Chapter 1. Transcriptional Profiling for DBA/2J glaucoma

Chapter 2. Developing Glaucoma Discovery Platform
   2.1 Maximizing the benefit of transcriptional profiling with Glaucoma Discovery Platform.
   2.2 Database construction.
   2.3 Developing the web interface.

Chapter 3. Performing searches and data retrieval
   3.1 Introduction.
   3.2 Home page/new search
   3.3 The results summary page
   3.4 Probe set details page
   3.5 Expression values page
   3.6 External Resources

Chapter 4. Worked example – assessing all genes in the TNF superfamily
   4.1 Introduction
   4.2 Performing detailed search
   4.3 Visualizing the results
   4.4 Interpreting the results

Chapter 5. References
Chapter 1: Transcriptional Profiling for DBA/2J glaucoma

Glaucoma is a complex, neurodegenerative disorder affecting 70 million people worldwide and is associated with the death of retinal ganglion cells (RGCs) and the associated degeneration of the optic nerve [1]. DBA/2J is a widely used mouse model of glaucoma that shows hallmarks of human glaucoma including age-related IOP elevation, optic nerve excavation and regional patterns of RGC loss [2-7]. DBA/2J mice develop glaucoma as a result of a disease of the iris that leads to an elevation in IOP. The disease of the iris is caused by mutations in two genes, Gpnmb\textsuperscript{R150X} and Tyrp1\textsuperscript{b}[2, 3]. DBA/2J-Gpnmb\textsuperscript{+} mice have a functioning Gpnmb gene and serve as a genetically matched control strain that does not develop glaucoma [8]. An important insult occurs to RGC axons at the optic nerve head in DBA/2J glaucoma [4]. However, other compartments of the RGC also are likely to undergo early changes in glaucoma such as the RGC soma [9] and their synapses [10]. The mechanisms involved in these early changes are not well understood.

We performed a gene expression profiling study of DBA/2J mice to investigate early changes in both the optic nerve head and retina separately for each eye (described in detail elsewhere, [11]). Briefly, the optic nerve head and retina were separately profiled using Mouse 430 v2 arrays (Affymetrix) for 50 DBA/2J eyes and 10 DBA/2J-Gpnmb\textsuperscript{+} controls. All data were processed and analyzed using MAANOVA [12]. DBA/2J eyes were initially grouped based on conventional morphological criteria including degree of axon damage (dataset 1: four groups, Figure 1.1A). However, these groups were not sensitive to identify early stages of disease that precede morphological damage. Therefore, hierarchical clustering, a method widely used in cancer biology, [4, 8, 13] was performed to identify groups of eyes undergoing early molecular changes. To our knowledge, this was the first time molecular clustering had been applied to the study of a complex disease. Eyes were clustered into different stages using both the expression profiles from the optic nerve head (dataset 2: five stages, Figure 1.1B) and the retina (dataset 3: four stages, Figure 1.1C). To identify differentially expressed genes for all three datasets, all possible pairwise comparisons were performed. In total, more than 70 different comparisons were generated and many thousands of differentially expressed genes identified. All raw data has been deposited in NCBI GEO (Accession number: GSE26299)
Figure 1.1: Schematic of the datasets available in the first release of Glaucoma Discovery Platform. Each of the three datasets provides a progression through glaucoma (from no glaucoma to severe glaucoma) defined either morphologically or molecularly [11]. (a) For each eye, the optic nerve head and retina were analyzed from 10.5 months old DBA/2J (white boxes) and D2-Gpnmb+ control strains (grey box). Stages of glaucoma were determined morphologically using axon damage assessment of the optic nerve just behind the orbit (see Methods and [8, 13, 14]). (b) Hierarchical clustering using the gene expression levels of selected glaucoma-relevant genes was used to group the optic nerve heads into 5 molecularly defined stages. Stages 1, 2 and 3 identify early stages of glaucoma not previously detectable using morphological analysis. Stages 4 and 5 contain eyes with moderate and severe axon damage respectively. (c) A similar hierarchical clustering was performed using the retina tissue. Four stages of disease were identified with stages 1 and 2 not previously detectable using morphological analysis. Stages 3 and 4 contain eyes with moderate and severe axons damage respectively. Optic nerve heads and retinas were assessed from the same set of eyes. However, different eyes make up the optic nerve head stages compared to the retina stages. This suggests a possible asynchrony of events in the optic nerve head compared to the retina. Therefore the optic nerve head dataset is most sensitive at identifying early molecular changes in the optic nerve head, whereas the retinal stages are most sensitive at identifying early molecular events in the retina.
Chapter 2: Developing Glaucoma Discovery Platform

2.1 Maximizing the benefit of transcriptional profiling with Glaucoma Discovery Platform.
Existing resources such as GEO and ArrayExpress were not suitable for interrogating and visualizing such as complex dataset. Therefore, to maximize the benefit of this study for us, and the wider scientific community, we developed Glaucoma Discovery Platform, a freely available web-based environment. Glaucoma Discovery Platform was developed using a suite of scripts we term Datga. To our knowledge, no other resource is available that provides this combination of user-friendly functionality provided by Glaucoma Discovery Platform.

