Chapter 1

Beyond the Gene List: Exploring Transcriptomics Data in Search for Gene Function, Trait Mechanisms and Genetic Architecture

Bregje Wertheim

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/48239

1. Introduction

Since the start of genomics research, genome-wide expression studies have been used prolifically as a tool to improve our understanding of the involvement of genes in various biological processes. Measuring gene expression patterns simultaneously across all the genes in the genome, i.e. transcriptomics, is a uniquely powerful technology to explore potential novel candidate genes for a particular process. This genome-wide approach has the huge advantage that we do not have to specify in advance which genes we believe to be involved, and as such, we are not limited by our current knowledge. Transcriptomics is an important first step to study traits that are under the control of several to many genes (i.e., polygenic traits) and responsive to external conditions and internal states (i.e., multifactorial traits).

The identification of potential novel candidate genes, however, is only a limited part of the power of transcriptomics. With this technology, the expression of thousands of genes is measured simultaneously. It provides a snapshot of all genes that are actively transcribed during a particular process. When we compare these measurements between conditions or treatments, those genes that are expressed at higher or lower level under a particular condition can be identified. As such, transcriptomics maximizes the awareness of effects anywhere in the genome, including those associated by costs, trade-offs and epistatic interactions. This could be viewed as a complication of transcriptomics data, because a change in expression does not necessarily reflect a causal relationship to the process of interest. In fact, however, it is also one of the major strengths of this technology. By combining various bio-informatic tools and resources, it is possible to obtain an insight into intricate gene-interaction networks, the regulatory control of traits, and the implications of a trait or process on the full phenotype.

In functional genomics, transcriptomics studies are typically a comparison between biological samples (e.g., a cell type, organ, individual, or group of individuals) that were collected under different conditions, to analyse which genes were up-regulated or down-regulated (i.e., were expressed at higher, respectively, lower levels) in response or relation to the condition. These conditions can be experimentally induced (e.g., treatment *versus* control, different dosages of a chemical, different food conditions or temperatures, etc.), or they represent different natural stages (e.g., diseased *versus* healthy, male *versus* female, different developmental stages or aged individuals, different genotypes, different tissues, different epigenetic profiles, etc). Including a proper control treatment or reference is crucially important for the interpretation of gene expression differences that results from such a comparison. There will always be a large number of genes expressed in any biological sample, and without control or reference, it is impossible to attribute expression of particular genes to the condition of interest. The purpose of transcriptomics is to reveal how the expression patterns *change* under different conditions.

Transcriptomics technology is used to characterize the composition of the messenger RNA (mRNA) pools from each biological sample. The mRNAs are the transcripts of a gene that carry the information encoded in the gene to the site of protein synthesis. When a particular mRNA is present in a biological sample, it implies that the corresponding gene was expressed, and a template is available for the synthesis of the protein product of that gene. The abundance of each mRNA in the pool represents the level of expression of the corresponding gene. By comparing the relative proportional representation of each mRNA in the total mRNA pool among the samples, we can identify which genes differed in expression in response or relation to the compared conditions. The most widely use technological platform for whole-genome expression studies are microarrays, although the sequencing of the transcriptome is rapidly increasing in popularity (Figure 1).

