and *extra-embryonic component*, form the top nodes. Child nodes represent more specialised sub-components such as *organ system* and *yolk sac*. Unlike the Gene

Ontology where duplicate nodes of the same term are allowed, the anatomical dictionary for each stage is a true tree because there can only be one instance of a term at any given stage.

Structures are not continuant across all stages. For example the term *yolk sac* exists only in TS9 to TS12. Before TS9 the embryo has not yet developed a yolk sac; whereas after TS12 it has differentiated to form other structures. Terms themselves are not unique unless linked to a TS. *Yolk sac* appears at TS9-12, although the same structure at each stage, is treated differently in the database, i.e. querying for the set of genes expressed in the *yolk sac* at TS9 is not the same as querying for the set of genes in *yolk sac* at TS10.

3.1.3.2 Gene Expression in Relation to Data Structure

The expression data of individual genes are represented in binary states 'On' and 'Off' to indicate whether the gene is being expressed at a certain stage and structure. Each gene entry in the database is related to a stage and a structure. Consequently, we may ask questions like 'Is gene X expressed in structure Y during stage Z?'. As a result, the database will always return us one off three options: 'On', 'Off' or 'Unknown'.

In this study, we will treat genes that do not have entries in the database as being not present ('Off'), instead of 'Unknown'. To what extent does this affect the accuracy of the data? Evidently, an implicit 'Off' is different from explicit 'Off'. The former assumes that the gene is not expressed simply because there is no entry in the database. The latter proves that a research team has actually performed some experiment to check for the presence of the gene, and confirms no such gene being detected. Ideally we should only include explicit 'Off'-genes in our data, however there are two obstacles to this:

- There is a practical limitation of not having enough data: The ratio of 'Off'- to 'On'-entries in the GXD database is extremely unequal. Apparently, the database curators are not considering it worthwhile to enter the gene data for 'Off'-genes[1].
- Using only explicit 'Off'-genes might also have negative implications on the data

[1] One has to remark at his point that experimental data from microarrays contains hundreds, possibly thousands, of genes. Often researchers might choose to enter only those where expression was found into the database, whereas genes which were investigated, but not found, get neglected.

analysis: If we were to enter only explicit 'Off'-genes, we risk creating an artificial bias of the data by assuming that far more genes are expressed throughout a whole process that actually present.

By considering only explicit 'On'-genes to be expressed we are employing some sort of a *Closed-Word Assumption*, which is a common stance to be taken in computational reasoning tasks [56].

3.1.3.3 Data Conversion

In a last small step, the data obtained from the database will also have to be transformed into a format recognized by *Fatigo+* (or any other gene annotation tool used) in order to be usable for the further analysis. The GXD database provides accession keys from a rich variety of databases and ontologies. We therefore directly query for the UniProt/SwissProt-ID of each gene found.

3.1.4 Computations to Be Performed

We know that some of the genes expressed in all of the structures in certain stages are involved in the mesenchymal-epithelium process. However it is not feasible nor helpful for the purposes of proving our hypothesis to investigate each and everyone of them indiscriminately. Each set of structures is located in different organs or locations of the embryo and therefore functionally different. Many genes will not be involved in the process of interest, since most genes will serve tissue-specific functions, which vary from tissue to tissue, organ to organ. Therefore, the fundamental approach of our strategy is to find the set of genes that are common to each set of tissues. By taking the set of genes that are common to all of them, we are able to isolate a smaller set that might show a strong indication to the mesenchymal-epithelial process by eliminating genes that perform tissue-specific functions.

Using this basic approach, we formulate several steps to achieve our objective. We want to gather the genes expressed in each set of tissues so that they collectively form a meaningful set of genes. In order to do this, we organise the structures to their respective set. Each set of tissues comprises a set of *Before* structures and a set of *After*

structures. We then perform four types of computations that each return us a different set of genes:

- *Computation 1*: 'On' genes in *Before* Structures

  The first computation is based on the assumption that, if a mesenchymal-epithelium transition is taking place, the structures will express a set of genes that are responsible for this process in the *Before* structures. Therefore, we obtain the set of genes from the *Before* structures so that every single gene that is expressed in one or more of these structures is included once and only once. Technically, this is the mathematical set known as the union of all the genes in the *Before* structures.

- *Computation* 2: 'On' genes in *After* Structures

  The second computation is based on a similar assumption, namely that, if the mesenchymal-epithelium transformation has recently taken place, the structures may still exhibit expression levels of genes related to the process, or genes that are responsible for the 'wrapping' up of the process. We obtain the union of all the genes in the *After* structures.

- *Computation 3*: 'On to Off' genes

  This is a more complex variation of the first computation, based on the assumption that if a mesenchyme-epithelium process is about to take place, the genes responsible for the process will be expressed only during the beginning and ongoing stages of the process, but should not be expressed once the process is complete. We begin by performing the same steps in computation 1 to obtain the 'On' genes in the *Before* structures. Next, we look to see whether any of the genes in this set are still being expressed in the corresponding *After* structures of the tissue. We do this by collecting all the genes that are expressed in the *After* structures. By comparing the two sets of genes and removing any gene that appears in both sets from our list, we construct a new list of genes that are expressed in, and only in the *Before* structures.

- *Computation 4*: 'Off to On' genes

  The last computation is based on the assumption that once the mesenchymal cells have completely differentiated into epithelial cells, there will be a new set

of genes in place that are responsible for the maintenance of these new cells. Therefore these genes should not be expressed before the process has begun, but only after the new epithelial-constructed structures are fully formed. Again, this computation is similar to that in computation 3. First we perform computation 2 to obtain the list of 'On' genes in the *After* structures. Next we compare these genes with the set of all 'On' genes from the corresponding *Before* structure and eliminate these from the list to give us a new list of genes that are expressed in, and only in the *After* structures.

The four computations outlined above will produce sets of genes whose functions are posited to be related to the developmental process of the structures where they are expressed. Some of the genes could also be transcription factors that regulate the expression of other mesenchymal-epithelial related genes. Furthermore, some of the members within this set of genes should also be related to the mesenchymal-epithelial process. Each of these computations reflect variations of our strategy, which in turns stems from our hypothesis. The first two computations for 'On' genes in the *Before* structures and *After* structures respectively are based on the general assumption that some genes expressed in the structures, regardless of their expression level prior or subsequent to the developmental process, should be involved in the developmental process itself. The last two computations, 'On to Off' and 'Off to On' are actually returning us a list of genes whose expression states have changed during the developmental process itself. We form these sets based on the reasoning that genes whose expression states remain stable throughout the developmental process are more likely to serve other on-going functions and hence are not so likely to be specific to the process of our interest.

Once we have performed the four computations above, we have four different sets of genes from each set of tissues. We then enter the next stage of our strategy, a higher level comparison between different sets of tissues. We formulate the fifth and final type of computation:

- *Computation 5*: Intersection between Sets of Tissues

  The computation for this involves first performing one type of computation (one of computations 1-4 above) on each of the sets of tissues we wish to intersect. Once we have done this, we compare two or more lists of genes and retain only

those genes that are common to all the lists in our comparison. In mathematical terms, this new set of genes is an intersection of sets. In order to obtain a complete set of data, it is desirable to perform all the possible intersections between the tissues. There is a total of

$$S = \sum_{k=2}^{N} \left(\begin{array}{c} N \\ k \end{array} \right)$$

distinct permutations of intersections of this at least two different sets out of N total sets. Although it is technically possible to intersect sets of genes that were calculated using different types of computations, such intersections would not return us any logically meaningful results, furthermore some of these intersections would be difficult to interpret from the viewpoint of our hypothesis. Therefore we impose a constrain so that only sets of genes that were obtained using the same methods are allowed to be intersected.

At this stage, our goal is to compare the genes from different tissues where the mesenchymal-epithelial transition is happening in order to refine our data even further so that genes involved in the tissue specific sub-processes of the larger on-going developmental process are removed from the final mesenchymal-epithelial candidate list. What we have prior to this stage are sets of genes that are very likely to be involved in developmental processes, but not all these genes will be involved in the mesenchymal-epithelium process. We hypothesise that the intersection set of developmental genes from various tissues will be more likely to be involved in the mesenchymal-epithelium process compared to the complementary sets.

Figure 3.1 shows an exemplary case how intersections can be performed to obtain a candidate set of mesenchymal-epithelial related genes.

3.2 Implementation of a Solution

We have implemented *Mouse Genome Intersector* (Fig. 3.2), a simple, Java-based tool with a graphical user interface (GUI). The tool allows to create a list of tissues consisting of structures which can be added by dynamically browsing the EMAP ontology for entries and querying GXD for the genes found in the respective structure. The user

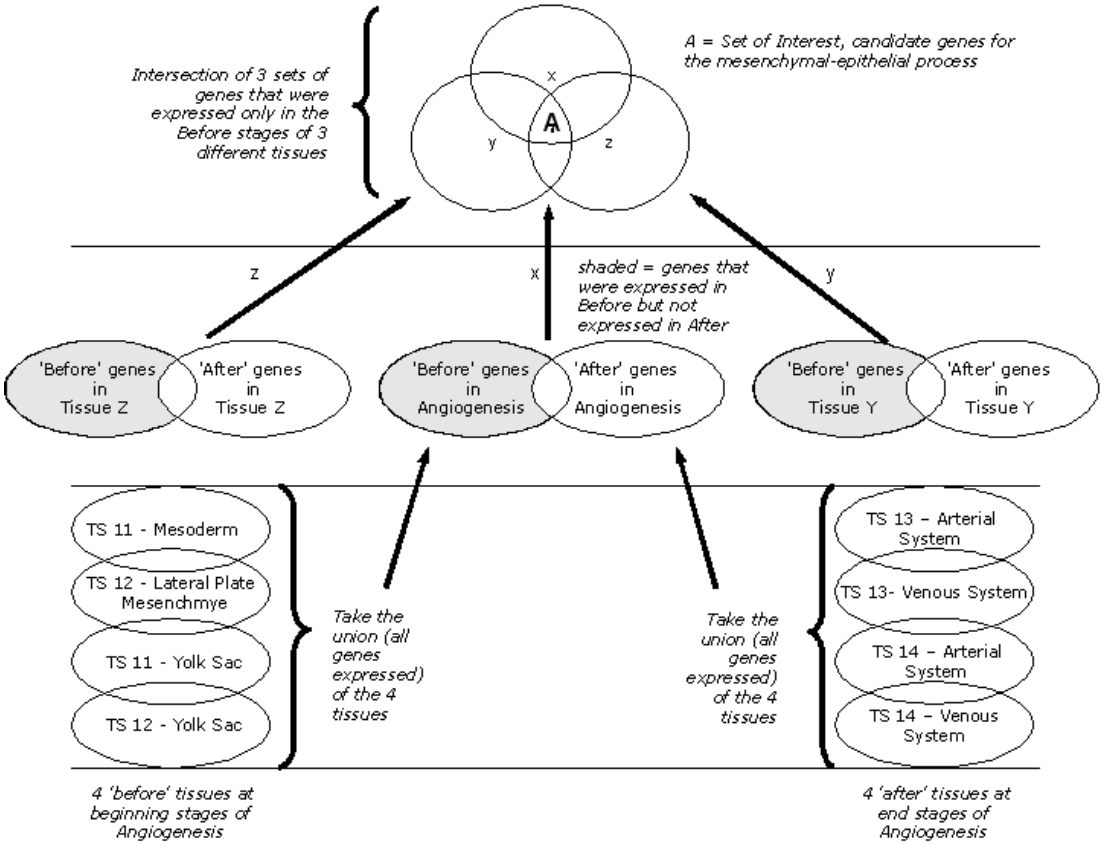


Figure 3.1: An example of how we can use different set operations on four sets of genes in order to obtain a candidate set of genes for the mesenchymal-epithelium process. We start by taking the union of all the genes in the *Before* structures and separately the union of all genes in the *After* structures and repeat this step for all tissues studied. We then calculate the set of genes which are in the *Before* but not in the *After* structures (computation 3) for each tissue. The intersection of these will yield our candidate set.

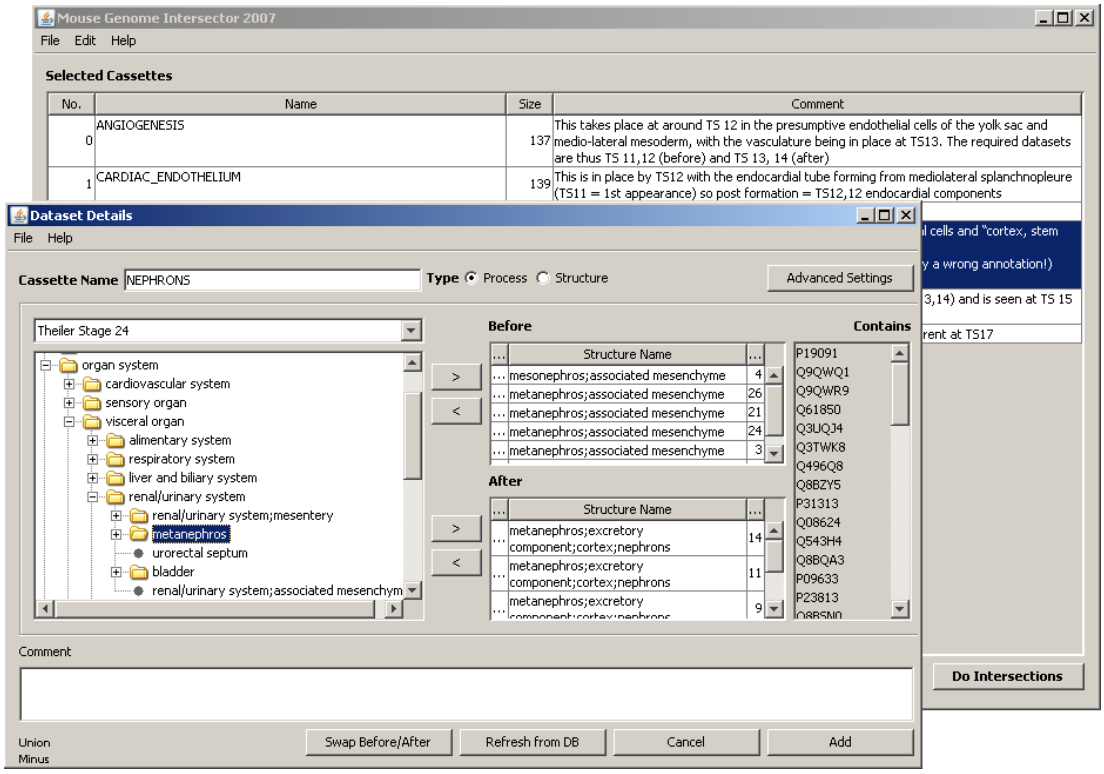


Figure 3.2: An exemplary screenshot from the Mouse Genome Intersector, the tool implemented in due course of this study. The windows shown are the main window (in the background) and the details of one `BeforeAfterCassette` in the foreground.

has the possibility to choose between different options in order to select which of the computations listed in Sec. 3.1.4 is to be performed. The program will then compute all the possible intersections of the chosen tissues and present the results to the user, who may chose to discard, further study or save individual results.

Figures 3.3 and 3.4, give UML class and sequence diagrams of the program respectively. The class diagram has been simplified in order to capture the principal relations between the important components of the software in a concise manner. The sequence diagram provides an overview about a typical series of interactions between the user, the program and the database, as they might be carried out in one round of experimental analysis. For further details of the program, please refer to Appendix A, where we provide full class diagrams, as well as screenshots and details about database queries used.

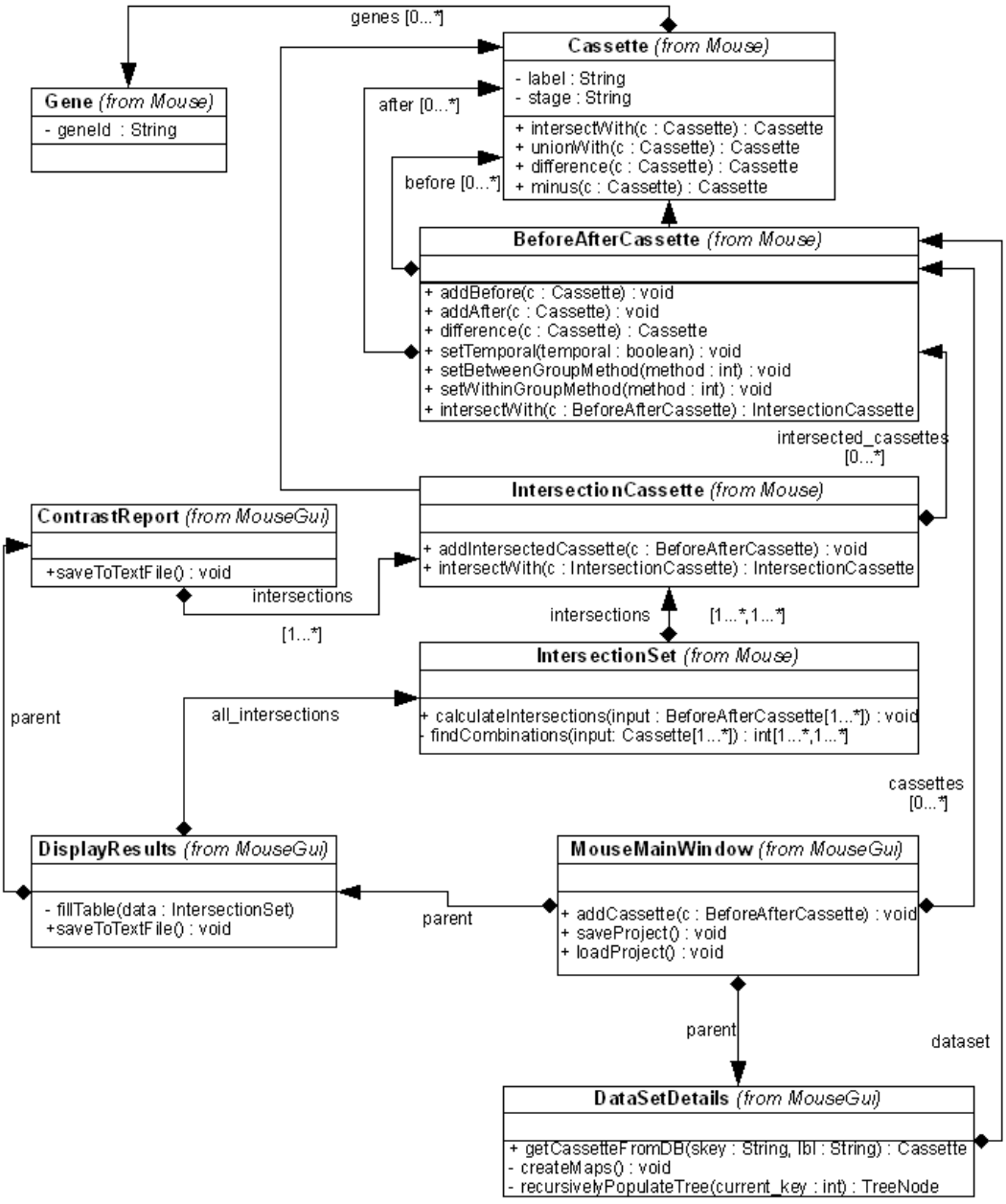


Figure 3.3: A massively simplified, conceptual UML class diagram of the main classes of the project. Note that several unimportant classes and numerous irrelevant methods (e.g. getter- and setter- methods or event-handling methods from the GUI) have been left out. Arrows without tails indicate inheritance relations, arrows with tails account for relational attributes.