2.2 Database construction.
To construct Glaucoma Discovery Platform, all pairwise comparisons for all 3 datasets (Morphological, Molecular ONH and Molecular retina) were incorporated into a series of data tables. The data tables architecture relates features of the three datasets including: raw intensity values; probe sets; samples; dataset groupings; comparisons of these groupings; statistics associated with the comparisons; and genetic annotations of the probe sets. A design file is generated to establish relationships between the raw data for each sample and the sample groups represented in the experimental design. Gene annotations to the probe sets are loaded from the Mouse Genome Database [15] using public reports available from the Mouse Genome Informatics (MGI) ftp site (http://www.informatics.jax.org). This provides the ability not only to search by gene symbols, but also by their synonyms/aliases.

2.3 Developing the web interface.
The web-interface was implemented using the Ruby on Rails web application framework (rubyonrails.org) plus some custom and public Javascript libraries. This interface leverages AJAX technology (asynchronous JavaScript and XML) to allow dynamic regeneration of plots in the same page view, while maintaining the main query panel. The database application infrastructure was implemented in a generic manner, allowing for its reuse for other analyzed microarray datasets, and to make it easier to load additional experimental results to the Glaucoma Discovery Platform. The web-based interactive search and retrieve component provides the ability to assess multiple gene(s) of interest in temporally and/or spatially defined developmental or disease stages. The output is returned in graphical and tabular format with statistically significant differences highlighted for easy visual analysis. Unlike NCBI GEO or EBI ArrayExpress, data for all probe sets for a given gene can be accessed as well as all data for individual samples within groups. Additionally, a bulk download component allows lists of differentially expressed genes to be retrieved as a series of tab delimited files. To facilitate access to additional functional information, for a given gene, links are provided to external resources including Entrez Gene, organism-specific databases such as Mouse Genome Informatics, UCSC Genome Browser and Online Mendelian Inheritance in Man (OMIM).
Chapter 3. Performing searches and data retrieval

3.1 Introduction. The web-based interface is divided into 4 main sections; (a) homepage/new search, (b) results, (c) probe set details and (d) expression values (Figure 3.1). A number of links are provided to selected external resources such as Mouse Genome Informatics [16], EntrezGene and the Online Medelian Inheritance in Man (OMIM) databases [17] and Kyoto Encyclopedia of Genes and Genomes (KEGG) [18] that allows users to access genetic, phenotypic and functional information for genes of interest (see External links below). The platform is also well supported by help pages.

3.2 Home page/new search (Figure 3.1a): The home page provides an overview of the database, links to the bulk download page, and the new detailed search box. The bulk downloads page allows all differentially expressed genes and associated information for individual comparisons to be downloaded in excel-friendly format. The home page also has a quick search feature that allows a single gene to be searched in all datasets. There are also links to internal help pages including instructions for accessing important gene information from external resources.

There are a variety of different ways to assess the data using the detailed search tool. Firstly, if a user has a set of genes they are interested in assessing, the official gene symbols (or MGI-recognized aliases) can be entered or pasted into the appropriate text box (Figure 4.1). Alternatively, there is a wild card (*) capability where, for instance, the results for all members of the tumor necrosis (Tnf) superfamily can be retrieved by searching for “Tnf*” (described in detail below and Figure 4). The user is able to select the dataset(s) of interest (see Figure 1.1), the tissue(s) of interest (retina or optic nerve head) and the reference group (e.g. the D2-Gpnmb+ control group). If no datasets are selected, the default is to search all datasets for all available tissues. It is possible to restrict the results by fold change and/or significance value (q value). The user can select whether the results are ordered by significance (lowest average q value across all groups) or by gene symbol (alphanumeric). The q value is a measure of the false detection rate and gives an indication of the significance of the fold change and is a standard statistic for microarray analysis [19]. The lower the q value, the more significant a fold change is considered to be. In our study, genes are considered differentially expressed, with respect to the reference, if the q value is less than 0.05 (roughly equivalent to a false detection rate of 5%). Finally, fold changes can be reported as either relative fold change (compared to reference), or as log2 fold change (compared to reference).