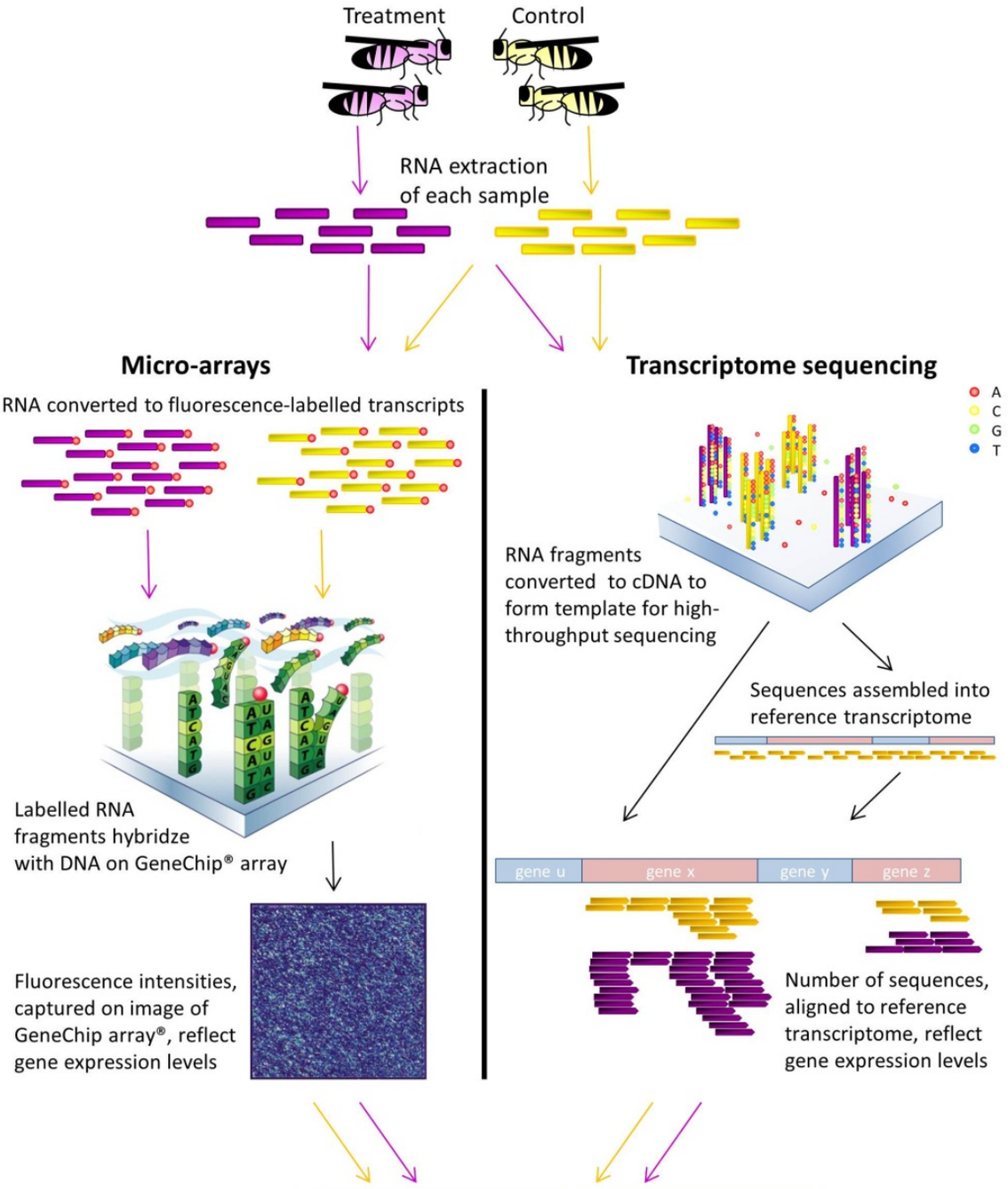
Microarrays are solid-based platforms (e.g., glass slides), containing millions of copies for thousands of 'reporter probes' that comprise part of the sequences of the genes in the genome. By binding (or 'hybridizing') fluorescent-labelled copies of the original mRNAs to the probes, measuring the label intensities for each position on the array, and associating these positions to their specific reporter probes, one can infer the presence and abundance of each transcript in the labelled RNA pool (Figure 1). It is assumed this representation is proportional to their abundances in the original mRNA samples. Microarrays are relatively cheap, and the tools to analyse the data have been developed, matured and tested. This makes microarrays an affordable and accessible platform for many applications [1]. After the initial introduction of expression arrays that reported only on known or predicted genes, tiling arrays were developed that contained reporter probes across the full genome, including the non-coding, non-translated and non-transcribed chromosomal regions. This enabled the identification of novel transcripts, including non-coding RNA genes, as well as a better characterization of splice variants and exons [2].

The latest developments in next-generation sequencing technologies are making transcriptome sequencing more affordable, and they provide a number of advantages over microarrays [3]. For this approach, the mRNA pool is converted into cDNA (either wholly

or after a partial digestion), which is then used as template for high-throughput sequencing. The generated sequence information is mapped to, or assembled into, a reference transcriptome, and the number of sequence copies generated for each gene is used to infer the number of mRNA copies in the original sample (Figure 1). Sequencing approaches provide more comprehensive information on the transcript characteristics (e.g. splice variants, mRNA sequence variations, gene fusions, etc.), they are not limited to the known or predicted genes of an organism or the genes represented on an microarray, and they avoid some problems inherent to slide-based technologies [4]. A downside of transcriptome sequencing is that the Quality Control and pre-processing and analysis procedures for these data have not yet fully matured, and the assembly of, or mapping against, the reference transcriptome requires substantial computing power, making this technology still less accessible.