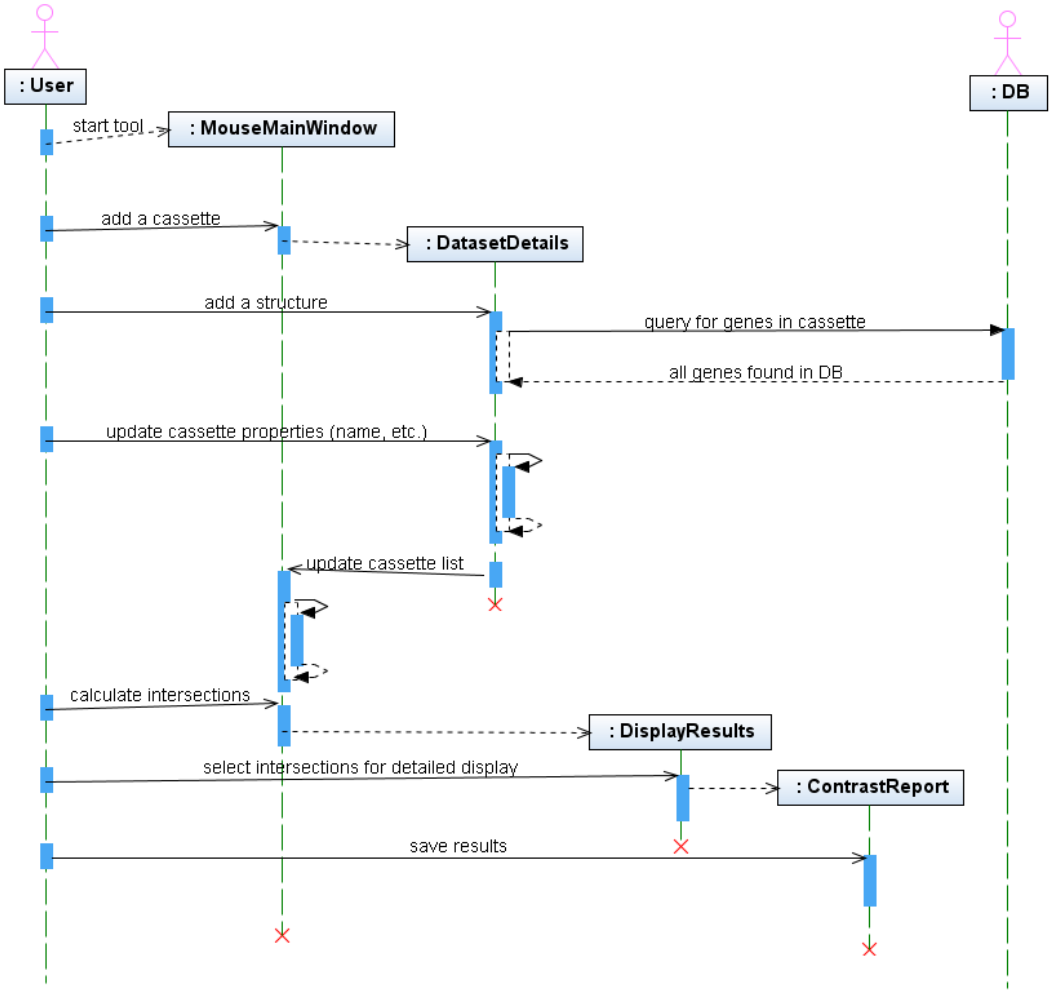


Figure 3.4: This UML sequence diagram depicts a typical flow of actions as they might be carried out by a scientist using the tool for creating a data set of gene cassettes, for intersecting these cassettes and then for analysing and saving the results.

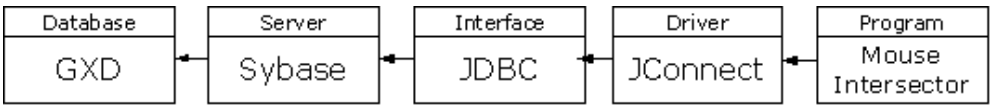


Figure 3.5: Illustration of the different stages of interaction between the program and the database.

3.2.1 Interaction with the Database

GXD is implemented in the Sybase Relational Database Management System [50]. The database supports direct SQL access for remotely connected programs. JDBC [57] is the Java API that we implement to connect with GXD. In order to actually establish and maintain a connection to the server side, we need a client-side adaptor called a driver. The program utilises a Type 4 driver, JConnect [58], which is a Java package provided by Sybase. Figure 3.5 depicts the sequence of interactions of the different components.

3.2.1.1 Accessing Relevant Data

Multiple tables are accessed in the query of a single gene. This part is a brief description of the tables that are involved in our query.

`ACC_Accession` is the master table that holds the MGI ID's for all existing gene and gene products. As its name suggests, the table holds the accession numbers all genes. An accession number is a unique identifier for a sequence. MGI and GXD creates and uses its own ID strings which all begin with the prefix 'MGI' followed by a unique string. Besides listing each gene by their internal MGI ID, `ACC_Accession` also contains the accession numbers or names that the gene is referred to by different sources, such as its UniProt/SwissProt and GenBank ID.

`GXD_Expression` is the table that stores the known gene expression states for all known sequences. The data in `GXD_Expression` is summarised from assays to accommodate efficient querying, so that a user can extract an 'On' or 'Off' reply for the presence of a gene in a structure. Each entry contains a gene, its expression state, and an assay key so that each discovery can be traced to its specific assay. The entry must also have a structure key, which is a unique number to indicate where and when the tissue was expressed.

`ACC_Accession` and `GXD_Expression` are the two primary tables that hold the information we need. Entries are linked in the table via a marker key, which is another unique number used to identify the same gene across different tables.

In order to specify the structure and stage where genes should be queried, we require a third table called `GXD_Structure`. This is the master table for the Mouse Anatomical Dictionary . It holds the terms of the entire dictionary. Each term is assigned a unique number that identifies both the stage and the structure. For example, the previously mentioned yolk sac at TS10 has structure key of 154 and *yolk sac* at TS12 has Structure key 402.

So a gene expression query begins with the selection of a stage and structure by the user. We then look up the structure key which identifies this stage and structure in `GXD_Structure`. Having identified the structure key, we search for all genes in `GXD_Expression` that belong to this structure key and retain only those genes where expression state is 'On'. However, at this point, the genes are not identified by any name other than the marker key. Hence for gene, we match the marker key to entries with corresponding marker keys in the `ACC_Accession` table. The corresponding `ACC_Accession` table entry will return us the proper gene names for all our genes in the form of MGI ID's. The result from this query is a list of genes where expression state is 'On' for the user selected structure and stage.

3.2.1.2 Converting MGI Accession ID's to UniProt ID's

As mentioned earlier, it is desirable to use the UniProt ID's as our gene naming convention in the program because it is more commonly used than the MGI ID's. Luckily, the `ACC_Accession` table also includes UniProt ID's for almost all entries, so we can query for the UniProt ID's instead. Only in a few instances, there will not be a UniProt ID available for an entry, in which case we use the MGI ID. If one of these happens to be in the studied candidate sets for the mesenchyme-epithelium transition process later on, special consideration will have to be paid to this entry.

3.2.1.3 Data Organisation

All the genes found in individual stages and structures will be stored in containers called `Cassettes`. Several of these cassettes taken together can make up a whole tissue (with *Before* and *After* structures). More details on the data containers will follow in the next section.

3.2.2 Core Classes

This section describes the classes involved in one round of analysis. A round of analysis is defined as the user executing the program, creating several lists of genes from several structures and developmental processes, and then performing the intersections. We also illustrate how the five types of computations discussed earlier on are implemented from an object oriented point of view. All the core classes implementing the underlying logic behind the computations are included in the `Mouse` package of the software. An entity-relationship diagram (ERD) depicting how the different modules are linked is shown in Fig. 3.6.

The different classes are:

- `Gene`

  The genes which we queried in the previous section are instantiated as objects of the class `Gene`. Each instance has the attribute `geneID` which stores the accession number of the gene. The class also has an `equals` method that checks to see whether two objects contain the same gene. This concept is important later on to ensure we have no duplicate genes in one list.

- `Cassette`

  `Cassette` is a class which contains a vector of genes. Each cassette can be viewed as a set of genes from a distinct origins. Genes can be grouped in different categories using the cassette class. The most common and basic cassette is one which lists the genes in a single tissue and stage. We can also create a cassette of genes from a structure by combining several cassettes of single tissue genes. The class defines four methods for these combinations:

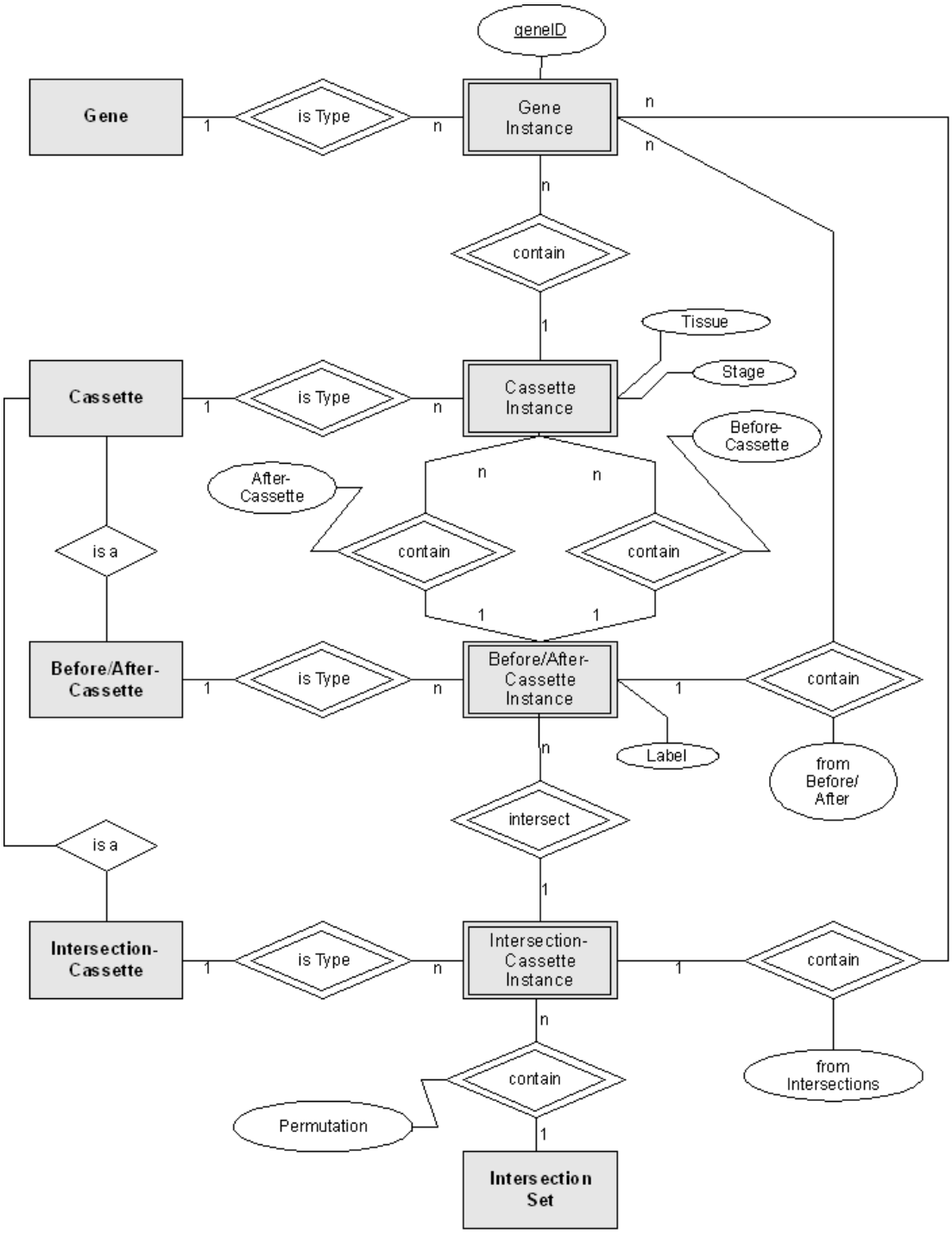


Figure 3.6: An entity relationship (ER) diagram of the `Mouse` package. Note that, proceeding from top to bottom, we encounter increasingly more complex classes entailing one or several instances of the object from the previous levels. For further details please refer to the text.

  - `intersectWith(Cassette)` returns a new cassette that contains the genes found in both and only both of the cassettes. This method implements the concept of what constitutes the common genes from a group of tissues.
  - `unionWith(Cassette)` returns a new cassette that contains the genes found in either one or both of the cassettes, removing double occurrences of the same gene. This method implements the concept of what constitutes collective genes from a group of tissues.
  - `minus(Cassette)` returns a new cassette that contains genes found in the main cassette, but not in the second cassette. This is the implemented concept of 'On to Off' genes. Alternatively, this method can also be used to find the 'Off to On' genes, depending on which object is calling the method and which object is the argument of the function.
  - `difference(Cassette)` computes a new cassette with all genes expressed in only one of the cassettes (exclusive or). Hence, it returns both the 'On to Off' and 'Off to On' genes.

- `BeforeAfterCassette`

  This class, as the name implies, contains a first vector of cassettes from the group of *Before* tissues, and a second vector of cassettes from the group of *After* tissues. The *Before* and *After* tissues must belong to the same developmental process, which is stored in the `label` attribute of the class. A user can add or remove cassettes from an object of this class, which calls the methods `addBefore()`, `removeBefore()`, `addAfter()` and `removeAfter()`. The `BeforeAfterCassette` is an inherited class of `Cassette`, thus it is similar in structure to a cassette. It differs from a cassette not only because it contains multiple cassettes, but also because it can compute the list of genes we are investigating in the developmental process by calling the four inherited methods previously described in the `Cassette` class. It stores this new list of genes in a vector attribute. Thus by specifying the *After* tissues, the *Before* tissues and the methods of computation, a `BeforeAfterCassette` class can return us the genes that are either 'On to Off', 'Off to On' or both in a given developmental process. If there are no cassettes of *After* tissues, it can simply return the set of genes that are 'On' from a group of tissues. Depending on whether the user specifies any *Before* and *After* cassettes, an object of this class can either be an instance of a developmental process or a simply a group of cassettes.

Theoretically it is possible to compute sets of genes in a developmental process using the `Cassette` class alone. However the user would have no way of deciphering how a list of developmental process genes was computed because the underlying information used to build and compute the list would be lost. If the list was computed from many structures it would be tedious to rebuild as well. The `BeforeAfterCassette` provides a framework for editing or referring to the group of *Before* and *After* tissues used in a computation. By implementing the `BeforeAfterCassette` class, which is a serialisable object in our program, all cassettes and computations linked to the developmental process can be loaded systematically by the program so that the user can trace the origin of the genes computed by this class, or edit the computation by adding and removing new cassettes.

- `IntersectionCassette`

  An object from this class stores the results from the intersection of two or more `BeforeAfterCassettes`. Again this class is inherited from `Cassette`, so it ultimately returns us a list of genes. However, the previous classes only stored genes from tissues, groups of tissues, or genes derived from a developmental process. By instantiating this class, two or more such cassettes can be intersected so that a new set of genes is returned by the intersection.

  For example, an object from this class would store the set of genes common to the three developmental processes in *Angiogenesis*, *Somites* and *Cardiac Endothelium*. The `IntersectionCassette` object only lists genes that are strictly common to all of the cassettes in the intersection, therefore it is called at the last stage of our analysis, only after we have used `Genes`, `Cassettes` and Before-After-Cassettes to return us the lists of genes from developmental processes.

- `IntersectionSet`

  This is the final class to implement in the analysis. It represents the set of all possible intersections from one analysis round. If there are, for example, five `BeforeAfterCassette` objects that have been defined by the user, this class stores the list of genes resulting from each and every intersection that can be performed between two or more of the objects. Thus it is essentially a list of all possible IntersectionCassette objects. There will be only one instance of this class at any one execution of the project.

The `IntersectionSet` defines a method called `calculateIntersections` that computes all different ways to intersect a given list of sets, and lists out all the possible permutations. For example for a list of sets (A,B,C,D) it would return us the permutations (AB), (AC), (AD), (BC), (BD), (CD), (ABC), (ACD), (BCD), (ABD) and $(ABCD)$. The program then proceeds to create an instance of `IntersectionCassette` for each permutation, thereby creating a set of IntersectionCassette objects.

By the end of instantiating the fifth class, we have obtained the final results for our analysis. The program is designed to enable the saving of individual as well as lists of `BeforeAfterCassette`s as serialised Java object files so that it can be reloaded as an entire project when the program is restarted. This allows the user to refer and edit the original source of computations. Furthermore, because the program is dynamically connected to the GXD database, we also implemented a `refresh` function which performs all the queries related to the loaded project or `BeforeAfterCassette` to obtain the most up-to-date genes from the database. This is especially useful if the user wants to check whether the database has been recently updated with any new genes relating to any of the structures being analysed without having to re-build an entire project.

3.3 Summary

Our methodology for finding genes responsible for a specific biological process (e.g. mesenchyme-epithelium transition) can be summarized briefly as follows:

1. Specify all tissues in which the studied process takes place by repeatedly adding the genes from all *Before* and *After* structures in this tissue.
2. Depending on which part of the tissues we want to study at the moment, retain only the Before or After component of the tissues, or define the genes of which of the two components is to be subtracted from the other one.
3. Calculate the intersections of all gene lists.
4. Analyse the found genes using an gene annotation tool on the basis of the GO, e.g. *Fatigo+*.

In the *Mouse Genome Intersector* software we have implemented a tool that allows us to easily carry out steps 1-3 and to produce the results in a format that can be used directly as an input for *Fatigo+*.

Chapter 4

Experiments and Results

This chapter will present the results obtained using our program and the results of our evaluations. There are two aspects to our evaluation. The first aspect is an computational analysis of the annotations for the sets of interest using *Fatigo+*. The second aspect is a biological review of the intersections to manually identify possible candidate genes for the mesenchymal-epithelial process. The latter section has been contributed by Dr. Jonathan Bard, an expert the field.

As the first type of analysis will form the bulk of this section, I shall outline the procedures for this section and our expectations. From our intersections, we hope to obtain substantial numbers of genes to analyse with *Fatigo+*. Next we hope to find significant terms for each of the intersections. The fact that an intersection is more related to one or more biological processes than a randomly selected list of genes is a good indication that the set as a whole has some specialised functionality. We also want to compare our intersections against other types of reference sets that could be more meaningful than random sets.

4.1 Computational Analysis

4.1.1 Results

In this section we will present the tissues and developmental process from which we derived our results.

	Tissues	
Process	**Before**	**After**
Angiogenesis	TS11: Mesoderm TS12: Lateral Plate Mesenchyme TS11-12: Yolk Sac	TS13-14: Arterial System TS13-14: Venous System
Cardiac Endothelium	TS11: Mesoderm TS12: Lateral Plate Mesenchyme TS11-12: Yolk Sac	TS12: Early Primitive Heart TS12: Primitive Heart Tube TS13: Endocardial Tube TS13: Outflow Tract, Endocardial Tube
Somites	TS12-20: Unsegmented Mesenchyme	TS12-22: Somite
Mesonephric Tubules	TS13: Interm. Mesenchyme TS14: Nephric Cord	TS15-16: Tubule
Nephrons	TS20: Mesonephros, Assoc. Mesenchyme TS21-25: Metanephros, Assoc. Mesenchyme	TS21-25: Metanephros, Execretory Component
Nephric Duct	TS12: Trunk Mesenchyme, Intermediate Mesenchyme TS13: Nephric Duct	TS15: Nephric Duct

Table 4.1: Overview of the different structures studied.