3.3 The results summary page (Figure 3.1b): Results are returned in both graphical and tabular format. Any gene names that were not recognized from the search page are listed and links to MGI (for clarification of official gene symbols) are given. A ‘summary of results’ table is shown indicating the number of genes found for each dataset/tissue selected on the search page. Below, there are separate tables and associated graphs. Each table contains the gene names searched, the official gene symbol (or description), the representative probe set, and the fold change with associated q value for each stage of glaucoma, compared to the chosen reference group. For ease of identification, significantly differentially expressed values are shown in white on a red background.
Genes are ranked in the table based on the chosen option in the detailed search tool (q value or gene symbol). The graph provides a visual display of the fold changes across all stages of disease for the genes of interest, up to a maximum of ten. For searches containing more than 10 genes, the first 10 genes in the table are displayed in the graph. In addition, check boxes are available to left of each gene to allow the user to select up to 10 genes of choice to view in the graph. Results from the table can be exported in an excel-friendly file format (.csv). Probe set details for individual genes or selected genes can be retrieved. External links to useful databases are also provided (see External resources below).

3.4 Probe set details page (Figure 3.1c): Each gene on the Affymetrix 430 v2 array is interrogated by a probe set of 11 probes and a gene can have multiple probe sets [20]. Each probe set corresponds to a particular region of a gene and finding that multiple probes sets for the same gene behave similarly can add confidence to a result. Alternatively, for those probe sets designed in alternatively spliced exons/transcripts, multiple probe sets can give insight into the behavior of splice variants [21]. The values for the representative probe set for each gene is displayed in the results summary page. It is possible to view details for all probe sets for a given gene, to view probe sets for selected genes or for all genes in the table. The probe set detail page provides a table/graph for each gene.

3.5 Expression values page (Figure 3.1d): Values provided in the summary pages are relative fold changes or log2 fold changes compared to a selected reference. Each stage of disease contains biological replicates, and the fold change for a given probe set (for a given gene) is calculated as the average of the normalized expression value of all replicates for a defined stage of disease, relative to the reference group. The normalized expression value reflects the relative abundance of the gene in the tissue interrogated (either retina or optic nerve head) with respect to all other genes. Those genes with the highest normalized expression value are most abundant. Conversely, those with the lowest normalized expression value are least abundant. An expression level of approximately 4 or less may be considered to represent a gene that is likely not expressed in the assessed tissue (i.e. is close to background levels).

The relative abundance of transcripts corresponding to particular probe sets is important information to access. For instance, a small fold change in some lowly expressed gene may have a greater biological outcome than a small fold change in some abundant genes. Also, within stages of disease (determined either morphologically or molecularly), individual biological replicates may behave differently. Knowing the variability in expression levels for different genes in eyes within groups is important. Genes with low variability within groups may be better targets for intervention strategies than variable genes. The expression values corresponding to each probe set for individual samples (replicates) can be accessed from the probe details page. The expression levels are displayed as a histogram. The average (±1 standard deviation) for the reference group is indicated.

Figure 3.1: Highlighting the main functions available within Glaucoma Discovery Platform. (a) The home page contains direct links to help pages, the detailed search tool, bulk downloads and a quick search
that interrogates all datasets for a single gene of interest. The detailed search tool enables individual genes, groups of genes or wild card searches (*). The bulk download option enables all differentially expressed genes for pairwise comparisons to be retrieved. (b) The gene results page provides both a tabular and graphical view of the expression levels for the searched genes in the different groups for each dataset selected. Access to gene information from external databases is provided (c) As multiple probe sets exist for many genes on the Affymetrix 430v2 array, the probe set page details the results for each probe for a chosen gene(s). (d) The normalized expression values for individual eyes in each group can be accessed.

3.6 External Resources: Selected external resources can be accessed directly from the results tables (Table 3.1). Resources were selected to allow users the maximum access to current information for genes of interest. A major resource for mouse-based expression is Mouse Genome Informatics (MGI). This resource provides the research community with information on genetics, genomics and biology of mice [16, 22]. EntrezGene and OMIM
are databases that form part of the Entrez system at the National Center for Biotechnology Information (NCBI) [17]. EntrezGene provides gene-relevant information such as transcript/protein sequences and links to relevant publications and genome browsers. OMIM provides a simplified disease-oriented description for a given gene including mutations that have been shown to cause diseases in humans. Finally, there are links to The Gene Ontology (GO) database and the Kyoto Encyclopedia of Genes and Genomes (KEGG), two databases that provide functional descriptions of genes. GO provides a controlled vocabulary of terms for biological process, cellular compartments and molecular function and is accessed through MGI [16]. KEGG uses known functional information to construct biologically relevant pathways [18]. Given the uniformity of gene symbols between the external resources and Glaucoma Discovery Platform it is possible to identify groups of genes of interest (such as genes in a given KEGG pathway or genes with the same GO term) in the appropriate database and paste these genes into the search tool in Glaucoma Discovery Platform.