In essence, both technological platforms yield data of very similar nature, although the information of sequencing approaches may be more specific and detailed than array-based approaches. After the specific pre-processing that each platform requires, the data can be analysed with similar methods, leading to a list or ranking of genes that show changes in expression patterns or transcript characteristics (e.g. splice variants) among the compared conditions. As such, the gene list provides a first step to identify the genes that potentially matter or are affected by a particular condition. A change in expression, however, is insufficient evidence for establishing a clear link between a gene and the trait of interest. At best, the genes on the list may be associated with the trait or condition of interest, while causality or direct involvement in the trait still needs to be established through additional empirical approaches.

Before discussing how gene lists can be generated or used for further analysis, it is important to emphasize that certain limitations are inherent to transcriptomics data. These limitations can be specific to the used platform, for instance microarrays can only report on the activity of genes that are known or represented on the array. Most limitations, however, are irrespective of the technology. As mentioned, genes that are differentially expressed are not necessarily causal to a particular trait or response. Moreover, not all the genes that are involved in a response or trait are detected by a changes in expression. Any post-transcriptional modifications or non-transcriptional processes (such as the re-directing of a transcription factor from its regular processes towards another function) are typically not detectable by a change in gene expression. A further precautionary note is warranted for the design, set-up and execution of any transcriptomics study. An essential requirement for associating changes in gene expression among different samples to a particular condition or treatment of interest, is to ensure that the only difference is the condition or treatment of interest. For example, the collection of control and treatment samples should be done simultaneously (e.g., not before infection and 12 hours after infection) by the same person, to avoid that circadian rhythms or handling effects differ between the samples. When such precautions would not be taken, genes responsive to the treatment would be confounded with genes responsive to these extraneous factors. It is impossible to resolve such confounding effects after the measuring of gene expression. The only way to avoid such



Gene ID	Transcript abundance in Treatment sample	Transcript abundance in Control sample
ABc00001	3778	3894
ABc00002	768	189
ABc00003	1087	1476

Figure 1. Schematic overview of transcriptomics approaches, using microarrays or transcriptome sequencing. Although the technologies differ, both approaches compare all the mRNAs in biological samples under different conditions, and provide quantifications of the abundance of all gene transcripts for each sample. *Images of GeneChips® courtesy of Affymetrix.*

issues is to take due care during experimental design, sample collection and sample preparation. Despite these limitations that are inherent to any transcriptomics technology, the resulting data does provide an array of possibilities for further meaningful analysis.

In this chapter, I will illustrate various ways in which transcriptomics data can be analysed, to identify novel candidate genes for the process of interest, and additionally, how to move beyond this list of candidate genes towards the molecular mechanisms and gene interaction networks of a trait. For these illustrations, I will mostly use transcriptomics data on the innate immunity in *Drosophila* larvae after parasitism. Our analysis on the transcriptomics during the acute immune response to infection by parasitic wasps [5], as well as between strains that differ genetically in resistance to these parasites [6], revealed a complex gene interaction network associated with defense mechanisms. The immune response to parasites is triggered by recognizing the invasion of the parasite, and comprises of the proliferation and differentiation of specialized blood cells that surround the parasite in a multi-layered capsule, and sealing the capsule with a layer of melanin. This melanotic encapsulation sequesters and kills the parasite [7]. By integrating the data from our studies with various resources and bioinformatics approaches, we gained a more comprehensive insight in the interactive and regulatory network of genes that are associated with the immune response to parasitism. We identified shared regulatory elements among genes that showed similar expression patterns, physiological costs associated with evoking the immune response, chromosomal positions that were associated with resistance traits and indications for epistatic gene-interactions. Combined, this information provided us with new insights on the mechanisms and complex genetic architecture of the innate immune response.