4.1.1.1 Developmental Processes and Tissues Investigated

We compare the gene expression data from six sets of tissues each undergoing a different developmental process. Each developmental process is related to a defined set of *Before* tissues and *After* tissues as listed in Tbl. 4.1. This is the gene pool, from which all the subsequent intersection results shall be obtained.

4.1.1.2 Categories of Results

As outlined before, we compute and store four separate lists of genes for each developmental process. A gene can belong to one or more of these lists, which are categorised based on whether the gene is expressed as:

1. 'On' genes from *Before* tissues
2. 'On' genes from *After* tissues
3. 'On to Off' genes from *Before* and *After* tissues
4. 'Off to ON' genes from *Before* and *After* tissues

Hence we obtain 24 separate lists of genes from the six developmental processes. This is the collective raw data on which we will perform the intersections. We will also

be utilising these lists of genes in our compilation of reference sets for evaluation, as demonstrated later on in this chapter. Please refer to Appendix B for a full list of all the genes in these lists.

The lists of genes for Nephrons, Nephric Duct and Mesonephric Tubules are disproportionately smaller than Angiogenesis, Cardiac Endothelium and Somites. This is probably because there were more tissues defined for each of the latter processes.

4.1.1.3 Intersections

We identify the genes that are common to more than one tissue by performing all possible intersections. Intersections are only allowed between lists falling under the same category (see previous section). For example, we do not compute the intersection of 'On' genes in Before Angiogenesis and 'On-Off' genes in Somites, since the genes resulting from cross-category intersections would be extremely difficult or even impossible to interpret for the purposes of inferring our mesenchymal-epithelial candidate genes.

Having this constraint in place, we then compute the various intersections that can take place for six developmental processes. Due to the small size of some of the pre-intersection lists, many of our intersections were empty sets. We also did not obtain any genes from intersections of four or more tissues. The majority of intersections which did return any results did not contain more than 5 genes. In general, intersections falling under the category 'On' genes in *Before* structures returned the most genes.

We summarise the results of our intersections in the Tbl. 4.2. The full list of genes found in the intersections can be found in the Appendix B. Permutations which did not yield any genes in any category (i.e. $A \cap C \cap S \cap Nd$) are excluded from the table. Intersections involving less than three tissues are also not listed since they will not be considered in our evaluation.

4.1.1.4 Candidate Sets

The aim of our evaluation is to find statistically significant terms for the intersected genes. We wish to run each separate intersection through *Fatigo+*. Each non-empty

	Intersection of 4 Tissues			**Intersection of 3 Tissues**								
Angiogenesis	A	A	A	A	A		A		A		A		A
Cardiac Endothelium	C	C	C	C	C	C		C	C			C	
Somites	S		S	S		S	S	S		S	S		
Mesonephric Tubules	M	M						M		M	M	M	M
Nephrons			N		N	N	N			N			
Nephric Duct		Nd							Nd			Nd	Nd
Category of genes													
ON genes in *Before*	2	2	3	34	8	3	3	2	2	2	2	1	1
ON genes in *After*							1			1	1		
ON to OFF				5	5				1				
OF to ON													

Table 4.2: Summary of the number of genes found in each of the intersections.

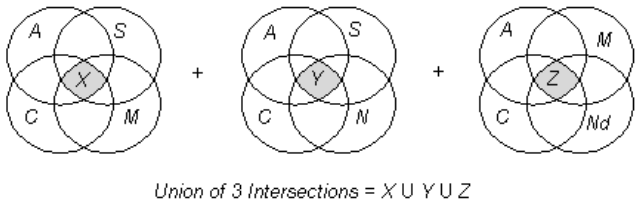


Figure 4.1: Illustration of the union of sets.

intersection involving three tissues and above is a separate candidate set to be considered.

Other than using individual intersections as candidate sets, we also form an 'overall' intersection set that comprises of the union of all intersections (cp. Fig. 4.1). Note that this is different from the union of all lists of genes themselves, because we first perform the intersections before taking the union itself. The difference is illustrated in Fig. 4.1 using the three intersections of 'On genes in *Before* tissues' shown in the Tbl. 4.2.

4.1.1.5 Reference Sets

The reference set we use depends on the candidate set it is being compared against. In general, for each candidate set, there are three types of comparisons we want to make:

Random Sets – Using a Java program and its inbuilt `random()` method, we compile sets of randomly selected genes taken from the same source as our results, i.e. the `GXD_Expression` table in GXD. We use three sizes of random sets contain-

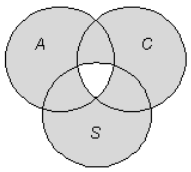


Figure 4.2: Illustration of complement sets.

ing 1000, 500 and 100 genes respectively. This is in part to cater to some of the smaller candidate sets. However, 100 genes may not be a statistically fair distribution of genes across various functional categories, thus random sets of 1000 and 500 genes are the most frequently used in our evaluation. In order to perform the same type of comparison more than once, we also compile second and third random sets for the sizes 1000 and 500.

Complement Sets – A complement set is a list of genes that are not in the intersection which forms the candidate set, but expressed somewhere in the tissues involved. There are different complement sets for each permutation and each category of gene expression data studied in our evaluation. For example, the permutation (A,C,S) returns two non-empty intersections - one for the list of common 'On genes in *Before* structures', and one for the list of common 'On to Off genes'. The reference set used for the former is all the 'On genes in *Before* structures' that are expressed in either one or two, but not all three sets of tissues involved in the developmental processes Angiogenesis, Cardiac Endothelium and Somites. The reference set for the latter follows the same theory but uses the 'On to Off' genes instead. Mathematically, the complement set described above can be presented as $(A \cup C \cup S)$ $(A \cap C \cap S)$, as illustrated in Fig. 4.2.

In theory, complement sets provide a more stringent reference than the random set, because we only consider genes originating from the tissues participating in the intersection we wish to compare. The purpose of complement sets as part of our comparisons is based on an issue for similar comparisons of microarray data in [26]. They point out that, if a gene is never included as a probe in a microarray set, its expression will never be detected, therefore using these genes in a reference set may prove inappropriate for the comparison. Similarly if a gene is never expressed in any of the tissues involved in our intersections, comparing our candidate sets against such genes could give an inaccurate, or over-optimistic picture of their functionality. Genes in our random sets are derived from the

entire mouse genome, some of which are never expressed in the tissues we are investigating. By comparing our candidate sets to the random sets we will be able to get an overall picture of how these genes are functionally different from genes expressed in the whole genome. However, random sets cannot provide us with an in depth view of how commonly expressed genes in several tissues are different from other genes in the tissues. By using complement sets for a part of our evaluation, we acknowledge that even if an intersection of genes is more functionally specialised then the entire mouse genome pool, it may not be any more specialised than other genes expressed in the tissues used to derived the intersection.

Combined Complement Sets – Obtaining these sets is a matter of joining the lists of genes from all six tissues. There are four combined complement sets – one for each of the four overall candidate sets. Each reference is a list of genes from one out of the four categories of genes. After joining these lists, we remove any genes that are found in the corresponding overall candidate set. The reasons for using these sets are similar to those discussed in for the simple complement sets. As a result, these lists are the inverse of the overall candidate set, i.e. genes that are only exclusively expressed in one of the tissues, or even if they are co-expressed, then it is only in two tissues.

4.1.2 Criteria and Expectations for Comparisons

First of all, we hope to obtain sets of interest that are large enough to withstand the statistical analysis of *Fatigo+*. We limit our sets of interests to intersections of three tissues and above. Although there were many intersections of two tissues that returned significantly larger sets of genes, it was deemed that such genes which are co-expressed in only two out of the six tissues are not 'common' enough to be worth investigating for our hypothesis.

4.1.2.1 Significant Terms

For each of the candidate sets, we will look for any significant annotations in the Biological Process ontology of the Gene Ontology. The idea of what constitutes a 'significant' term is to see whether a given set contains more genes related to a certain

function than one would expect to find and to ensure that this difference is not just a matter of random fluctuations (cp. Sec. 2.3.1.3).

A term that is annotated at a higher or lower percentage than the expected value might qualify for significance. *Fatigo+* does not set a threshold for how large the difference in percentage must be in order for a term to be considered significant. Quantitative values of percentages are not a reliable indicator for determining whether it is a significant enrichment. This largely in part due to the fact that *Fatigo+* allows candidate and reference sets of varying sizes, this is particularly true for our case because there is a vast difference between the sizes of our candidate sets and reference sets.

In hind sight, most of the significant terms obtained in our comparisons possess percentage differences of at least 10% and above. In rare cases where the difference is smaller than 5%, the expected percentage is also smaller than 10%, in which case a difference of even 2 or 3% is considered significant enough to qualify the term. Cases of such assessments are not uncommon because *Fatigo+* uses p-values obtained by Fisher's Exact Test and adjusted by FDR in conjunction with percentages to measure for significance.

For all comparisons in the subsequent section, we use an FDR-adjusted p-value of 0.05 as a cut off point for determining significant terms. This is also the default value used by *Fatigo+*. Terms that are larger than 0.05 are discarded regardless of how large the difference in percentages.

We only record the occurrence of non-redundant significant terms in our comparisons. The candidate set could be related to multiple terms where each of these terms is actually a more specific instance of the previous term. For example, a set could have an enrichment for all three terms, *Ear Development* [GO:0043583], *Inner Ear Development*[GO:0048839] and *Inner Ear Morphogenesis* [GO:0042472], the last term being a grand-child of the first. In this instance, *Fatigo+* reports all three of them, but selects the term at the deepest node to include in the summary of non-redundant terms. This simplifies our task of sifting through the terms for each comparison to find the most specialised instances of related biological processes.

4.1.3 Comparisons

We run four rounds of comparison for each of the four categories of genes where there are non-empty intersections of three tissues or more. Due to the sparse data returned from some of our categories, we also ran the same comparisons for intersections of two tissues just to see whether anything interesting would turn up. The next four sections shall be a summary of each comparison .

4.1.3.1 Intersections vs. Random Sets(1000, 500, 100)

1. *ON Genes in Before Tissues* – Significant terms were only found in two intersections, $A \cap C \cap S$ and $A \cap C \cap N$. However it is noted that the sets of *Before* tissues for Angiogenesis and Cardiac Endothelium are exactly the same, hence we are effectively only taking the intersection of two sets of genes.
2. *ON to OFF Genes* – no significant results.
3. *ON Genes in After Tissues* – no significant results.
4. *OFF to ON Genes* – no significant results.

4.1.3.2 Intersections vs. Complement Sets

1. *ON Genes in Before Tissues* – no significant results.
2. *ON to OFF Genes* – no significant results.
3. *ON Genes in After Tissues* – no significant results.
4. *OFF to ON Genes* – no significant results.

4.1.3.3 Changing Genes vs. Static Genes

1. *ON-OFF and OFF-ON genes in the a set of developmental process tissues* – no significant results.
2. *Intersections of ON-OFF/OFF-ON genes vs. all ON-OFF/OFF-ON genes from the same tissues but not in the intersections themselves* – no significant results.

4.1.3.4 Functional Profiles of Common Genes and Non-Common Genes

One problem with the previous comparisons was that most of our individual intersections were simply too small to return any statistically significant results. This does not necessarily mean that the genes themselves are not more functionally related to some processes more than others.

In this section, we discuss a means of employing our results to obtain some meaningful results. The most obvious way of using the data from small candidate sets is to combine all the lists to form a larger set. We take the precautionary measure of acknowledging beforehand that if we do obtain any significant terms from such a combined list, such terms can only apply to the entire set of genes as a whole and should not be assumed to be related to any individual intersection.

For this task, we focus on the category of 'On genes in *Before* tissues' because: (1) the intersections from this category returned us the most results to work with, and evaluations of individual intersections proved more promising than the intersections from other categories, (2) from a biological aspect evaluation (which shall be discussed in detail in Sec.4.2) genes from this category were deemed the most interesting.

The method we use to combine the intersections to form what we shall call the 'combined intersection set' has already been discussed in Sec. 4.1.1.5 above. It is worth mentioning again that all of the genes found in this set are common to three or four of the tissues, thus we are actually investigating the collective set of co-expressed genes. We compare this candidate set against five different random sets, two of these have 500 genes in them, and the remaining three consist of 1000 genes each. It has to be noted here that the five random sets are not disjoint, i.e. genes found in one set can occur in other sets. However since the gene pool we use for our random sets consists of virtually the entire mouse genome, the chances and frequencies of re-occurring genes across sets are very low. The results that we obtain here are probably reproducible using disjoint random sets.

The combined intersection set contains 43 genes. While this is still disproportionate in size to the reference sets, they proved large enough to capture a list of significant terms in each round. As is expected when different reference sets are used, not all five comparisons returned the exactly same terms. We recorded the significant terms for each of the five rounds. In each case, only the non-redundant terms were noted.

This resulted in a compilation of only the most specific biological processes where our candidate set was enriched. If one of the later comparisons introduced a less specific process as a new significant term, we check each of the previous comparisons to see whether this process was a significant term albeit in the redundant list, and record it accordingly. Using this reiterative method, we created a functional profile of our co-expressed genes using a total of 51 biological processes in which our overall candidate set were enriched.

One challenge of interpreting the results at this stage is that a majority of these terms do not show any obvious indication to the mesenchymal-epithelial transformation, although we did not investigate each significant term in depth because such an attempt would not be feasible. Given enough resources, the functional profile we formed with the 51 terms alone may be sufficient for a biological expert to tell whether our co-expressed genes as a whole, contain any significant terms related to the mesenchymal-epithelial process. However, the terms themselves are not informative enough for a layman to evaluate objectively. Relying on manual assessment alone would also be very time consuming even for the most experienced biologist, and ultimately not an objective evaluation from an informatician's point of view until irrefutable evidence from laboratory experiments could be produced.

The problem we faced is also compounded by the fact that any random selection of genes could also yield significant terms when compared against another random set. In short, simply proving that our set of co-expressed genes are functionally enriched is not enough. Thus at this stage although we know that our co-expressed genes were definitely enriched in various biological processes, we cannot interpret the results to contribute anything towards our hypothesis.

In order to overcome this problem, we proposed overlapping the functional profile we have against a different functional profile to see whether we could get a contrast. The advantages of such a comparison are two fold: First of all, even a layman will be able to spot an obvious difference between the two profiles; secondly, such a comparison is more objective and less time consuming than exhaustively investigating each term manually. The final question then is what type of functional profile should we choose?

The most obvious choice at first would be a functional profile of randomly selected genes. However this would not contribute any further information to what we

already know - that our candidate set contains enriched terms. What we need at this point is some way of telling whether our co-expressed genes are any different from the not co-expressed genes in the same tissues (*complementary genes*). When we ran the comparison directly between individual intersections as candidates and corresponding complement as references which we described in Sec. 4.1.3.2, no significant terms were found. At this point it is tempting to assume that the two groups of genes are no different from one another. However, from an objective point of view without any a priori speculation, it is still the most logical choice to contrast our functional profile against.

The term 'complementary genes' is a little misleading because the list may contain genes originating only from one set of tissues, or they may be genes occurring in not more than two tissues at the most. In any case, these genes are less 'common' than our co-expressed genes which have to belong to at least three tissues. The list of these genes is actually the combined complement set as presented in Sec. 4.1.1.5. There is a total of 250 genes in this set.

We constructed the functional profile for our combined complement using the same iterative process discussed above. In order to ensure consistency, we also used the same five random sets in the previous comparison as reference sets here. Approximately half of the significant terms annotated to this set were already included in the previous set. However, the five rounds of comparisons introduced a total of 34 new terms to our existing compilation of non-redundant terms used for the first functional profile. We also ensure by repeating the previous five comparisons for co-expressed genes to ensure that none of these 34 terms are significant.

As a result, the two combined functional profiles use a total of 81 non-redundant terms. Fig. 4.3 shows the two functional profiles side by side as well as a comprehensive list of all the non-redundant terms and ten comparisons in total for both sets of genes. A cursory glance at the table is sufficient to tell that the two profiles are different. While both sets share approximately a third of the 81 biological processes listed, each set shows a distinct pattern of functional enrichment. It is also important to note that the set of co-expressed genes and the set of complements are disjoint, so the occurrence of one gene in both sets as a factor in overlapping of terms is impossible.

The set of co-expressed genes have eleven significant terms not found in the other set. Out of these eleven terms, seven terms that have been highlighted in the table are

Candidate Set:	Combined Intersection Set					Combined Non-Intersection Set				
Reference Set of Random Genes:	500	500	1000	1000	1000	500	500	1000	1000	1000
No Significant GO Terms										
1 mesoderm morphogenesis		X	X							
2 gland development			X	X						
3 bone remodeling			X	X						
4 determination of left/right symmetry	X	X	X	X	X					
5 central nervous system development	X	X	X	X	X					
6 compartment specification	X	X	X	X	X					
7 odontogenesis (sensu Vertebrata)	X	X	X	X	X					
8 formation of primary germ layer	X	X	X	X	X					
9 cell fate determination	X	X	X	X	X					
10 tissue morphogenesis	X	X	X	X	X					
11 cellular morphogenesis during differentiation		X	X	X	X					
12 BMP signaling pathway	X	X	X	X	X			X		
13 sensory organ development	X	X	X	X	X				X	
14 brain development		X	X	X	X					
15 epidermis morphogenesis	X			X	X					
16 mesoderm development		X	X		X					
17 epidermis development	X		X	X						
18 inner ear development	X		X	X						
19 blood vessel morphogenesis	X	X	X	X	X			X	X	
20 Notch signaling pathway			X		X					
21 positive regulation of cell differentiation	X		X	X						X
22 enzyme linked receptor protein signaling pathway	X	X	X	X	X		X	X		X
23 embryonic pattern specification	X	X	X	X	X		X	X		X
24 cell fate commitment	X	X	X	X	X	X	X	X		X
25 positive regulation of transcription	X	X	X	X	X	X	X		X	X
26 anatomical structure formation	X	X	X	X	X		X	X	X	X
27 tissue development	X	X	X	X	X		X	X	X	X
28 gastrulation	X	X	X	X	X		X	X	X	X
29 transmembrane receptor protein serine/threonine	X	X	X	X	X		X	X	X	X
30 respiratory tube development		X	X		X		X	X		
31 angiogenesis		X	X					X		
32 negative regulation of cell differentiation	X	X	X	X	X	X	X	X	X	X
33 embryonic morphogenesis	X	X	X	X	X	X	X	X	X	X
34 negative regulation of cellular process	X	X	X	X	X	X	X	X	X	X
35 regionalization	X	X	X	X	X	X	X	X	X	X
36 tube development	X	X	X	X	X	X	X	X	X	X
37 regulation of cell differentiation	X	X	X	X	X	X	X	X	X	X
38 branching morphogenesis of a tube	X	X	X	X	X	X	X	X	X	X
39 organ morphogenesis	X	X	X	X	X	X	X	X	X	X
40 anterior/posterior pattern formation	X	X	X	X	X	X	X	X	X	X
41 somitogenesis	X		X	X	X	X		X	X	X
42 blood vessel development	X	X		X	X		X	X	X	X
43 segmentation		X	X	X	X	X		X	X	X
44 cell development	X								X	
45 lung development			X				X			
46 tissue remodeling			X					X	X	X
47 heart development	X	X		X		X	X	X	X	X
48 RNA biosynthetic process		X			X	X	X	X	X	X
49 regulation of cell proliferation		X				X	X	X	X	
50 transcription, DNA-dependent		X				X	X	X	X	
51 skeletal development			X				X	X	X	X
52 regulation of MAPK activity							X	X		X
53 hormone metabolic process								X	X	X
54 embryonic heart tube development								X	X	X
55 induction								X	X	X
56 neurogenesis						X	X	X	X	
57 transforming growth factor beta receptor signaling						X	X	X		X
58 ureteric bud development						X	X	X		X
59 positive regulation of cell proliferation						X	X	X	X	
60 G-protein coupled receptor protein signaling path						X	X		X	X
61 morphogenesis of embryonic epithelium						X		X	X	X
62 cell-cell signaling during cell fate commitment							X	X	X	X
63 ureteric bud branching							X	X	X	X
64 regulation of transcription, DNA-dependent							X	X	X	X
65 ion transport						X	X	X	X	X
66 carbohydrate transport						X	X	X	X	X
67 embryonic development (sensu Vertebrata)						X	X	X	X	X
68 cell migration							X	X		
69 embryonic development (sensu Mammalia)						X			X	
70 mesenchymal cell development						X			X	
71 negative regulation of cell proliferation							X		X	
72 common-partner SMAD protein phosphorylation						X				X
73 generation of neurons						X				
74 generation of precursor metabolites and energy							X			
75 vasculogenesis									X	
76 negative regulation of MAPK activity									X	
77 embryonic limb morphogenesis									X	
78 establishment and/or maintenance of epithelial ce									X	
79 neural crest cell development									X	
80 vitamin A metabolic process									X	
81 embryonic organ development									X	
82 epithelial cell differentiation									X	
83 regulation of cell migration										X
84 negative regulation of transferase activity										X
85 heart looping										X