<table>
<thead>
<tr>
<th>Database</th>
<th>Acronym</th>
<th>Brief description</th>
<th>Home page address</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouse Genome Informatics</td>
<td>MGI</td>
<td>Gene catalogues, links to additional resources including phenotype data, genome browsers and SNP databases</td>
<td><a href="http://www.informatics.jax.org/">http://www.informatics.jax.org/</a></td>
</tr>
<tr>
<td>The Gene Ontology@MGI</td>
<td>GO</td>
<td>Controlled vocabulary for functional attributes</td>
<td><a href="http://www.informatics.jax.org/searches/GO_form.shtml">http://www.informatics.jax.org/searches/GO_form.shtml</a></td>
</tr>
</tbody>
</table>

**Chapter 4. Worked example - assessing all genes in the TNF superfamily**

In this section, we describe the workflow for extracting the data relevant to members of the TNF superfamily. This serves both to reinforce key features of the database described above and as a worked example that could be followed to interrogate any gene (or group of genes) of interest.

**4.1 Introduction.** TNF (formerly TNFα) has been suggested to be important in retinal ganglion cell loss during glaucoma [23, 24]. Therefore, we were interested to assess all members of the TNF superfamily in both the molecular ONH dataset (Figure 2.1B), most sensitive for predicting early disease changes in the optic nerve head), and the molecular
4.2 Performing detailed search (Figure 4.1). First, we determined the best search option, in this case, a wild card search using ‘TNF*’ will retrieve all TNF superfamily members. Although a wild card search takes a little longer to return the results, it does not require any a priori knowledge about which genes are present in the TNF superfamily. Second, we selected the datasets and tissues of interest; molecular ONH dataset, tissue ‘ONH’ and molecular retina dataset, tissue ‘retina’. Finally, we selected the reference group; in this case, D2-Gpnmb control eyes. In this example, we did not limit our search based on fold change or q value, and chose the default options for the layout of the results page (display expression differences as fold change rather than log2 of fold change, and ordered genes based on significance of gene expression differences (q value) rather than gene symbol).

Figure 4.1. Performing a detailed search for members of the tumor necrosis factor (TNF) superfamily. From the search tools box on the home page, ‘TNF*’ was entered into the ‘Wild card Gene Symbol’ search box. Both the ‘molecular ONH’ dataset (tissue ‘ONH’) and molecular retina (tissue ‘retina’) were selected. No limits were set for fold change or q value and the results are to be returned as fold change (FC - not log2 FC) in gene symbol order.

4.3 Visualizing the results (Figure 4.2). The results were returned below the original search box, allowing straightforward modification of search options if necessary. First, the summary table indicates that 50 genes met the criteria of the search for both the molecular ONH and molecular retina datasets (Figure 4.2A). These 50 genes have ‘TNF’ in either their official gene symbol or in any approved alias (MGI). Below the summary
of results table, detailed results for each dataset are displayed. In this example, we assess the results for the molecular ONH dataset, tissue ONH. First, the top 10 most significant genes are visualized in the graph (Figure 4.2B). The results for all 50 genes are shown in the table, the top 10 most significant of which are shown in Figure 5C. Of the 50 genes, 34 are differentially expressed in at least one stage in the molecular ONH dataset. To download all expression data in the table, use the ‘Download as comma separated values (CSV)’ option below the table. Expression values (and associated q values) for the 50 genes across all stages of disease are exported in excel-friendly format.