2. Constructing a list of genes with differential expression

The fundamental purpose of a transcriptomics experiment is to identify the genes with changed expression under a particular condition, which is done by comparing the abundance measurements for each gene transcript among the biological samples. Depending on the platform used, these abundance measurements are derived from fluorescence intensity measurements captured in digital images of the microarrays, or the counts of the number of transcripts for sequencing approaches (Figure 1). These measurements, however, are not only reflecting the biologically interesting variation in gene expression under the different conditions, but also a substantial level of technical variation that is introduced during the preparation and measuring of the samples. This includes, for example, residues of reagents that create a background signal on microarrays, short fragments of RNA that bind non-specifically to microarray probes or cannot be uniquely mapped to a reference genome, slight differences in RNA doses for the different samples, or slight differences among samples/batches in the efficiency of the molecular techniques. Some of these aspects affect whole samples, while others are specific to particular genes. To perform the meaningful comparisons on the variation in gene expression measurements, it is typically essential to first eliminate the bias introduced by technical variation as much as possible.

The raw intensity measurements first need to be pre-processed to deal with the technical variation, normalized to scale all samples to the same range, and combined into a single expression value per gene per sample for comparisons. Many different approaches have been developed for the pre-processing and normalization of microarrays, and subsequent studies have tried to determine the optimal strategy to remove the noise without introducing bias. Some approaches are outperforming others and consensus has been mostly reached for the commonly used microarray platforms, although full consensus for all microarray platforms is still lacking [8]. Also for transcriptome sequencing, normalization is important to address deviations due to slight differences in doses, the gene length and GC-content. The exploration of the best pre-processing and normalization approaches for transcriptome sequencing are still being established (e.g., [4,9]).

To statistically test for changes in gene expression, biological replication is essential. Having multiple biological units for each condition enables the estimation of variation within and between conditions, which allows for the partitioning of all variation into noise (i.e., technical and random variation), and the biologically interesting variation reflecting the changes in gene expression patterns. Technical replications are sometimes also incorporated in the platform or analysis, for example by repeating the same probes on a microarray, by applying a dye-swap on samples, or by testing the same samples twice. Although this can increase the accuracy and sensitivity for the estimation of technical variation, it is generally not as important as biological replication is for increasing the sensitivity and power of the analysis. The minimum number of replications that is required for any transcriptomics experiment depends, among others, on the objective of the experiment, the required sensitivity, the type of microarray or sequencing method used, the experimental design, and the number of treatment groups [10]. Measuring gene expression across a time course may also be a powerful way to increase the power of the analysis, as well as providing a means to determine the sequence of action for genes.

For the statistical analysis of transcriptomics data, many different alternatives are available. Most tests developed for microarray data or transcriptome sequencing are essentially modifications of more standard statistical tests [8]. To identify the genes showing differential expression (i.e., differences in expression level) among treatments or conditions, many of the statistical procedures consist of some form of variance analysis and test whether the variance in expression patterns among treatments or conditions exceeds the variance between biological replicates within a treatment. The most commonly used tests include (modifications of) t-tests, ANOVAs, regression analysis, mixed models and generalized linear models. The modifications for these tests are primarily to increase power for the often small sample sizes, and to avoid violation of the assumptions for the parametric tests, in particular the assumptions of a Normal distribution and independence among measurements. Modifications include methods to shrink variance estimates (using combined information on variance for the large number of measurement on a single sample), permutation approaches and empirical Bayesian methods. Similar to the best choice for the number of biological replicates, the best statistical approach depends on the objective of the experiment, the transcriptomics platform used, the experimental design, the number of treatment groups and the number of replicates per treatment.

Not only statistical significance, but also the magnitude of a change in expression (or the 'fold change') between conditions is often provided, sometimes as an auxiliary for biological significance. Fold changes are typically provided at a log2 scale, so that the fold changes are centred around zero, and a doubling or halving of expression level in the treatment compared to the control would result in an equal deviation from zero. These fold change data can be plotted to visualize the differentially expressed genes, either in relation to the average expression level of that gene (MA-plot, Figure 2a), or in relation to the statistical significance (volcano plot, Figure 2b). It should be realized, however, that fold changes are fickle indicators of biological significance. Firstly, depending on the position and role of a particular gene in a regulatory network (e.g., a central transcription factor, or a direct regulator of transcriptional activity), a small fold change may have large biological implications. Large fold changes could be primarily expected at the margins of these networks, which may involve the final effectors of the response while that may reveal little about the key regulators of the response. Secondly, microarrays typically only detect large fold changes in the intermediate range of expression values. Low levels of expression may be below the detection limit of the array, and background noise or corrections may obscure any changes in the expression of such genes. High levels of expression may result in saturation of the probes, vastly underestimating the actual fold changes. Transcriptome sequencing approaches would not be biased towards these intermediate expression levels, but instead, could suffer from exaggerated fold-change estimates for genes not expressed, or expressed at very low level, in one sample or both samples (when the denominator approaches zero).