Figure 4.3: The functional profiles of the collective set of co-expressed genes (in red) and the complementary genes (in blue).

the strongest indicators that our co-expressed gene set possesses a different functional profile from exclusively expressed gene set. All seven significant terms occur in each of the comparison for the former set and not once in any of the latter set's comparison. Similarly, the exclusively expressed gene profile contain three such terms:

- Seven terms most likely to be enriched only for the co-expressed set:
  - Level 6:

    GO:0007417 central nervous system development

    GO:0001709 cell fate determination
  - Level 7:

    GO:0007417 central nervous system development

    GO:0001709 cell fate determination

    GO:0048729 tissue morphogenesis
  - Level 8:

    GO:0042475 odontogenesis (sensu Vertebrata)
- Three terms most likely to be enriched only for the complementary gene set:
  - Level 5:

    GO:0006811 ion transport

    GO:0008643 carbohydrate transport

    GO:0043009 embryonic development (sensu Vertebrata)

One inference we make here is: the more frequent a significant term appears across comparisons using different random sets, the more likely that the candidate set is enriched with the term. For example, the term *Skeletal Development* [GO:0001501] is only found significant in one of the comparisons for the co-expressed gene set. We attribute the single occurrence to the inherent vagaries of each random set, rather than any kind of indication that our co-expressed genes are related to skeletal development. There are other cases of such anomalies in both sets of genes. While the contribution of such single and dual occurrences contribute to form the overall functional profile of the set, such terms are not worthwhile investigating separately to see whether they have any relation to the mesenchymal-epithelial transformation. This is based on the reasoning that if a candidate set has a robust percentage of genes related to a significant

term, then it should be picked up by *Fatigo+* using the adjusted *p*-values every single time, regardless of the random set we use.

Another observation is that both sets of genes contain many of the same significant terms. There are 10 biological processes that are returned as significant in all five comparisons for each of the two candidate sets. Approximately 40% of the genes in both sets are annotated to *Organ Morphogenesis* [GO:0009887] . These, and several other overlapping terms with similar values give a strong indication that both sets of genes share some common functions.

With this hindsight, one explanation for why the comparisons using the co-expressed genes as candidates versus the complementary genes as references failed to return us any significant terms is, because both sets shared many of the same terms as proven here via their functional profiles. Following up on this line of reasoning, the aforementioned comparisons also failed to pick up any of the five terms listed here as 'very likely to be significant' because either (1) their significance in smaller sets went undetected by *Fatigo+* after the adjusted *p*-values were considered, or/and (2) the significance of these terms depend on what kind of set is used as reference.

4.2 Biological Interpretation

This section, which was contributed almost unedited by Dr. Jonathan Bard (Department of Biomedical Sciences, University of Edinburgh), shall highlight the biological implications of the results we obtained from our intersection experiments.

The original reason for doing the project was that there was not enough literature on mesenchyme-epithelial transitions available and there were no tools available to investigate the genes involved in this process. It was therefore not possible to assess what genes might emerge from an analysis of GXD prior to developing the computational tools. GXD, it should be said, is populated by assays on individual genes that have interested individual researchers, and not by high-throughput data; hence it has only sporadic coverage of transcriptomes, the sets of genes expressed by individual tissues at particular stages (N.B. All background material is in [59]).

As a general observation, the data in GXD on genes involved in the mesenchyme-epithelium transition turned out not to be as comprehensive as we had hoped. We had looked for:

1. Common genes expressed **before** the transition – *some were found.*
2. Common genes expressed **after** the transition – *only two were found.*
3. Genes that needed to be switched off before development could take place. *None were necessarily expected and* ***nothing of obvious interest emerged*** *from the analysis, but time has not permitted a full analysis of the results.*
4. New genes produced as a result of transcriptional activity on the stage before development – *there were none*.

The genes expressed just before the transition (1) fall into three interesting functional classes:

- *Signalling pathway proteins*: It looks as if common genes include those involved in the *BMP* (*SMADs*), *wnt* (*Dll*, *DACT1*, etc.) and *notch-delta* (*jagged*, *dll*, etc.) pathways. There are also slight indications that the *eph-ephrin* pathway may play a role. The *notch-delta* and *eph-ephrin* pathways require direct contact between cells (a property of epithelial cells), the *wnt* pathway is very short range

(a few cell diameters) while the *BMP* pathway is relatively long-range (tens of cell diameters).

- *Retinoic acid blocking protein*: It is noticeable that cellular retinoic binding protein 1 (*CRAPB1*) is expressed in paraxial mesoderm, nephrons, early heart and presumptive blood vessels. However, *CRABP1* is very widely expressed during early development. This protein seems to mop up retinoic acid, a potent activator, and stops it getting to the nucleus. One possibility is that retinoic acid may block the mesenchyme-epithelium transition, but it is probably better to view *CRABP1* as a housekeeping gene.

- *Transcription regulation proteins*: This is a particularly important class of protein as it is responsible for the production of new proteins and hence for generating change. These proteins are therefore well represented in the database. The subtraction analysis yielded the following Uniprot IDs as being commonly represented:

  - **Q543E8 = Meox1**

    Meox1 was expressed in the early heart, presumptive blood vessels, somites and mesonephric mesenchyme – its presence has not been looked for in the early kidney or other potential tissues undergoing an mesenchymal-epithelial transition.

  - **Q3UGA1 = Cited1** and **Q6PGA9 = Cited2**

    Cited 1, 2 may well be involved in transcriptional regulation – they are already known to be expressed in early heart, prevascular material and unsegmented paraxial mesoderm, but there was no entry for metanephric mesenchyme or mesonephros. A follow-up of this gene in PubMed showed that [60] say that *cited1* and *cited2* are both expressed in cells about to undergo an mesenchyme-epithelium conversion, but that their deletion has no effect, presumably because of redundancy.

  - **Q8CCU9 = Lhx1**

    LHX1 is known to be expressed in prevascular material, the early heart, the nephric duct and the mesonephros. It has not been reported either way for paraxial mesoderm, and is not present in the metanephric mesenchyme (MM), but is in the ureteric bud of the metanephros (the immediately adjacent tissue that interacts with the MM).

The results as a whole show that the methodology works in principal, but that there are too few genes currently in GXD for a full analysis to be done and that the analysis produced, for the first time, a set of candidate transcriptional activators for the mesenchyme-to-epithelium conversion.

Chapter 5

Discussion

The results returned in Sec. 4.1.3.4 show that genes co-expressed in three or more *Before* sets of tissues have a distinctly different functional profile from the rest of the genes found in these tissues. We established seven significant terms of the co-expressed set and three for the complement set that are the strongest indicators for their respective profiles. The universal characteristic of these terms is that they are all absent for either one of the sets but always present for the other set. Six of the terms from the co-expressed set are descendants of the node 'Developmental Process', whereas only one such term was found in the complement set.

5.1 Interpretation of Results

How these developmental processes are linked to the mesenchymal-epithelial process is unclear. In this section we highlight issues that would be worthwhile investigating further from a biological aspect. All percentages and p-values mentioned henceforth are averages taken across all five comparisons.

5.1.1 Set of Co-expressed Genes

Compartment Specification [GO:0007386] is annotated to roughly 8% of our co-expressed genes, the expected percentage from the random sets is less than 0.5%. GO defines the term as:

> The regionalization process by which embryonic segments are divided into compartments that will result in differences in cell differentiation.

The keyword 'regionalization' here is itself another significant term, i.e. *Regionalization* [GO:0003002], that was common to both the co-expressed and complement genes. Both terms are involved in cell differentiation, and defining the spaces or areas of cells where this takes place. Since most of the cells in our *Before* tissues are in the process of developing into more specialised tissues or organs, it is not surprising to see *Regionalization* annotated in both sets. For example, this process would take place in the yolk sac and mesoderm (two important tissues in our *Before* sets), where cells are differentiating into more specific types of cells. More interestingly, *Compartment Specification* is a grandchild of *Regionalization* linked by the `is-a` relation. Therefore the former process is a more specialised instance of the latter. There are several other less obvious cases where both sets are annotated to a more general process, whereas only one of them is involved in a lower level related process.

Before we proceed to discuss the implications of the *Compartment Specification* case, it is important to note that such occurrences are not because our sets of genes happen to have multiple terms referring to the same process. We took every precaution in our comparisons to ensure that such redundancy would not happen by taking only the most specific processes possible in our table of 'non-redundant' terms. Rather, this phenomenon happens only when one set is linked to a less specific term, while the other set is linked to a more specific term, in which case we have no choice but to include both terms even if they are parent-child nodes because in this case the more general term is not redundant since it indicates the lowest level of annotation for one set.

When a term occurs with a very high frequency in one set as is the case of *Compartment Specification*, there are several inferences one can make:

1. There is one general biological process that is happening in all the tissues.
2. There are some co-expressed and non-co-expressed genes responsible for this process (as shown by the functional profile table, cp. 4.3).
3. There is a more specific instance of this process that is happening in all the tissues where the genes are co-expressed.

4. Only co-expressed genes are responsible for this second process. (cp. 4.3)

In order to make these inferences, we assume that each set of intersected tissues has a roughly equal number of genes with the annotation 'Regionalization', this may not necessarily be the case since our findings were based on combined sets of genes from multiple tissues, thus we have no way of ascertaining which tissues the genes originate from. This would be one point worth investigating in the future work. However, if our assumption is true, then all four inferences are valid. We can then ask the question: Is it possible that these co-expressed genes could also be related to the mesenchymal-epithelial process, since we know that it is a biological process related to cell differentiation, and our analysis has already proven that our set of genes are responsible for one such process?

Formation of Primary Germ Layer [GO:0001704] is defined by GO as:

> The formation of the ectoderm, mesoderm and endoderm during gastrulation.

About 11% of the genes from our co-expressed set are linked to this term, whereas the overall expected value is smaller than 0.5%. We know that mesenchymal cells are a derivative of the mesoderm. At this point we do not know for sure whether the genes responsible for the entire mesenchymal-epithelial transformation is responsible for the entire process from start to beginning, or whether different genes are responsible for different parts of it. Again, there are several questions worthwhile investigating:

- Is it possible that with our co-expressed set we have managed to isolate the genes that are responsible for the first stage of the mesenchymal-epithelial transformation? Since none of the comparisons for the complement set returned this term, there is probably a common set of genes responsible for this process.
- On the other hand, since the mesenchymal-epithelial process has not been documented in GO, could the same set of genes be responsible for both functions?
- If there is the same set of genes responsible for this function are being co-expressed, and we have managed to capture this set in our intersections, is it likely that the same intersections have also managed to capture the set responsible for mesenchymal-epithelial transformation?

The same questions would also apply for *Cell Fate Determination* [GO:0001709]. Answering them would again require investigating the individual genes which were annotated to this process, and tracing the tissues from which they were expressed from our set of co-expressed genes. After doing this, the genes would have to be subject to laboratory tests or curation from genome sources other than the GO in order to answer the questions above.

5.1.2 Set of Complementary Genes

One important observation is that many of the significant terms found in the complement set contain the keyword (or derivatives of) 'epithelial': *Establishment and/or Maintenance of Epithelial Cell Polarity* [GO:0045197] , *Epithelial Cell Differentiation* [GO:0030855], *Morphogenesis of Embryonic Epithelium* [GO:0016331]. There is only one occurrence each for the first two terms, but *Morphogenesis of Embryonic Epithelium* was detected in four out of five comparisons. GO defines it as:

> The process by which the anatomical structures of embryonic epithelia are generated and organized. Morphogenesis pertains to the creation of form.

The high frequency of these terms is generally not encouraging news for our investigation, since we would have hoped to find they occur in our co-expressed set. Nevertheless we have to concede that in all cases, especially *Morphogenesis of Embryonic Epithelium*, it would seem that these functions are highly related to the mesenchymal-epithelial transformation. None of these terms were found across all five sets, although four occurrences for the last term is still a high number, the fact that none of the comparisons for the co-expressed set returned any of these genes is also indicative that more comparisons are likely to confirm that the set of complement genes are in fact more functionally related to the terms than the co-expressed genes.

5.1.3 Conclusion

In conclusion, our findings are subject to a wide variety of questions based on conjectures that need to be further investigated to provide any real answers. What we have established is that our co-expressed set of genes are functionally different from the genes which are not co-expressed. Amongst the functions exclusive to the co-expressed

genes, a number of them are specialised instances of developmental processes. On the other hand, functions that are exclusive only to the expressed genes include keywords that are highly indicative of their relation to the epithelial-mesenchymal transformation.

5.2 Analysis of Problems

In this section, we will try to shed some light on the reasons why we obtained such dissatisfying results. We hope to provide some insights that might help future research to be more successful than the one at hand.

5.2.1 Lack of Data

Throughout the comparisons, we noted that many results returned lay just below the accepted threshold for the adjusted p-value. Many annotations had values like 0.05 and 0.06 and are therefore rejected by *Fatigo+*. Unadjusted p-values were also far better than their FDR adjusted counterparts.

Since many of our intersections contained only one, two or three genes, it is highly unlikely that such sets will return any significant terms, regardless of the types of reference sets we use. In rare cases where we managed to find any significant terms, these same results were not reproducible using different random sets and therefore had to be rejected as well. There are several explanations for our lack of data. The most obvious reason is that the GXD tables we use do not contain high throughput data. Thus the number of genes we obtain even before intersections are very small to begin with. Annotation analysis tools like *Fatigo+* are mainly designed with high throughput data in mind, thus our candidate sets are not large enough to withstand the statistical tests of such tools.

A second contributing factor is due to the low number of tissues in some of our developmental processes. For example, there is only one *After* tissue in our set for Nephric Duct. Therefore some of our pre-intersection sets consist of only 10 genes or less. The size of intersection are dependent on the sizes of the sets involved. At best, an intersection can only return as many genes as there are in the smallest set.

5.2.2 Weaknesses of the Functional Profile Approach

Our findings are primarily based on the functional profiles of the co-expressed genes and the complement genes. There are several inherent disadvantages of this comparison. Firstly, the functional profiles do not tell us which tissues the genes are originating from. For instance, we only know that the nature of the sets are either 'co-expressed' or 'not co-expressed'. In the case of the former, we also combined intersections of three and four in the same list. Therefore, our comparison fails to make a distinction between co-expressed genes in three and four tissues respectively.

Secondly, the data in this kind of set can lead to misleading conclusions. For example, the comparison shows that 40% of the genes are related to *Organ Morphogenesis*. This does not necessarily mean 40% of genes in each and every different set of tissues have the annotation. It could be the case that the bulk of the 40% were contributed from one set. The problem is not so easily missed in the case of our co-expression sets, where we know that each gene must occur in three tissues at the very least. However it is especially hard to detect in the case of our complement sets, where each gene is expressed in at most two tissues. Thus it is difficult to make inferences that rely on such 'universal' judgements of the tissues. More investigation needs to be done in this area, which shall be discussed in the future work section.

5.2.3 Relying on the Gene Ontology

The biggest challenge we faced with trying to relate significant terms to the mesenchymal-epithelial process hinged on the fact that we were only relying on the Gene Ontology for information. As discussed in the Sec. 2.2, we know that the GO does not have any consistent documentation of how terms are derived and related to each other. Interesting annotations may have went undetected in the generic pool of non-outstanding terms.

Furthermore, it is difficult to pinpoint any one branch of biological process where our process of interest is most likely to occur, other than determining that any process falling under Level 4, *Developmental Process* is a 'preferable' indication. This problem is compounded by the fact that the process of interest is common to so many different types of tissues. As a result, our comparisons returned annotations situated in

a variety of branches that are related to different locations, such as the heart, kidney, more specific ones like blood vessels, and even highly unlikely places like the brain, teeth and skeleton.

At this point, one may pose the question: Is it correct to discount significant terms like *Skeletal Development* and *Lung Development* as having absolutely no relation to the set of genes we are interested in locating? Although one could argue that it is highly unlikely that the same gene responsible for transforming mesenchymal cells to epithelium could also be performing other biological processes that are so radically different, ultimately one cannot prove this unless supported by laboratory evidence. Apart from referring to the definitions of each term provided by the Gene Ontology, we did not have any other reliable source of data to guide us in this aspect.

5.3 Future Work

In this section we propose several tasks that can be undertaken to improve the validity of our results so that more robust interpretations and inferences can be made from our data.

5.3.1 Relation of Mesenchymal-Epithelial Transformation to the Significant Terms

First of all, we need to establish a systematic method for determining whether an existing GO term is related to the mesenchymal-epithelial transformation. At this point, we are relying on the occurrences of certain keywords, or a very rudimentary understanding of the process to infer any possible relations. Without any proper knowledge or research in the area, such a method is subject to a lot of speculation, and is clearly not suitable for objective evaluation.

Since the mesenchymal-epithelial transformation has not been documented *per se* in GO, and since the genes that are responsible for this process have not been identified, it would be highly unlikely that any annotation-based analysis that relies on the curation of literature would be helpful in this aspect. The most effective method would be a manual assessment of each term by an informed biological expert. If the

task of relating mesenchymal-epithelial function to existing GO terms is achievable, it would also be desirable to set a scale of such 'relatedness' so that further statistical tests can utilise this scale to rate the desirability of each term in relation to our process of interest.

5.3.2 De-Constructing the Functional Profile

Because we combined all our intersections for this comparison, certain inferences that were made in the discussion can only be validated by performing various traces on our results. In order to answer questions like 'Is it true that 40% of the genes expressed in every individual sets of tissues (i.e. Angiogenesis, Cardiac Endothelium, Somites etc.) are related *Organ Morphogenesis*?', we need to reverse engineer the procedures used to obtain our functional profiles so that we can trace the results of each comparison back to their source of data.