Figure 4.2 Results of the wild card search TNF*. (a) A summary of the results is provided indicating the number of genes identified in each of the tissues and datasets searched. (b-c) Below the summary table, the detailed results are returned in graphical (b) and tabular format (c). For the molecular ONH dataset, the relative fold change (with respect to the chosen reference) and the q value for the representative probe set for a gene is provided. The ten most significant differentially expressed genes in the TNF superfamily are shown. Links are provided for each gene to selected external resources. These are MGI – by clicking on the gene symbol (black arrow), UCSC (by clicking the representative probe ID, blue arrow) EntrezGene, OMIM, GO and KEGG (database (DB) links, far left). Details of all probe sets for a given gene can be accessed (green arrow).
4.4 Interpreting the results (Figure 4.3). *Tnf* shows only a modest increase in expression in stage 4 (1.2 fold, q value = 0.0413) in the molecular ONH dataset. The five most significant genes are *Tnf*, alpha-induced protein 8-like 2 (*Tnfaip8l2*), *Tnf* superfamily receptor 1a (*Tnfrsf1a*, formerly *Tnfr1*) and 1b (*Tnfrsf1b*, formerly *Tnfr2*), *Tnf* alpha-induced protein 2 (*Tnfaip2*) and *Tnf* superfamily member 9 (*Tnfrsf9*) (Figure 4.2C). *Tnfrsf1a* and *Tnfrsf1b* are two major receptors for *Tnf*, although in the molecular ONH dataset, *Tnfrsf9* has the highest fold change overall, with a 2.9 fold expression difference compared to D2-Gpnmb+ controls in stage 3. Of direct relevance to DBA/2J glaucoma, *Tnfrsf9* is involved in the proliferation of monocytes that are precursors of microglia and macrophages [25-27]. This information was retrieved from either OMIM or Entrez Gene. Microglia/macrophages have been shown to increase in the optic nerve head and retina in early in glaucoma [11, 28].

As described previously, many genes are represented by multiple probe sets on the Affymetrix 430 v2 array. *Tnfrsf9* is represented on the array by three probe sets (1460469_at, 1428034_a_at and 142481_at) with ‘1460469_at’ being shown in the results table as the most significant probe set of the three. Given the potential importance of all probe sets, full information can be accessed through the probe set details page (Figure 3C and 6). Interestingly, only 1460469_at is differentially expressed in the molecular ONH dataset, tissue ONH, the other two probe sets show no significant difference compared to the D2-Gpnmb+ control group (Figure 4.3A). At least three alternative transcripts are predicted for *Tnfrsf9*, and each probe set is designed specifically to each of the alternative transcripts (Figure 4.3B). This information can be accessed through a direct link to the UCSC genome browser. 1460469_at specifically corresponds to the major transcript (most widely studied transcript), whereas 1428034_a_at and 142481_at specifically correspond to two minor transcripts (least studied transcripts). This analysis would suggest that only the major transcript is expressed in the optic nerve head and this is reflected in expression values of the probe set 1460469_at. The other two transcripts are likely not expressed in the optic nerve head.

To assess the normalized expression levels of the three probe sets in the optic nerve head it is necessary to access the expression values page (Figure 4.3C). The normalized raw intensity values for individual eyes for 1460469_at range between 4 and 7. In contrast, the expression levels in individual eyes for 1428034_a_at and 142481_at are all less than 4 (data not shown). Values in this range are considered to be close to background intensity and it is likely only the major transcript is expressed in the optic nerve head.
Figure 4.3 Analyses of multiple probe sets for Tnfrsf9 predict only the major transcript is expressed in the optic nerve head. (a) The table from the probe set details page for the Tnfrsf9 gene (a graphical view is also available— not shown). Three probe sets are available for this gene. Only one probe set, 1460469_at, is differentially expressed in the optic nerve head. The remaining two probes show no difference with respect to D2-Gpnmb+ control group. The three probes sets are designed to three alternative transcripts for the Tnfrsf9 gene. The information can be accessed by clicking on probe set identifier (arrow). (b) The design information is displayed in the UCSC genome browser. The three predicted transcripts for Tnfrsf9 are shown with exons as vertical boxes connected by introns (arrow heads indicate direction of transcription). These correlate with the three probe set designs below. 1460469_at is designed to the major transcript (greatest amount of biologically verified supporting data) for Tnfrsf9 (arrowed). Little is known about the two other alternative transcripts. The normalized expression values for each eye in the cluster groups can be accessed from below the table shown in (a) (green arrow). (c) The normalized expression values (intensity values) for all stages for the 1460469_at probe set. The average of the reference group is shown (purple line) along with two standard deviations above and below the average (grayed area). The normalized raw intensity values range from between 4 and 7. This graph shows the variability of Tnfrsf9 expression in individual eyes. The normalized raw intensity values in all eyes for the other two probes are below 4 (data not shown) and suggests that only the major transcript is expressed in the optic nerve head.
Chapter 5. References

22. Shaw DR: Searching the Mouse Genome Informatics (MGI) resources for information on mouse biology from genotype to phenotype. *Curr Protoc Bioinformatics* 2009, Chapter 1:Unit1 7.