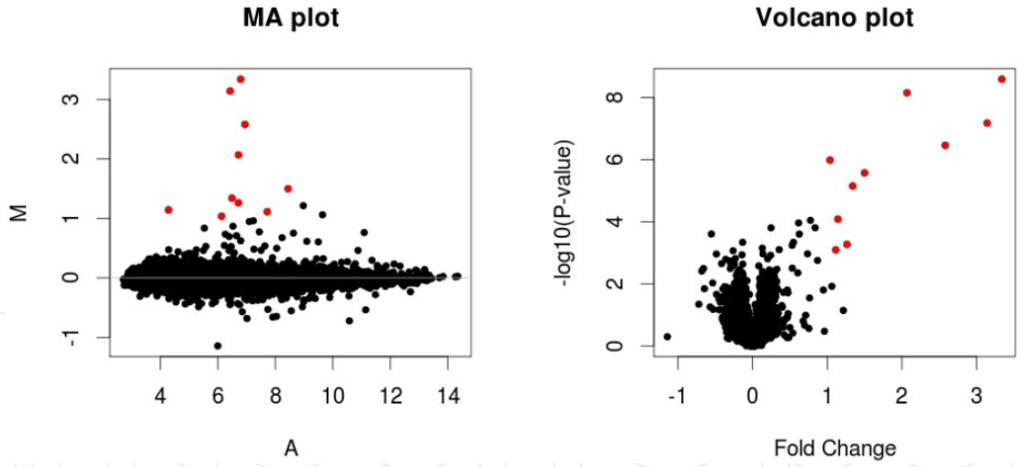


Figure 2. Plots that summarize the fold-change differences in gene expression between two conditions. a) MA plots portray for each gene the average gene expression across the two conditions on the x-axis (A), and the log2 fold change difference in expression between the two conditions on the y-axis (M). b) Volcano plots portray for each gene the log2 fold change in difference of expression between the two conditions on the x-axis (Fold Change) and the statistical significance for the *t*-test on expression measurements between the two conditions on the y-axis (–log10 *P*-value). The presented data is on *Drosophila* larvae 12 hours after being parasitized and control larvae (that had not been parasitized) [5]. The 'outliers' in both plots represent genes that differed in expression between the two conditions. In red are the genes that both scored a *P*-value < 0.001 and had at least a 2-fold change in expression between the two conditions. Applying these combined criteria for assigning significance would exclude several 'outlier' genes with high average expression levels (a) and/or with low p-values (b).

Finally, to determine the genes with significant differences in expression among conditions or treatments, a statistical correction needs to be applied for the large number of statistical tests for each experiment (i.e., multiplicity or multiple testing). In a transcriptomics experiments, several thousands of genes are tested, and each gene is analysed for differences in expression among conditions. In statistics, we normally use a type I error rate of $\alpha = 0.05$, which means that we accept that in 5% of cases where we rejected the null hypotheses (H_0: no differences among conditions) and called something 'significantly' different, the observed difference was purely by chance. When we do not correct the type I error rate while performing thousands of statistical tests (i.e., one for each gene), this would result in hundreds of genes called significantly differentially expressed, while these differences were merely by chance. Genes that are deemed differentially expressed while they are not, are *false positives*. Genes that are deemed not differentially expressed while they are, are *false negatives*. Correcting for false positives in large scale experiments is needed to avoid including many erroneous calls, but it needs to be carefully balanced by controlling for false negatives to ensure optimal sensitivity and accuracy of the analysis.

The typical statistical correction for false positives in non-genomic experiments with multiple testing is a Bonferroni correction, which divides α by the number of statistical tests applied to the data. This approach, however, is often too conservative (i.e., accepting the null hypothesis H_0, while it was false) for the thousands of tests in transcriptomics analyses, and would result in a large number of false negatives. The most widely used correction for multiple testing in transcriptomics analysis is a False Discovery Rate (FDR) correction, which attempts to provide a more even balance between false positives and false negatives. Several FDR approaches exist, but they generally adjust or replace the *P*-value for significance to reflect the likely proportion of false positives among the genes that are called significant. For example, when we would identify 100 genes with an FDR adjusted *P*-value (P_{adj} or *q*-value) of <0.05, we would on average expect less than 5 of these genes to be false positives [11]. The acceptance level for significance used with FDR often ranges from P_{adj} <0.001 to <0.10, depending on the desired sensitivity and accuracy, the sample size (i.e., power) and the estimated numbers of genes with differential expression.