In order to do this, we need to systematically trace each significant term back to its source. The first step is to find, for each significant term, all the genes that are actually annotated with it. After this, for each gene, we record the set or sets of tissues where they were expressed. We can then combine these two aspects of information to give us a more accurate functional view of each intersection. It would be interesting to find out whether the functional profile of individual properties are correlated to the overall functional profile we already have. This is probably more likely for the larger intersections but may not always be the case for smaller intersections.

5.4 Critical Assessment

Statistically the hypothesis clearly needs to be rejected for each individual intersection as we did not obtain any significant terms in most cases. Our functional profiles, however, managed to prove that the genes obtained from our intersections are significantly different in overall function than genes not included in the intersections. A large portion of the differing functions are instances of developmental processes.

Overall, we failed to establish a strong relationship between the intersections and the mesenchymal-epithelial transformation. Although we managed to find several in-

teresting aspects in our results both from a computational analysis and a biological assessment, the evidence presented in both instances were not solid enough to validate our hypothesis.

This project has managed to raise several issues related to the mesenchymal-epithelial transformation that could be worth further investigation. Several candidate genes for the process were also identified.

Appendix A

The Mouse Genome Intersector Tool

In this chapter, we shall give a few more details about the program implemented. We start with giving UML class diagrams of the two packages, `Mouse` and `MouseGui`, that the tool consists of, continue with a list of SQL queries used by the program and conclude with a small selection of screenshots which provide a short walk-through of the program's graphical user interface (GUI).

A.1 Complete UML Class Diagrams

We have earlier presented a massively simplified version of the UML class diagram in Figure 3.3. For the sake of conciseness, this had been crucially truncated and, on the other hand, a few additional links had been invented to clarify the mutual dependencies of the individual classes. For completeness, we now present the entire class diagram of the `Mouse` package, as well as a more detailed version of the class diagram for the `MouseGui` package.

Figure A.1 shows the complete UML class diagram of the `Mouse` package of the software. This package holds all the underlying logic for the operations on gene sets performed using the GUI, i.e. it provides a container for storing distinct `Genes` (`Cassette`), a higher level container storing two lists of `Cassettes` for *Before* and *After* structures, namely `BeforeAfterCassette`, and another high level container, called `IntersectionCassette`, which is used to compute the intersections of (other) cassettes while remembering where the intersected genes came from. The `IntersectionSet` class is used to calculate all possible permutations of intersection of N tissues and store the results of each of them.

Figure A.2 shows a slightly simplified version of the UML class diagram of the `MouseGui` package. Unfortunately, due to the vast amount of variables and methods that happen to occur with GUI's, we again had to concede to leave out a few less important classes and methods. Consequently, some irrelevant classes (small dialogues, help, ...) and purely GUI-related methods and fields (buttons, labels, event handlers,

...) have been dropped for the sake of clarity. The whole GUI is built around the `MouseMainWindow` class, from which several other windows can be opened according to which state of the analysis process the program is in at the moment.

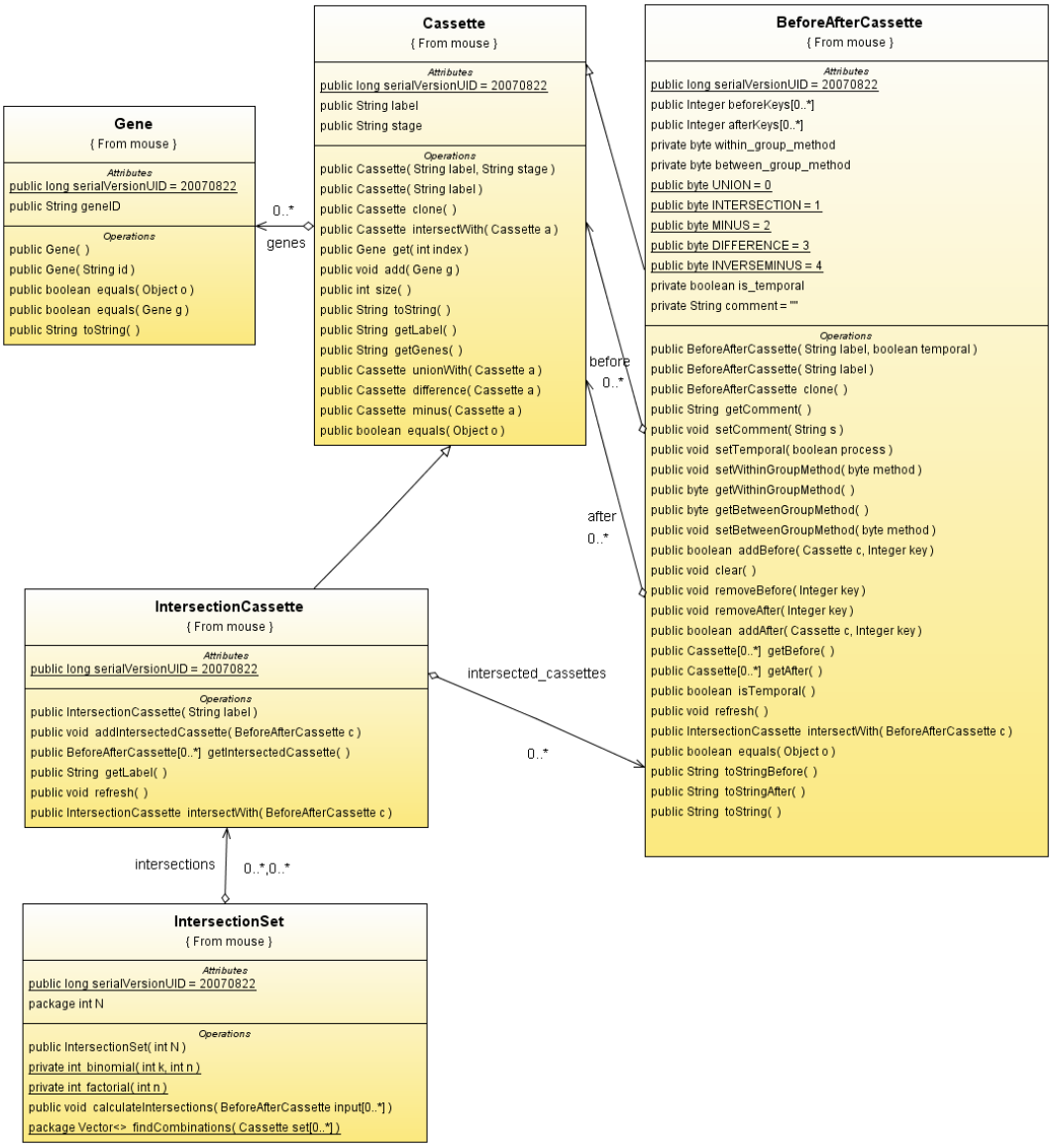


Figure A.1: UML Class Diagram of the `Mouse` package.

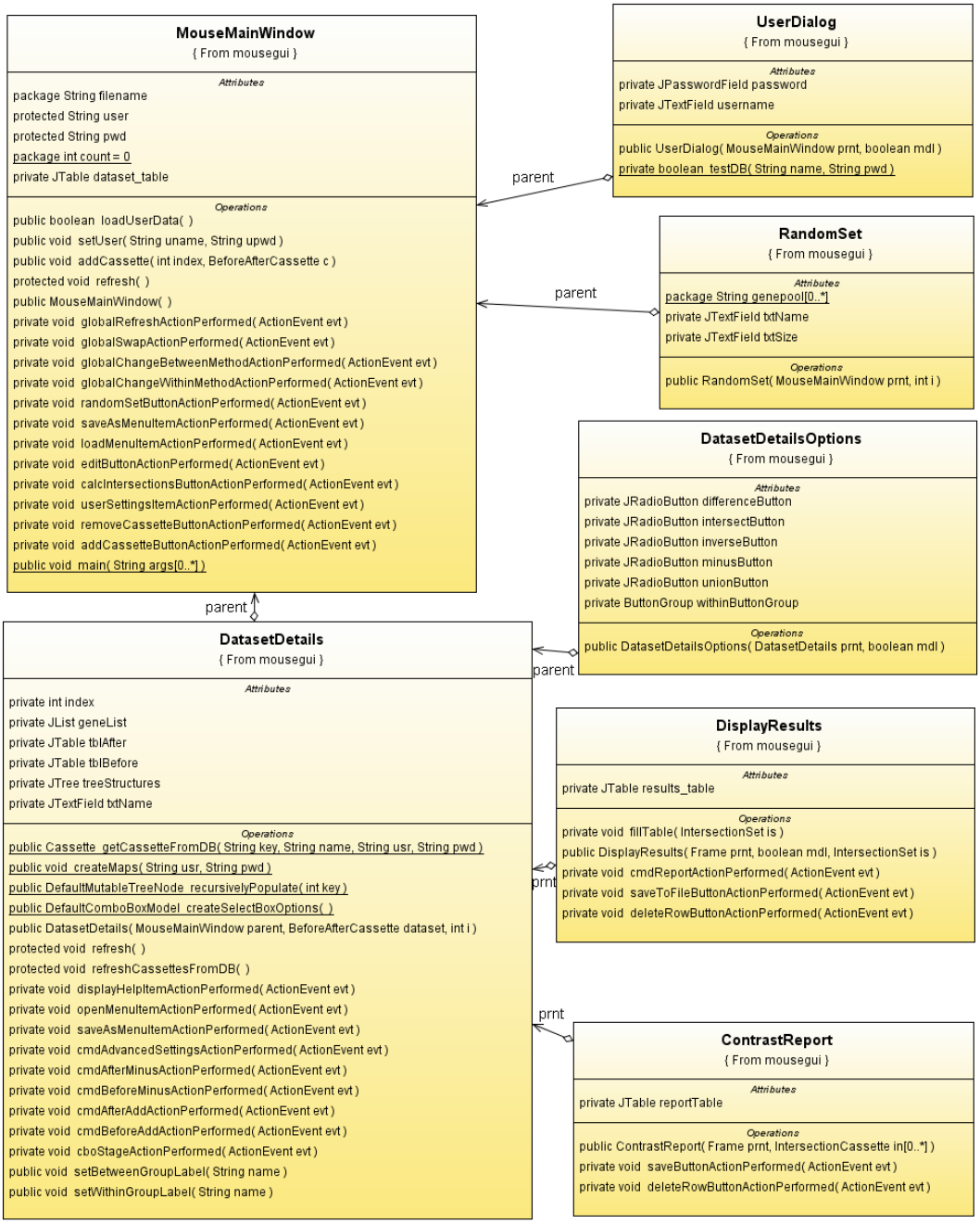


Figure A.2: Simplified UML Class Diagram of the `MouseGui` package.

A.2 SQL Queries

The whole program actually uses only three different queries for the interaction with the GXD database. The first one is used to create a map of child/parent relations between biological structures in the database, for quicker accessibility of the tree browsing feature later on:

```
SELECT _Structure_key, _Parent_key, printName, edinburghKey
FROM GXD_Structure
```

The second query is employed to select the root node of the tree corresponding to the stage which has been selected in the GUI (the question mark (`?`) in the code will be dynamically replaced with the key of the stage in question):

```
SELECT _Structure_key
FROM GXD_Structure
WHERE _Stage_key = ? AND _Parent_key IS NULL
```

The third query is used to actually retrieve the the genes expressed in a certain structure. Note that – again – the question mark (`?`) will be replaced dynamically by the key of the structure in question. Furthermore, this query will be recursively repeated for all the structure's children in order to find all genes expressed (some genes have only been identified in substructures, but are hence also expressed in the parent structures).

```
SELECT MAX(a.accID)
FROM
     MRK_Mouse_View m,
     ACC_Accession a,
     GXD_Expression e,
     GXD_Structure s
WHERE
     m._Marker_key = a._Object_key AND
     a._MGIType_key = 2 AND
     a._LogicalDB_key in (13, 41, 1) AND
     a.preferred = 1 AND
     a._Object_key = e._Marker_key AND
     e.expressed = 1 AND
     e._Structure_key = s._Structure_key AND
     s._Structure_key = ? AND
     NOT EXISTS
     (SELECT 1
      FROM GXD_AlleleGenotype g
      WHERE e._Genotype_key = g._Genotype_key
     )
GROUP BY a._Object_key
```

A.3 Screenshots

The following screenshots are taken at several stages during a typical analysis run using the *Mouse Genome Intersector* tool. The fully functional tool as well as the source code is available from the author on request.

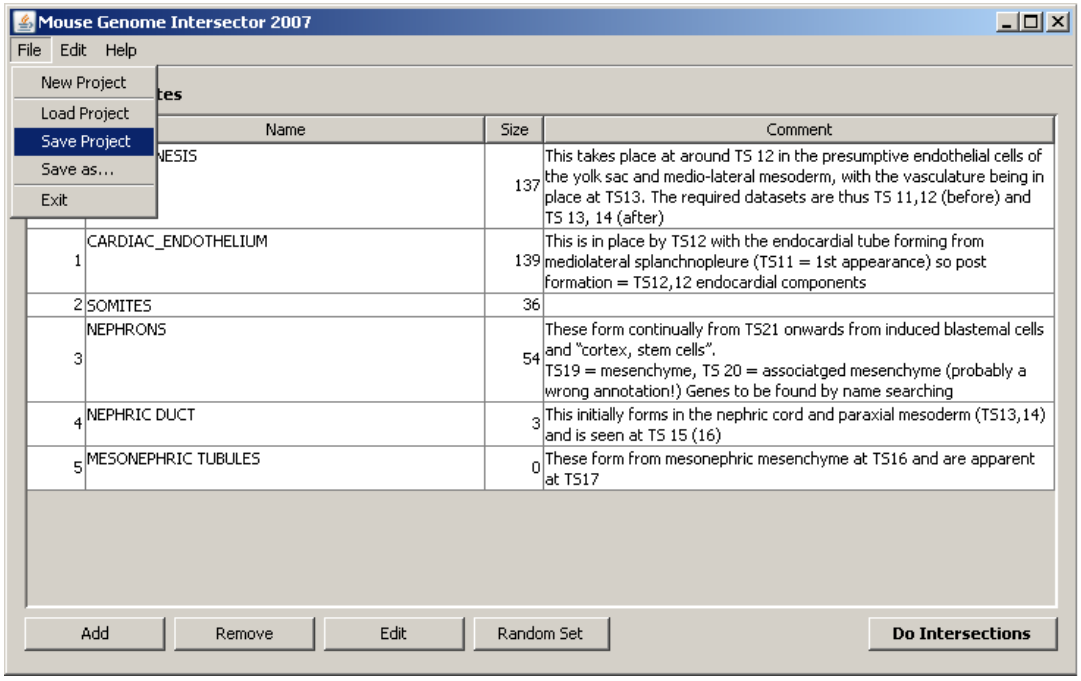


Figure A.3: `MainWindow`: Once the user starts the program he will be brought to this window. The user can now choose between loading an existing project or creating a new one by adding adding tissues to the empty project. Of course, he may save his progress at any stage. The user can change the properties of all the individual datasets on a global level by choosing the methods used for intersecting gene cassettes and applying the changes to all tissues in the current project. It is also possible to update the structures and genes in the project with most up-to-date values from the GXD database by just a few clicks.

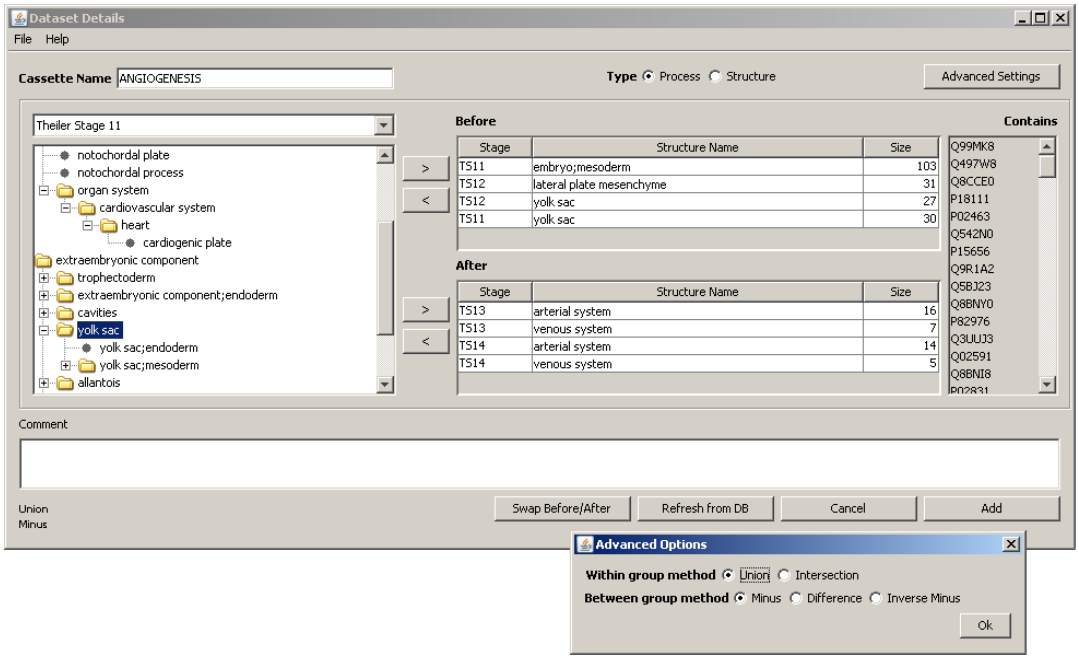


Figure A.4: `DatasetDetails` and `DatasetDetailsOptions`: This window represents the basic interface for adding new or modifying existing tissues. The user can browse the EMAP ontology for structure terms and add the corresponding gene cassettes to either the *Before* or *After* tissues. Alternatively, the user may choose to enter only one list of cassettes by setting the BeforeAfterCassette to be a 'structure' (rather than a 'process'). It is also possible to change the intersection methods used for this dataset in an advanced options dialogue. Note, how the currently selected methods are displayed in the lower left part of the window. Worthwhile mentioning, the gene list on the right is updated on-the-fly with every cassette or option changed in the process.

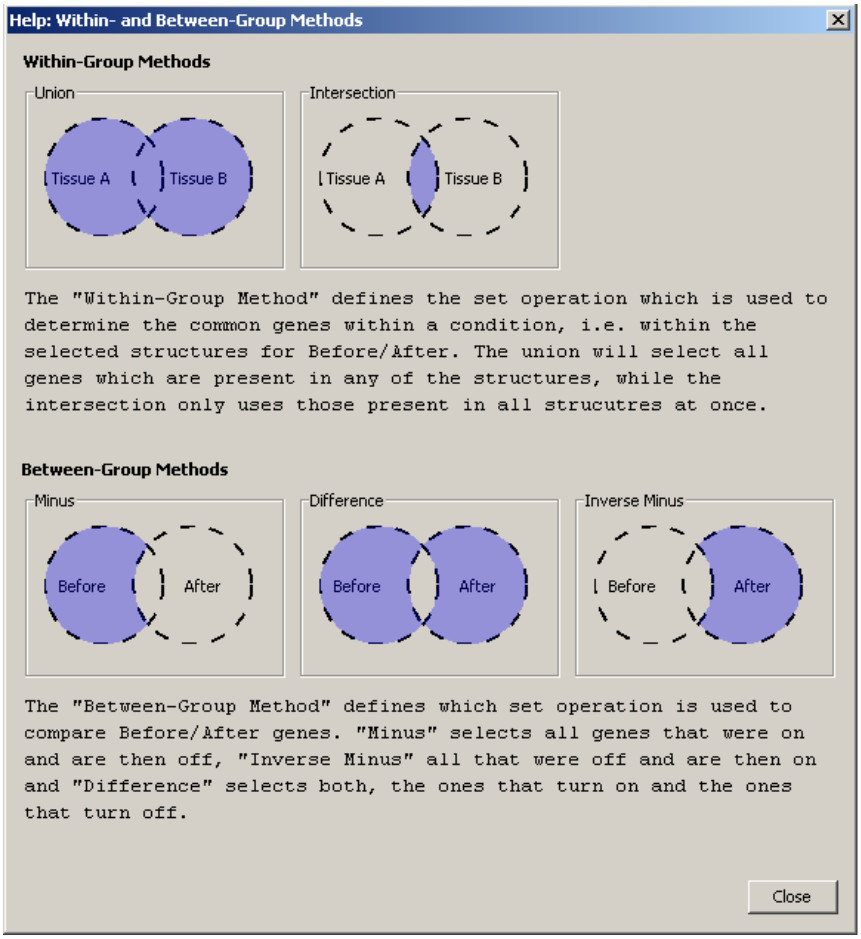


Figure A.5: `DatasetDetailsHelp`: Since we realized that the different options for Within- and Between-Group methods can be rather confusing, we also added a small help dialogue which summarizes the differences between the different options. This dialogue can be accessed from the `DatasetDetails` window (cp. Fig. A.4).

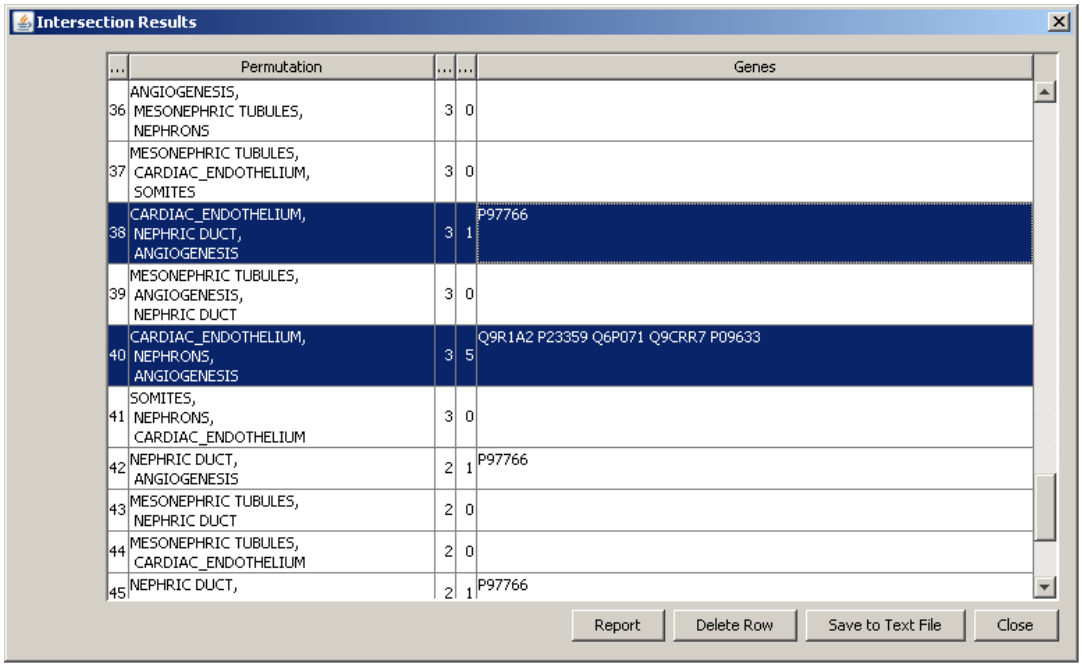


Figure A.6: `DisplayResults`: Once the user has added all datasets his is interested in, he can click the button labelled 'Do Intersections' in the main window, to be forwarded to this window. In a table, the results of all possible permutations of intersections between the selected datasets are displayed. It is possible to sort the results according to different criteria (arguably most useful, by the number of tissues intersected or by the number of genes in the intersection). The user can select to save individual results to text files, remove uninteresting rows from the table or to display a special report form contrasting the results from a number of selected rows (cp. Fig. A.7).

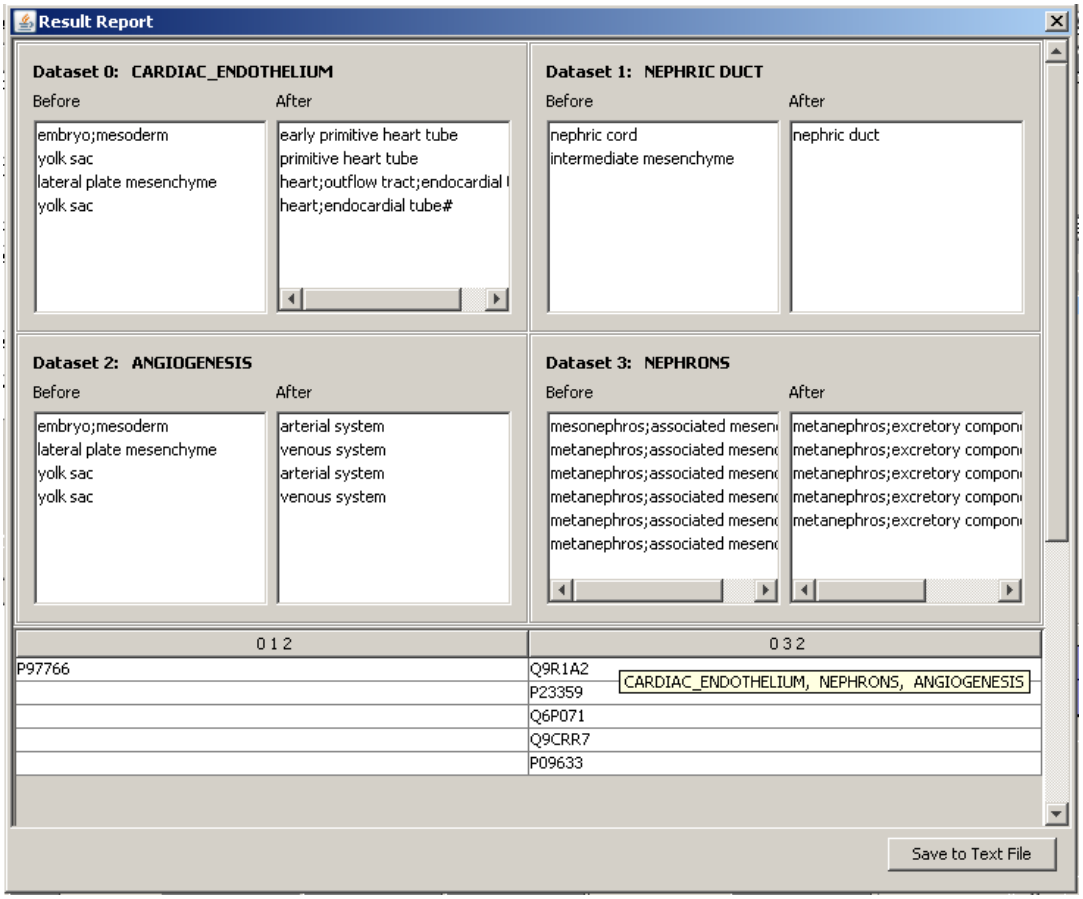


Figure A.7: `ContrastReport` (incl. `ContrastDataset`): This window provides a concise overview over a number of selected results. The datasets involved are briefly summarized in the upper part of the window and below we find a table with each columns corresponding to one intersection and each row corresponding to one gene. This report can also be saved to a text file.

Appendix B

Full List of Experimental Results

In this chapter, a comprehensive overview about all genes found in the different stages of our analysis shall be given. The interested reader may study the lists himself and draw his own conclusions. In summary, we have studied genes in four different settings:

- Genes that are expressed in *Before* structures (Tab. B.1).
- Genes that are expressed in *After* structures (Tab. B.3).
- Genes that are turned off during the process, i.e. that are expressed in *Before*, but not expressed in *After* (Tab. B.5).
- Genes that are turned on during the process, i.e. that are not expressed in *Before*, but expressed in *After* (Tab. B.7).

In each of these scenarios, we then looked at all possible intersections of three or more of the gene lists and recorded all non-empty gene sets found (Tbl. B.2, B.4, B.6). There were no genes found in any of these intersections for genes that turn on during the process!

For the sake of conciseness, we will abbreviate the full tissue names in the following in some places. The abbreviations used are: **A** = Angiogenesis, **C** = Cardiac Endothelium, **M** = Mesonephric Tubules, **N** = Nephrons, **Nd** = Nephric Duct and **S** = Somites.

Table B.1: This table lists all genes expressed in the *Before* structures of the individual tissues. Note that *Angiogenesis* and *Cardiac Endothelium* share the same *Before* structures and hence express exactly the same genes (yet the order is different due to random processes involved in data retrieval). Furthermore, genes which have been tagged as 'unknown' by *Fatigo+* have been marked with an asterisk.

Angiogenesis	**Cardiac Endothelium**	**Somites**	**Nephrons**	**Nephric Duct**	**Mesonephric Tubules**
Q99MK8	Q99MK8	Q80X37	P19091	Q04744	Q8CCU9
Q497W8	Q497W8	Q6P071	Q9QWQ1	P48540	Q543E8
Q3ULR1	Q3ULR1	P22935	Q9QWR9	Q8BM75	Q3UY25
Q8CCE0	Q8CCE0	Q9R1A2	Q61850	Q8CCU9	Q9WVG7
P18111	P18111	Q8BNI8	Q3UQJ4	P97766	Q9QWR9
P02463	P02463	Q9R2A7*	Q3TWK8		Q61850
Q542N0	Q542N0	Q8K428	P31310		Q8BSP4
P15656	P15656	Q4FK48	P31311		Q8R4A3
Q9R1A2	Q9R1A2	Q5SDA2	Q496Q8*		Q8BM75
Q5BJ23	Q5BJ23	Q543E8	Q8BZY5		
Q8BNY0	Q8BNY0	Q6PFV7	P31313		
Q8VI87	Q8VI87	Q8BSB3	Q08624		
P82976	P82976	Q08EF2	P32043		
Q3UUJ3	Q3UUJ3	Q99KA8	Q543H4		
Q02591	Q02591	Q8C5P2	Q8BQA3		
Q9R1X2	Q9R1X2	Q9CRD0	P09633		
Q8BNI8	Q8BNI8	Q9EPZ6	P28359		
P02831	P02831	Q80U19	P23813		
Q61681	Q61681	Q8BZA0	Q8BSN0		
Q3ZAX9*	Q3ZAX9*	Q58EU7	Q8VHP0		
P09632	P09632	Q9WUL2*	Q3U1N3		
Q8BV11*	Q8BV11*	Q60756	Q62438		
Q3UJB6	Q3UJB6	Q499J8	Q8C2P1		
Q8VCD0	Q8VCD0	P70658	Q8CC31		
Q9CRX6	Q9CRX6	Q6R5E9	O55222		
Q6P8P3	Q6P8P3	O55233	Q9Z0Y6		
Q9EQ12	Q9EQ12	Q6PGA9	Q9ER74		
Q9R2A7*	Q9R2A7*	Q3UGA1	Q9CTF6		
Q925V3	Q925V3	Q61553	Q920C1		
Q8K428	Q8K428	Q8R4A3	Q9R0R2		
Q8VD35	Q8VD35	Q9DBB1	Q3KQI1*		
Q9CU96	Q9CU96	Q8BKG3	Q6P071		
Q8CGH8	Q8CGH8	Q9CWM2	Q9WVF5		
Q9CY80	Q9CY80	Q7TS73	Q9R1A2		
Q8C765	Q8C765	Q8CAT6	Q9CZD6		
Q4FK48	Q4FK48	Q3ZAX9*	Q5BLJ8		
Q78ZW9	Q78ZW9	Q8BV11*	Q91ZN8		
Q7TQ06	Q7TQ06	Q6GTZ3	Q199A7		
P26687	P26687	Q9Z197	Q80ZS9		
Q8VCV6	Q8VCV6	Q91YX2	Q62219		
Q3UMZ6	Q3UMZ6	Q922Z8	P23359		
Q80ZL6	Q80ZL6	Q3UMZ6	Q8VIK0		
Q8CCU9	Q8CCU9	Q8BSP4	P48540		
P19137	P19137	Q3V1C5	Q9Z1W4		
Q8K1X3*	Q8K1X3*	Q68EF7	Q80UW0		
P20263*	P20263*	Q9QZX5*	Q925H1		
Q5SDA2	Q5SDA2	Q9D7K8	Q6AZB0		
Q543E8	Q543E8	Q9CXC9	Q6PCX9		
P23359	P23359	Q62392	Q6PAS4		
Q5SRD8	Q5SRD8	Q9Z2C5	Q8CD68		
Q9WUL2*	Q9WUL2*	Q6PEB3	Q3ULR1		
Q80TC1	Q80TC1	Q6PFZ9	Q3UNK5		
Q6PFV7	Q6PFV7	Q80UL5	Q6PFV7		
Q60756	Q60756	Q9QX13	Q811W8		
Q9DCA0	Q9DCA0	Q9WV93	O35253		
P51655	P51655	Q9QXV8	Q8BUN5		
Q8BSB3	Q8BSB3	Q9JLL3	Q9CRR7*		
Q3UQH0	Q3UQH0	Q9JLF7*	Q80ZV9		
Q923Z1	Q923Z1	Q9R205	Q9JIW5		
Q62318	Q62318	Q9JHX2	Q8CDB8		
P97766	P97766	Q9EQW1	Q91ZD6		
Q8CC31	Q8CC31	Q8BHS3	Q99LS8		
P70658	P70658	Q9CZM1*	Q99J48		

Table B.1: Expressed genes in *Before* structures (cont.)

Angiogenesis	**Cardiac Endothelium**	**Somites**	**Nephrons**	**Nephric Duct**	**Mesonephric Tubules**
Q8BTM5	Q8BTM5	Q8R3I2	O09009		
Q9QXX0	Q9QXX0	Q8K0C8	Q3UPI0		
Q8R381	Q8R381	Q99J68			
Q9D7K8	Q9D7K8	Q9CYI8			
Q99KA8	Q99KA8	Q91ZJ5			
Q91VZ3*	Q91VZ3*	Q91WG3			
O55127	O55127	Q9QX46			
Q9CXC9	Q9CXC9	Q8BQI5			
Q3UND5	Q3UND5	Q99K36			
Q80UL7	Q80UL7	Q8BNY0			
O55233	O55233	P47856			
Q9JJL1	Q9JJL1	Q4L141*			
Q9EQM2	Q9EQM2	Q8BS81			
Q6PFZ9	Q6PFZ9	Q99LA0			
Q544L3	Q544L3	Q9JL41*			
Q8R5G0	Q8R5G0	P49817			
Q6PGA9	Q6PGA9	Q9DCR3			
Q9Z0E2	Q9Z0E2	Q8C313			
Q9R1X4	Q9R1X4	Q80WX0*			
Q80UL5	Q80UL5	Q9CU49			
Q3UGA1	Q3UGA1	Q3UUM9			
Q61271	Q61271	Q62318			
Q9WVC6	Q9WVC6	Q8C4U3			
Q8K0H5	Q8K0H5	P97401			
Q60688	Q60688	Q80UF3			
Q61583	Q61583	Q71V68			
Q9QXP9	Q9QXP9	Q9QXX0			
Q3V1F2	Q3V1F2	Q544L9			
Q9WUI0	Q9WUI0	O08574			
Q9QXN0	Q9QXN0	Q3UND5			
Q9Z0Z7	Q9Z0Z7	Q922L1			
Q8R357	Q8R357	Q8BSU4			
Q9JI57	Q9JI57	Q9QZR5			
Q8R4A3	Q8R4A3	Q9WVM0			
Q9JHX2	Q9JHX2	Q8VCN8*			
Q8R044*	Q8R044*	Q9QXV9			
Q80U19	Q80U19	Q9WTP2			
Q91XQ5	Q91XQ5	Q9R001			
Q9ESD2	Q9ESD2	Q9QWR9			
Q8QZV2	Q8QZV2	Q61850			
Q6P071	Q61639	Q9Z138			
Q8QZY0	P09631	Q9ER74			
Q91YX2	P09633	Q9WTK0			
Q8BS64	Q545I7	Q9DC72			
Q5U3K8	Q5EEX1	Q9D1X9			
Q64280	Q9QUM0	Q8BKT2			
Q8C5P2	Q6LEB3	Q91YE5			
Q61080	Q8CCN5	Q80SY4			
Q9ES03	Q8BLF7	Q7TMY7			
P57785	P14246	Q811G8			
Q921T1	Q9CRR7*	Q80TF3			
Q544Z2	Q9QXT5	Q99LW6			
Q9D2A8	Q922E0	Q8CGH8			
Q3U223	Q544Z2	O55003			
Q9JKQ8	Q3ULW0	Q5SQB3			
Q60636	Q920W1	Q8BSS2			
Q7TMX8	Q9D8L6				
P17439	Q6P071				
Q9R1L3	Q8QZY0				
Q9CZK5	Q91YX2				
Q9R2A1	Q8BS64				
Q9QUM0	Q5U3K8				
Q5SQP3	Q64280				
Q925F5*	Q8C5P2				
Q922E0	Q61080				
Q8VIL9*	Q9ES03				
Q9JK33	P57785				

Table B.1: Expressed genes in *Before* structures (cont.)

Angiogenesis	**Cardiac Endothelium**	**Somites**	**Nephrons**	**Nephric Duct**	**Mesonephric Tubules**
Q543R9	Q921T1				
MGI:106910*	Q9D2A8				
Q9CRR7*	Q3U223				
Q9D5V4	Q9JKQ8				
Q91X98	Q60636				
Q9Z1Z8	Q7TMX8				
Q3KP84	P17439				
Q61639	Q9R1L3				
P09631	Q9CZK5				
P09633	Q9R2A1				
Q545I7	Q5SQP3				
Q5EEX1	Q925F5*				
Q6LEB3	Q8VIL9*				
Q8CCN5	Q9JK33				
Q8BLF7	Q543R9				
P14246	MGI:106910*				
Q9QXT5	Q9D5V4				
Q3ULW0	Q91X98				
Q920W1	Q9Z1Z8				
Q9D8L6	Q3KP84				
150	**150**	**119**	**65**	**5**	**9**

Table B.2: The table lists all possible intersections of three or more of the gene lists in Tbl. B.1. Note, that the list of genes for *Angiogenesis* (A) and *Cardiac Endothelium* (C) in the *Before* structures was identical, hence many of the intersections were redundant and have been left out in this table. Furthermore, genes which have been tagged as 'unknown' by *Fatigo+* have been marked with an asterisk.

$A \cap C \cap M \cap S$	$A \cap C \cap M \cap Nd$	$A \cap C \cap N \cap S$	$A \cap C \cap Nd$	$A \cap C \cap S$	$A \cap C \cap M$	$A \cap C \cap N$	$M \cap N \cap S$
Q543E8	Q8CCU9	Q6P071	Q8CCU9	Q6P071	Q8CCU9	P09633	Q9QWR9
Q8R4A3		Q9R1A2	P97766	Q9R1A2	Q543E8	Q8CC31	Q61850
		Q6PFV7		Q8BNI8	Q8R4A3	Q6P071	
				Q9R2A7*		Q9R1A2	
				Q8K428		P23359	
				Q4FK48		Q3ULR1	
				Q5SDA2		Q6PFV7	
				Q543E8		Q9CRR7*	
				Q6PFV7			
				Q8BSB3			
				Q99KA8			
				Q8C5P2			
				Q80U19			
				Q9WUL2*			
				Q60756			
				P70658			
				O55233			
				Q6PGA9			
				Q3UGA1			
				Q8R4A3			
				Q3ZAX9*			
				Q8BV11*			
				Q91YX2			
				Q3UMZ6			
				Q9D7K8			
				Q9CXC9			
				Q6PFZ9			
				Q80UL5			
				Q9JHX2			
				Q8BNY0			
				Q62318			
				Q9QXX0			
				Q3UND5			

Table B.2: Intersections of expressed genes in *Before* structures (cont.)

$A \cap C \cap M \cap S$	$A \cap C \cap M \cap Nd$	$A \cap C \cap N \cap S$	$A \cap C \cap Nd$	$A \cap C \cap S$	$A \cap C \cap M$	$A \cap C \cap N$	$M \cap N \cap S$
				Q8CGH8			
2	**1**	**3**	**2**	**34**	**3**	**8**	**2**

Table B.3: This table lists all genes expressed in the *After* structures of the individual tissues. Furthermore, genes which have been tagged as 'unknown' by *Fatigo+* have been marked with an asterisk.

Angiogenesis	**Cardiac Endothelium**	**Somites**	**Nephrons**	**Nephric Duct**	**Mesonephric Tubules**
Q544Z2	Q8VI87	Q80X37	P31310	Q04744	P18111
Q60753	Q921T1	P18111	P31311	Q8BNI8	Q8BM75
Q61614	Q9QVP4	Q5BJ23	P09631	Q3ZAX9*	
Q8C2P1	Q9R074	Q9Z1Z8	P09023	Q8VHP0	
Q61824	Q8BS64	Q9CT20	P32043	Q8C8Q7	
Q8C6E4	Q80UL7	Q8BNI8	P28359	Q8CCU9	
Q3UQJ4	Q9CWL2	Q9R2A7*	P10628	Q3UTY8	
Q9Z1Z8	Q3UGA1	Q9CU96	Q3UMQ3	Q9QWR9	
Q71V84	Q8CJ69	Q8CGH8	Q8R1P3	Q9DCI0	
Q8VCD0	Q3UNK5	Q3UTY8	Q3V0Z9	Q80TF3	
Q922E0	Q925V3	Q544Z2	Q8C2P1		
Q9D7K8	Q8BLK4	Q543E8	Q8CC31		
Q9CRD0	Q61272	Q60756	Q923S6		
Q9QXT5	Q9CXK3	Q8R381	Q9ER74		
Q9R1X2	Q8VCD0	Q9WV08	Q8K428		
Q91XQ5	Q3KP84	Q9QXN0	Q6PCM9		
Q8VI87	P19123	Q61553	Q99L24		
Q80UL5	Q6ZWX2	Q9ER74	Q6PFV7		
Q9WV08	Q8VHX6	Q9JHX2	Q7TQI8		
Q8BM75	Q9ES03	Q80U19	Q08EF2		
Q8R4A3	Q99MV5	Q9R1A2	Q9QXX0		
Q0PHV7*	Q99K17	Q8BNY0	Q9WV93		
Q3U223	Q922E0	P47806	Q9DBX7		
Q9EPN2	Q3UE22	Q3UJB6	Q9D1D6		
Q6GUA3	Q9D7K8	Q923F4	Q6PCX9		
Q3ULR1	MGI:1344335*	Q8K428	Q61045		
Q8BHZ7	Q9D2T3	Q8K4Q2	Q199A7		
	Q99KE3	Q9CZK7	Q91ZD6		
	Q921D9	Q4FK48	Q9D586		
		Q921T1	Q99MU3		
		Q80TC1	Q9QZM3		
		Q6PFV7	Q920W1		
		Q61115	Q62433		
		Q9D2N4	O35437		
		Q7TNN4			
		Q3UZ96			
		Q6R5E9			
		Q3UGA1			
		Q3UY25			
		Q9QWR9			
		Q61850			
		Q3V1F2			
		Q80UW0			
		Q9JI57			
		Q8BTF1			
		Q8BKG3			
		Q8BM75			
		Q08EF2			
		Q9EPZ6			
		Q8BYL5			
		P02831			
		Q9CXK3			
		Q8CCE0			
		Q6P071			

Table B.3: Genes expressed in *After* structures (cont.)

Angiogenesis	Cardiac Endothelium	Somites	Nephrons	Nephric Duct	Mesonephric Tubules
		Q9ESV5*			
		P12979			
		Q5BLJ8			
		Q8CFN5*			
		Q8BSB3			
		Q78DU3			
		O35392			
		MGI:2443183			
		Q496Q8*			
		Q58EU7			
		Q9WUL2*			
		Q8C2P1			
		O55233			
		Q6PGA9			
		Q3ULW0			
		Q61477			
		Q9JII6			
		Q8CD62			
		Q8CF69			
		Q3ZAX9*			
		Q61685			
		Q6TDG5*			
		Q80ZL6			
		Q8BSP4			
		Q9QXA3			
		Q9JLL0			
		Q8R4A3			
		Q9EQW1			
		Q8BHZ7			
		Q91XQ5			
		Q5SYC7			
		Q8BZA0			
		Q0PHV7*			
		Q61272			
		Q60636			
		Q9D748			
		Q3URE6			
		Q3UBP0			
		Q9QX46			
		Q8BSI9			
		P39061			
		Q99LW4			
		Q3USF7			
		Q99K36			
		Q542N0			
		Q8C399			
		P47856			
		P10284			
		P09024			
		Q8BQ25			
		Q4L141*			
		Q6Q533			
		Q8VHZ2			
		Q7TPG3			
		Q8BS81			
		Q99LA0			
		Q9Z2D6			
		Q9JL41*			
		P49817			
		Q62219			
		Q9DCR3			
		Q8C313			
		Q3V1C5			
		P52955			
		Q499J8			
		Q5SSV4			
		Q60753			

Table B.3: Genes expressed in *After* structures (cont.)

Angiogenesis	Cardiac Endothelium	Somites	Nephrons	Nephric Duct	Mesonephric Tubules
		Q80WX0*			
		Q9CU49			
		Q62318			
		Q99KF3			
		P97401			
		Q80UF3			
		Q71V68			
		Q9QXX0			
		O08574			
		Q3UND5			
		Q922L1			
		O35085			
		Q3TV22			
		Q9QZR5			
		Q9D8U0			
		Q545G3			
		Q9WVM0			
		Q8VCN8*			
		Q3U8E7			
		Q8C5P2			
		Q9WTS6			
		Q9CS91			
		Q3UMW5			
		Q4VBC9			
		Q9JLL3			
		Q9R0R4			
		Q9WUZ7			
		Q9WTK0			
		Q9CS81			
		Q9DBB1			
		Q9CRD0			
		Q9D1D6			
		Q7TQ32			
		Q7TT16			
		Q9EPK5			
		Q8BGT1			
		Q9D8G2			
		Q9DC72			
		Q8BKT2			
		Q8R3I6			
		Q9ESD2			
		Q91YE5			
		Q6AZB0			
		Q8CCN5			
		Q6T264			
		Q8R0M1			
		Q7TMY7			
		Q8QZV2			
		Q811G8			
		Q68BG2			
		MGI:2183691*			
		P10628			
		P09632			
		Q8BQA3			
		Q9CUQ8			
		Q3TYD9			
		Q80YP5			
		Q8CC06			
		Q545G1			
		Q8C6B1			
		Q61045			
		Q3UUM9			
		Q8C4U3			
		Q640N1			
		O08665			
		Q9JHE6			
		Q6NZL8			

Table B.3: Genes expressed in *After* structures (cont.)

Angiogenesis	Cardiac Endothelium	Somites	Nephrons	Nephric Duct	Mesonephric Tubules
		Q8BIA3			
		Q66PY1			
		P28481			
		Q99M23			
		Q9CXC9			
		Q8BG74			
		Q9ERH7			
		Q3U223			
		Q9Z0F1			
		P31311			
		Q8BV11*			
		P51655			
		Q8QZY9			
		Q9QXV8			
		Q9WVF2			
		Q811E4			
		Q8JZL1			
		Q3TZM2			
		Q9ERF6			
		Q8BYL6			
		Q99LW6			
		Q8BQ62			
		Q9JJY2			
		Q6DI71*			
		Q543H4			
		P09633			
		Q3UMQ3			
		Q05DD2*			
		Q3UVH8			
		Q6P6L3			
		P54823			
		Q61809			
		Q8C2A8			
		Q9EQW7			
		Q80VZ1			
		Q3UZB9			
		Q9JLL4*			
		O88876			
		Q62433			
		Q9WTP2			
		Q9Z1S5			
		MGI:1861674*			
		Q9QZ26			
		Q99N11			
		Q9JM62			
		Q9CWM2			
		Q8CFG0			
		Q8K1S7			
		Q99N43			
		Q8K350			
		Q8C4C4			
		Q9DBJ9			
		Q3TZE7			
		P22935			
		P82976			
		Q8C9K1			
		Q9JL91			
		Q99LS8			
		Q9D6N4			
		P26687			
		Q9D080			
		Q8R228			
		Q8K3G6			
		Q76LY2			
		Q8VHX6			
		Q3ZAZ1			
		Q61824			

Table B.3: Genes expressed in *After* structures (cont.)

Angiogenesis	Cardiac Endothelium	Somites	Nephrons	Nephric Duct	Mesonephric Tubules
		Q6NXW7			
		Q99LQ5			
		Q9Z139			
		Q9Z138			
		Q9R0G8			
		Q9JLI1			
		Q9JJZ5			
		Q9ERP3			
		Q99PV5			
		Q571J2			
		Q6P5N9			
		Q61080			
		Q8CJ69			
		Q8BYQ8			
		Q8BMD9			
		Q8BLU0			
		Q3UYE4			
		Q9JI91			
		Q99N13			
		Q78TF3			
		Q60932			
		Q9QZX5*			
		Q9Z0Z7			
		Q9CZP5			
		Q8K2P6			
		Q61321			
		Q9QX23			
		Q9DC41			
		Q8VHX3*			
		Q0VEJ7			
		Q9CZ19			
		Q6PCW9*			
		Q91YX2			
		Q05DE6*			
		Q8BRJ5			
		Q8C6B5			
		Q62232			
		Q8C766			
		Q9Z2G6			
		Q8VHZ3			
		Q9QXV9			
		Q8BSS2			
		Q8CFR7			
		Q9D677			
		Q8BZ84			
		Q9EPR5			
		Q9JKB3			
		Q8R419			
		Q8CJG1			
		Q8R3Q7			
		Q8CJF9			
		MGI:2676881*			
		Q3KP84			
		Q8R145			
		Q9CUZ5			
		Q9D030			
		Q80W09			
		Q80UF6			
		Q9WUU4			
		Q9JKV7			
		Q8CGN4*			
		P23813			
		P70217			
		P70323			
		Q810F8			
		Q8BSF9			
		Q921F1			

Table B.3: Genes expressed in *After* structures (cont.)

Angiogenesis	**Cardiac Endothelium**	**Somites**	**Nephrons**	**Nephric Duct**	**Mesonephric Tubules**
		Q80TF3			
		P97938			
		Q8C7A7			
		P09631			
		Q3TX21			
		O70373*			
		Q9D9B2			
		O55127			
27	**29**	**331**	**34**	**10**	**2**

Table B.4: The table lists all possible intersections of three or more of the gene lists in Tbl. B.3.

$A \cap M \cap S$	$A \cap N \cap S$
Q8BM75	Q8C2P1
1	**1**

Table B.5: The table lists the UniProt ID's (or the MGI Accession Key, if there is no UniProt ID available) of all genes that were expressed in the *Before* structures, but have turned off in the *After* structures.

Angiogenesis	**Cardiac Endothelium**	**Somites**	**Nephrons**	**Nephric Duct**	**Mesonephric Tubules**
Q99MK8	Q99MK8	Q5SDA2	P19091	P48540	
Q497W8	Q497W8	Q99KA8	Q9QWQ1	Q8BM75	
Q8CCE0	Q3ULR1	P70658	Q9QWR9	P97766	
P18111	Q8CCE0	Q7TS73	Q61850		
P02463	P18111	Q8CAT6	Q3UQJ4		
Q542N0	P02463	Q6GTZ3	Q3TWK8		
P15656	Q542N0	Q9Z197	Q496Q8*		
Q9R1A2	P15656	Q922Z8	Q8BZY5		
Q5BJ23	Q9R1A2	Q3UMZ6	P31313		
Q8BNY0	Q5BJ23	Q68EF7	Q08624		
P82976	Q8BNY0	Q9D7K8	Q543H4		
Q3UUJ3	P82976	Q62392	Q8BQA3		
Q02591	Q3UUJ3	Q9Z2C5	P09633		
Q8BNI8	Q02591	Q6PEB3	P23813		
P02831	Q9R1X2	Q6PFZ9	Q8BSN0		
Q61681	Q8BNI8	Q80UL5	Q8VHP0		
Q3ZAX9*	P02831	Q9QX13	Q3U1N3		
P09632	Q61681	Q9WV93	Q62438		
Q8BV11*	Q3ZAX9*	Q9JLF7*	O55222		
Q3UJB6	P09632	Q9R205	Q9Z0Y6		
Q9CRX6	Q8BV11*	Q8BHS3	Q9CTF6		
Q6P8P3	Q3UJB6	Q9CZM1*	Q920C1		
Q9EQ12	Q9CRX6	Q8R3I2	Q9R0R2		
Q9R2A7*	Q6P8P3	Q8K0C8	Q3KQI1*		
Q925V3	Q9EQ12	Q99J68	Q6P071		
Q8K428	Q9R2A7*	Q9CYI8	Q9WVF5		
Q8VD35	Q8K428	Q91ZJ5	Q9R1A2		
Q9CU96	Q8VD35	Q91WG3	Q9CZD6		
Q8CGH8	Q9CU96	Q8BQI5	Q5BLJ8		
Q9CY80	Q8CGH8	Q544L9	Q91ZN8		
Q8C765	Q9CY80	Q8BSU4	Q80ZS9		
Q4FK48	Q8C765	Q9R001	Q62219		
Q78ZW9	Q4FK48	Q9D1X9	P23359		
Q7TQ06	Q78ZW9	Q80SY4	Q8VIK0		
P26687	Q7TQ06	O55003	P48540		
Q8VCV6	P26687	Q5SQB3	Q9Z1W4		
Q3UMZ6	Q8VCV6		Q80UW0		

Table B.5: Genes that turn off during the processes (cont.)

Angiogenesis	**Cardiac Endothelium**	**Somites**	**Nephrons**	**Nephric Duct**	**Mesonephric Tubules**
Q80ZL6	Q3UMZ6		Q925H1		
Q8CCU9	Q80ZL6		Q6AZB0		
P19137	Q8CCU9		Q6PAS4		
Q8K1X3*	P19137		Q8CD68		
P20263*	Q8K1X3*		Q3ULR1		
Q5SDA2	P20263*		Q3UNK5		
Q543E8	Q5SDA2		Q811W8		
P23359	Q543E8		O35253		
Q5SRD8	P23359		Q8BUN5		
Q9WUL2*	Q5SRD8		Q9CRR7*		
Q80TC1	Q9WUL2*		Q80ZV9		
Q6PFV7	Q80TC1		Q9JIW5		
Q60756	Q6PFV7		Q8CDB8		
Q9DCA0	Q60756		Q99LS8		
P51655	Q9DCA0		Q99J48		
Q8BSB3	P51655		O09009		
Q3UQH0	Q8BSB3		Q3UPI0		
Q923Z1	Q3UQH0				
Q62318	Q923Z1				
P97766	Q62318				
Q8CC31	P97766				
P70658	Q8CC31				
Q8BTM5	P70658				
Q9QXX0	Q8BTM5				
Q8R381	Q9QXX0				
Q99KA8	Q8R381				
Q91VZ3*	Q99KA8				
O55127	Q91VZ3*				
Q9CXC9	O55127				
Q3UND5	Q9CXC9				
Q80UL7	Q3UND5				
O55233	O55233				
Q9JJL1	Q9JJL1				
Q9EQM2	Q9EQM2				
Q6PFZ9	Q6PFZ9				
Q544L3	Q544L3				
Q8R5G0	Q8R5G0				
Q6PGA9	Q6PGA9				
Q9Z0E2	Q9Z0E2				
Q9R1X4	Q9R1X4				
Q3UGA1	Q80UL5				
Q61271	Q61271				
Q9WVC6	Q9WVC6				
Q8K0H5	Q8K0H5				
Q60688	Q60688				
Q61583	Q61583				
Q9QXP9	Q9QXP9				
Q3V1F2	Q3V1F2				
Q9WUI0	Q9WUI0				
Q9QXN0	Q9QXN0				
Q9Z0Z7	Q9Z0Z7				
Q8R357	Q8R357				
Q9JI57	Q9JI57				
Q9JHX2	Q8R4A3				
Q8R044*	Q9JHX2				
Q80U19	Q8R044*				
Q9ESD2	Q80U19				
Q8QZV2	Q91XQ5				
Q6P071	Q9ESD2				
Q8QZY0	Q8QZV2				
Q91YX2	Q61639				
Q8BS64	P09631				
Q5U3K8	P09633				
Q64280	Q545I7				
Q8C5P2	Q5EEX1				
Q61080	Q9QUM0				
Q9ES03	Q6LEB3				

Table B.5: Genes that turn off during the processes (cont.)

Angiogenesis	**Cardiac Endothelium**	**Somites**	**Nephrons**	**Nephric Duct**	**Mesonephric Tubules**
P57785	Q8CCN5				
Q921T1	Q8BLF7				
Q9D2A8	P14246				
Q9JKQ8	Q9CRR7*				
Q60636	Q9QXT5				
Q7TMX8	Q544Z2				
P17439	Q3ULW0				
Q9R1L3	Q920W1				
Q9CZK5	Q9D8L6				
Q9R2A1	Q6P071				
Q9QUM0	Q8QZY0				
Q5SQP3	Q91YX2				
Q925F5*	Q5U3K8				
Q8VIL9*	Q64280				
Q9JK33	Q8C5P2				
Q543R9	Q61080				
MGI:106910*	P57785				
Q9CRR7*	Q9D2A8				
Q9D5V4	Q3U223				
Q91X98	Q9JKQ8				
Q3KP84	Q60636				
Q61639	Q7TMX8				
P09631	P17439				
P09633	Q9R1L3				
Q545I7	Q9CZK5				
Q5EEX1	Q9R2A1				
Q6LEB3	Q5SQP3				
Q8CCN5	Q925F5*				
Q8BLF7	Q8VIL9*				
P14246	Q9JK33				
Q3ULW0	Q543R9				
Q920W1	MGI:106910*				
Q9D8L6	Q9D5V4				
	Q91X98				
	Q9Z1Z8				
137	**139**	**36**	**54**	**3**	**0**

Table B.6: The table lists all the intersections of three or more of the gene lists in Tbl. B.5 that were not empty. The only gene which was tagged 'unknown' by *Fatigo+* (Q9CRR7*) has been marked with an asterisk.

$A \cap C \cap N$	$A \cap C \cap Nd$	$A \cap C \cap S$
Q9R1A2	P97766	Q5SDA2
P23359		Q99KA8
Q6P071		P70658
Q9CRR7*		Q3UMZ6
P09633		Q6PFZ9
5	**1**	**5**

Table B.7: The table lists the UniProt ID's (or the MGI Accession Key, if there is no UniProt ID available) of all genes that were not expressed in the *Before* structures, but have turned on in the *After* structures (i.e. are expressed in the *After* tissues). Furthermore, genes which have been tagged as 'unknown' by *Fatigo+* have been marked with an asterisk.

Angiogenesis	**Cardiac Endothelium**	**Somites**	**Nephrons**	**Nephric Duct**	**Mesonephric Tubules**
Q60753	Q9QVP4	P18111	P09631	Q8BNI8	P18111

Table B.7: Genes that turn on during the process (cont.)

Angiogenesis	**Cardiac Endothelium**	**Somites**	**Nephrons**	**Nephric Duct**	**Mesonephric Tubules**
Q61614	Q9R074	Q5BJ23	P09023	Q3ZAX9*	
Q8C2P1	Q9CWL2	Q9Z1Z8	P10628	Q8VHP0	
Q61824	Q8CJ69	Q9CT20	Q3UMQ3	Q8C8Q7	
Q8C6E4	Q3UNK5	Q9CU96	Q8R1P3	Q3UTY8	
Q3UQJ4	Q8BLK4	Q3UTY8	Q3V0Z9	Q9QWR9	
Q71V84	Q61272	Q544Z2	Q923S6	Q9DCI0	
Q9CRD0	Q9CXK3	Q8R381	Q8K428	Q80TF3	
Q9WV08	P19123	Q9WV08	Q6PCM9		
Q8BM75	Q6ZWX2	Q9QXN0	Q99L24		
Q0PHV7*	Q8VHX6	P47806	Q7TQI8		
Q9EPN2	Q99MV5	Q3UJB6	Q08EF2		
Q6GUA3	Q99K17	Q923F4	Q9QXX0		
Q8BHZ7	Q3UE22	Q8K4Q2	Q9WV93		
	MGI:1344335*	Q9CZK7	Q9DBX7		
	Q9D2T3	Q921T1	Q9D1D6		
	Q99KE3	Q80TC1	Q61045		
	Q921D9	Q61115	Q9D586		
		Q9D2N4	Q99MU3		
		Q7TNN4	Q9QZM3		
		Q3UZ96	Q920W1		
		Q3UY25	Q62433		
		Q3V1F2	O35437		
		Q80UW0			
		Q9JI57			
		Q8BTF1			
		Q8BM75			
		Q8BYL5			
		P02831			
		Q9CXK3			
		Q8CCE0			
		Q9ESV5*			
		P12979			
		Q5BLJ8			
		Q8CFN5*			
		Q78DU3			
		O35392			
		MGI:2443183			
		Q496Q8*			
		Q8C2P1			
		Q3ULW0			
		Q61477			
		Q9JII6			
		Q8CD62			
		Q8CF69			
		Q61685			
		Q6TDG5*			
		Q80ZL6			
		Q9QXA3			
		Q9JLL0			
		Q8BHZ7			
		Q91XQ5			
		Q5SYC7			
		Q0PHV7*			
		Q61272			
		Q60636			
		Q9D748			
		Q3URE6			
		Q3UBP0			
		Q8BSI9			
		P39061			
		Q99LW4			
		Q3USF7			
		Q542N0			
		Q8C399			
		P10284			
		P09024			
		Q8BQ25			

Table B.7: Genes that turn on during the process (cont.)

Angiogenesis	Cardiac Endothelium	Somites	Nephrons	Nephric Duct	Mesonephric Tubules
		Q6Q533			
		Q8VHZ2			
		Q7TPG3			
		Q9Z2D6			
		Q62219			
		P52955			
		Q5SSV4			
		Q60753			
		Q99KF3			
		O35085			
		Q3TV22			
		Q9D8U0			
		Q545G3			
		Q3U8E7			
		Q9WTS6			
		Q9CS91			
		Q3UMW5			
		Q4VBC9			
		Q9R0R4			
		Q9WUZ7			
		Q9CS81			
		Q9D1D6			
		Q7TQ32			
		Q7TT16			
		Q9EPK5			
		Q8BGT1			
		Q9D8G2			
		Q8R3I6			
		Q9ESD2			
		Q6AZB0			
		Q8CCN5			
		Q6T264			
		Q8R0M1			
		Q8QZV2			
		Q68BG2			
		MGI:2183691*			
		P10628			
		P09632			
		Q8BQA3			
		Q9CUQ8			
		Q3TYD9			
		Q80YP5			
		Q8CC06			
		Q545G1			
		Q8C6B1			
		Q61045			
		Q640N1			
		O08665			
		Q9JHE6			
		Q6NZL8			
		Q8BIA3			
		Q66PY1			
		P28481			
		Q99M23			
		Q8BG74			
		Q9ERH7			
		Q3U223			
		Q9Z0F1			
		P31311			
		P51655			
		Q8QZY9			
		Q9WVF2			
		Q811E4			
		Q8JZL1			
		Q3TZM2			
		Q9ERF6			
		Q8BYL6			

Table B.7: Genes that turn on during the process (cont.)

Angiogenesis	Cardiac Endothelium	Somites	Nephrons	Nephric Duct	Mesonephric Tubules
		Q8BQ62			
		Q9JJY2			
		Q6DI71*			
		Q543H4			
		P09633			
		Q3UMQ3			
		Q05DD2*			
		Q3UVH8			
		Q6P6L3			
		P54823			
		Q61809			
		Q8C2A8			
		Q9EQW7			
		Q80VZ1			
		Q3UZB9			
		Q9JLL4*			
		O88876			
		Q62433			
		Q9Z1S5			
		MGI:1861674*			
		Q9QZ26			
		Q99N11			
		Q9JM62			
		Q8CFG0			
		Q8K1S7			
		Q99N43			
		Q8K350			
		Q8C4C4			
		Q9DBJ9			
		Q3TZE7			
		P82976			
		Q8C9K1			
		Q9JL91			
		Q99LS8			
		Q9D6N4			
		P26687			
		Q9D080			
		Q8R228			
		Q8K3G6			
		Q76LY2			
		Q8VHX6			
		Q3ZAZ1			
		Q61824			
		Q6NXW7			
		Q99LQ5			
		Q9Z139			
		Q9R0G8			
		Q9JLI1			
		Q9JJZ5			
		Q9ERP3			
		Q99PV5			
		Q571J2			
		Q6P5N9			
		Q61080			
		Q8CJ69			
		Q8BYQ8			
		Q8BMD9			
		Q8BLU0			
		Q3UYE4			
		Q9JI91			
		Q99N13			
		Q78TF3			
		Q60932			
		Q9Z0Z7			
		Q9CZP5			
		Q8K2P6			
		Q61321			

Table B.7: Genes that turn on during the process (cont.)

Angiogenesis	**Cardiac Endothelium**	**Somites**	**Nephrons**	**Nephric Duct**	**Mesonephric Tubules**
		Q9QX23			
		Q9DC41			
		Q8VHX3*			
		Q0VEJ7			
		Q9CZ19			
		Q6PCW9*			
		Q05DE6*			
		Q8BRJ5			
		Q8C6B5			
		Q62232			
		Q8C766			
		Q9Z2G6			
		Q8VHZ3			
		Q8CFR7			
		Q9D677			
		Q8BZ84			
		Q9EPR5			
		Q9JKB3			
		Q8R419			
		Q8CJG1			
		Q8R3Q7			
		Q8CJF9			
		MGI:2676881*			
		Q3KP84			
		Q8R145			
		Q9CUZ5			
		Q9D030			
		Q80W09			
		Q80UF6			
		Q9WUU4			
		Q9JKV7			
		Q8CGN4*			
		P23813			
		P70217			
		P70323			
		Q810F8			
		Q8BSF9			
		Q921F1			
		P97938			
		Q8C7A7			
		P09631			
		Q3TX21			
		O70373*			
		Q9D9B2			
		O55127			
		Q9JJ37			
14	**18**	**248**	**23**	**8**	**1**

Bibliography

[1] J.P. Thiery. Epithelial-mesenchymal transitions in development and pathologies. *Current Opinion in Cell Biology*, 15(6):740–746, 2003.

[2] R. Kalluri and E.G. Neilson. Epithelial-mesenchymal transition and its implications for fibrosis. *Journal of Clinical Investigation*, 112(12):1776–1784, 2003.

[3] J.P. Thiery and J.P. Sleeman. Complex networks orchestrate epithelial-mesenchymal transitions. *Nature Reviews Molecular Cell Biology*, 7(2):131–42, 2006.

[4] D. Zipori. Mesenchymal stem cells: harnessing cell plasticity to tissue and organ repair. *Blood Cells, Molecules and Diseases*, 33(3):211–215, 2004.

[5] RU de Iongh, E. Wederell, FJ Lovicu, and JW McAvoy. Transforming growth factor-beta-induced epithelial-mesenchymal transition in the lens: a model for cataract formation. *Cells Tissues Organs*, 179(1-2):43–55, 2005.

[6] J. Bard. Personal Communication, 2007.

[7] C.L. Chaffer, E.W. Thompson, and E.D. Williams. Mesenchymal to Epithelial Transition in Development and Disease. *Cells Tissues Organs*, 185:7–19, 2007.

[8] P. Ekblom. Developmentally regulated conversion of mesenchyme to epithelium. *The FASEB Journal*, 3(10):2141–2150, 1989.

[9] B. Christ and C.P. Ordahl. Early stages of chick somite development. *Anatomy and Embryology*, 191(5):381–396, 1995.

[10] W. Zhang, QD Morris, R. Chang, O. Shai, MA Bakowski, N. Mitsakakis, N. Mohammad, MD Robinson, R. Zirngibl, E. Somogyi, et al. The functional landscape of mouse gene expression. *Journal of Biology*, 3(5):21, 2004.

[11] H.K. Lee, A.K. Hsu, J. Sajdak, J. Qin, and P. Pavlidis. Coexpression Analysis of Human Genes Across Many Microarray Data Sets, 2004.

[12] O.G. Troyanskaya, K. Dolinski, A.B. Owen, R.B. Altman, and D. Botstein. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). In *Proceedings of the National Academy of Sciences*, volume 100, pages 8348–8353. National Academy of Sciences, 2003.

[13] DJ Allocco, IS Kohane, and AJ Butte. Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics*, 5(1):18, 2004.

[14] MR Carlson, B. Zhang, Z. Fang, P.S. Mischel, S. Horvath, and S.F. Nelson. Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC Genomics*, 7(1):40, 2006.

[15] J.M. Stuart, E. Segal, D. Koller, and S.K. Kim. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science*, 302(5643):249–255, 2003.

[16] Q. Sheng, Y. Moreau, F. De Smet, K. Marchal, B. Home, et al. *Data Analysis and Visualization in Genomics and Proteomics*, chapter Advances in cluster analysis of microarray data, pages 153–173. Wiley Publishing, New York, NY, USA, 2005.

[17] A. Laegreid, T.R. Hvidsten, H. Midelfart, J. Komorowski, and A.K. Sandvik. Predicting Gene Ontology Biological Process From Temporal Gene Expression Patterns. *Genome Research*, 13(5):965–979, 2003.

[18] B. Smith, J. Williams, and S. Schulze-Kremer. The ontology of the gene ontology. In *Proceedings of the AMIA Annual Symposium*, pages 609–613, 2003.

[19] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, and J.T. Eppig. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[20] F. Azuaje, H. Wang, and O. Bodenreider. Ontology-driven similarity approaches to supporting gene functional assessment. In *Proceedings Of The Eighth Annual Bio-Ontologies Meeting*, 2005.

[21] Wikipedia. Ontology — Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Ontology, 2007. [Online; accessed 13-August-2007].

[22] T.R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, 43(5/6):907–928, 1995.

[23] P.V. Ogren, K.B. Cohen, and L. Hunter. Implications of compositionality in the gene ontology for its curation and usage. *Pacific Symposium on Biocomputing*, 10:174–185, 2005.

[24] PW Lord, RD Stevens, A. Brass, and CA Goble. Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–1283, 2003.

[25] M. Ashburner, C.A. Ball, J.A. Blake, H. Butler, J.M Cherry, J. Corradi, K. Dolinski, JT Eppig, M. Harris, and DP Hill. Creating the gene ontology resource: design and implementation. *Genome Research*, 11(8):1425–1433, 2001.

[26] P. Khatri and S. Draghici. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21(18):3587, 2005.

[27] R.J. Cho, M. Huang, M.J. Campbell, H. Dong, L. Steinmetz, L. Sapinoso, G. Hampton, S.J. Elledge, R.W. Davis, and D.J. Lockhart. Transcriptional regulation and function during the human cell cycle. *Nature Genetics*, 27:48–54, 2001.

[28] M.Z. Man, X. Wang, and Y. Wang. POWER_SAGE: comparing statistical tests for SAGE experiments. *Bioinformatics*, 16(11):953–959, 2000.

[29] L. Fisher and G. Van Belle. *Biostatistics*. Wiley Publishing, New York, NY, USA, 1993.

[30] M.E. Stokes, G.G. Koch, and C.S. Davis. *Categorical Data Analysis Using the Sas System*. SAS Publishing, Cary, NC, USA, 2000.

[31] S. Draghici, P. Khatri, R.P. Martins, G.C. Ostermeier, and S.A. Krawetz. Global functional profiling of gene expression. *Genomics*, 81(2):98–104, 2003.

[32] R. Díaz-Uriarte, F. Al-Shahrour, and J. Dopazo. The Use Of GO Terms To Understand The Biological Significance Of Microarray Differential Gene Expression Data. In *Proceedings of the Third Conference for Critical Assessment of Microarray Data Analysis*, 2002.

[33] S. Zhong, L. Tian, C. Li, K.F. Storch, and W.H. Wong. Comparative analysis of gene sets in the gene ontology space under the multiple hypothesis testing framework. In *Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB'04)*, pages 425–435, Washington, DC, USA, 2004. IEEE Computer Society.

[34] S. Draghici. *Data analysis tools for DNA microarrays*. Chapman & Hall/CRC Boca Raton, 2003.

[35] S. Draghici, P. Khatri, P. Bhavsar, A. Shah, S.A. Krawetz, and M.A. Tainsky. Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Research*, 31(13):3775–3781, 2003.

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

[37] B. Zhang, D. Schmoyer, S. Kirov, and J. Snoddy. Gotree machine (gotm): a web-based platform for interpreting sets of interesting genes using gene ontology hierarchies. *BMC Bioinformatics*, 5, 2004.

[38] F. Al-Shahrour, P. Minguez, J. Tarraga, I. Medina, E. Alloza, D. Montaner, and J. Dopazo. FatiGO+: a functional profiling tool for genomic data. Integration

of functional annotation, regulatory motifs and interaction data with microarray experiments. *Nucleic Acids Research*, 2007.

[39] F. Al-Shahrour, P. Minguez, J. Tarraga, D. Montaner, E. Alloza, J.M. Vaquerizas, L. Conde, C. Blaschke, J. Vera, and J. Dopazo. BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. *Nucleic Acids Research*, 34(Web Server issue), 2006.

[40] T. Beissbarth et al. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, 20(9):1464–1465, 2004.

[41] S. Zhong, K.F. Storch, O. Lipan, M.C. Kao, C.J. Weitz, and W.H. Wong. GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in Gene Ontology space. *Applied Bioinformatics*, 3(4):261–264, 2004.

[42] F. Al-Shahrour, R. Díaz-Uriarte, and J. Dopazo. Fatigo: a web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics*, 20(4):578–580, 2004.

[43] P.L. Ross, Y.N. Huang, J.N. Marchese, B. Williamson, K. Parker, S. Hattan, N. Khainovski, S. Pillai, S. Dey, S. Daniels, et al. Multiplexed Protein Quantitation in Saccharomyces cerevisiae Using Amine-reactive Isobaric Tagging Reagents. *Molecular & Cellular Proteomics*, 3(12):1154–1169, 2004.

[44] A. Rodriguez, S. Griffiths-Jones, J.L. Ashurst, and A. Bradley. Identification of Mammalian microRNA Host Genes and Transcription Units. *Genome Research*, 14(10 a):1902–1910, 2004.

[45] D.R. Rhodes and A.M. Chinnaiyan. Integrative analysis of the cancer transcriptome. *Nature Genetics*, 37:31–37, 2005.

[46] W. Ning, C.J. Li, N. Kaminski, C.A. Feghali-Bostwick, S.M. Alber, Y.P. Di, S.L. Otterbein, R. Song, S. Hayashi, Z. Zhou, et al. Comprehensive gene expression profiles reveal pathways related to the pathogenesis of chronic obstructive pulmonary disease. In *Proceedings of the National Academy of Sciences*, volume 101(41), pages 14895–14900. National Acadademy of Sciences, 2004.

[47] K. Theiler. *The House Mouse: Atlas of Mouse Development*. Springer-Verlag, New York, NY, USA, 1989.

[48] J.B.L. Bard, M.H. Kaufman, C. Dubreuil, R.M. Brune, A. Burger, R.A. Baldock, and D.R. Davidson. An internet-accessible database of mouse developmental anatomy based on a systematic nomenclature. *Mechanics of Development*, 42:111–120, 1998.

[49] MedicaLook. Blood vessels — MedicaLook human anatomy. http://www.medicalook.com/human_anatomy/organs/Blood_vessels.html, 2007. [Online; accessed 20-August-2007].

[50] M. Ringwald, J.T. Eppig, D.A. Begley, J.P. Corradi, I.J. McCright, T.F. Hayamizu, D.P. Hill, J.A. Kadin, J.E. Richardson, and O. Journals. The Mouse Gene Expression Database (GXD). *Nucleic Acids Research*, 29(1):98–101, 2001.

[51] DP Hill, DA Begley, JH Finger, TF Hayamizu, IJ McCright, CM Smith, JS Beal, LE Corbani, JA Blake, JT Eppig, et al. The mouse Gene Expression Database (GXD): updates and enhancements. *Nucleic Acids Research*, 32(Database issue):568–571, 2004.

[52] M. Ringwald, RA Baldock, J. Bard, MH Kaufman, JT Eppig, JE Richardson, JH Nadeau, and D Davidson. A database for mouse development. *Science*, 265:2033–2034, 1994.

[53] Bard J.B.L., Baldock R.A., Kaufman M., and Davidson D. Graphical gene-expression database for mouse development. *European Journal of Morphology*, 35(1):32–34, 1997.

[54] D. Davidson, J. Bard, R. Brune, A. Burger, C. Dubreuil, W. Hill, M. Kaufman, J. Quinn, M. Stark, and R. Baldock. The mouse atlas and graphical gene-expression database. *Seminars in Cell and Developmental Biology*, 8(5):509–517, 1997.

[55] R.A. Baldock, J.B.L. Bard, A. Burger, N. Burton, J. Christiansen, G. Feng, B. Hill, D. Houghton, M. Kaufman, J. Rao, et al. EMAP and EMAGE. *NeuroInformatics*, 1(4):309–325, 2003.

[56] J. Minker. On Indefinite Databases and the Closed World Assumption. In *Proceedings of the 6th Conference on Automated Deduction*, pages 292–308. Springer-Verlag, London, UK, 1982.

[57] G. Reese. Database programming with JDBC and JAVA. *O'Reilly Java Series*, 1997.

[58] Sybase. *JConnect for JDBC*. Sybase, Inc., USA, 1997.

[59] M. Gilbert. *Developmental Biology*. Sinauer Press, Sunderland, MA, USA, 7 edition, 2007.

[60] S. Boyle, Shioda. T., A.O. Perantoni, and A. de Castaestecker. Cited1 and Cited2 are differentially expressed in the developing kidney but are not required for nephrogenesis. *Developmental Dynamics*, 236:2321–2330, 2007.