The end result of all pre-processing steps, normalisation, statistical analyses and corrections for false positives is a list or ranking of genes that significantly changed expression in response or relation to the different conditions that were compared. This lists contains potential candidate genes that may be actively involved in the process of interest. However, many genes are also included in the list that are only indirectly associated with the response or process of interest. Moreover, the gene list does not contain *all* the (candidate) genes that are involved in the process, but only these that could be detected by transcriptomics and under the particular experimental conditions (e.g. time points during the response, sample sizes, technological platform) and analysis choices (e.g. normalization approach, acceptance thresholds for significance). Finding gene expression changes in a transcriptomics experiments is not required, nor sufficient, evidence for the function of a gene or its involvement in a biological process. It is, however, a valuable starting point for further analysis.

3. Standard explorations of the gene list

The first inspection of a gene list typically is to link the gene names to what is known, predicted and published about these genes, both in terms of the function of the gene (product), the protein family or protein domains that the gene codes for, and the signal transduction pathways in which it participates. For model species and other species for which the full genomic sequence is available, repositories exist that combine several sources of information on individual genes (for example, see "www.nature.com/scitable/content/Genomics-Databases-744357" for a list of species-specific repositories [12]). The annotation of genes is mostly following a controlled vocabulary or restricted terminology. For functional annotations, Gene Ontology (GO) is a widely used vocabulary. Gene Ontology describes the genes and their products (e.g., the proteins for which a gene codes) within three main Ontology domains: Molecular Function, Biological Process and Cellular Component. Genes can be described at various hierarchical levels using this GO terminology, ranging from broad over-arching themes to very specific descriptions. Descriptions of protein domains are often inferred based on sequence similarity to other organism, for example using the InterPro terminology. Since many proteins are involved in several biological processes or contain more than one functional domain, genes (or gene products) have often different GO annotations across the three GO domains and different IP annotations (Table 1).

Gene Name (symbol)	Gene Ontology Annotation	InterPro Annotation
αPS4	Cellular Component: Integrin complex Biological Process: Cell adhesion Biological Process: Cell-matrix adhesion Biological Process: Heterophilic cell-cell adhesion Molecular Function: Cell adhesion molecule binding Molecular Function: Receptor activity	Integrin alpha chain Integrin alpha beta-propellor Integrin alpha-2 Integrin alpha chain, C-terminal cytoplasmic region, conserved site FG-GAP
lectin-24A	Cellular Component: - Biological Process: Galactose binding Molecular Function: -	C-type lectin C-type lectin fold
Thiolester containing protein II (TepII)	Cellular Component: Extracellular space Biological Process: Antibacterial humoral response Biological Process: Defense response to gram-negative bacterium Biological Process: Phagocytosis, engulfment Molecular Function: Endopeptidase inhibitor activity Molecular Function: Peptidase inhibitor activity	Terpenoid cylases/protein prenyltransferase alpha-alpha toroid Alpha-macroglobulin, receptor-binding Alpha-2-macroglobulin, N-terminal Alpha-2-macroglobulin, N-terminal 2 A-macroglobulin complement component Alpha-2-macroglobulin, conserved site Alpha-2-macroglobulin, thiol-ester bond-forming

Table 1. Examples of gene annotations, using the vocabulary of the Gene Ontology (GO) and InterPro (IP). Annotations are provided for three genes that were differentially expressed during the immune response of *Drosophila* after infection by parasites [5]. The GO annotations describe the function and

process that have been reported for the protein, and the IP annotations describe the protein domains. Genes that are involved in different processes, or coding for proteins with multiple functional domains, may contain a variety of annotations. Many genes, however, are not fully annotated.

The abundance and reliability of annotation information is highly variable among genes and species: some genes are well studied and annotations are solidly supported by empirical evidence, while other genes are not annotated, only partially annotated or annotations are based only on unconfirmed computer predictions or non-traceable author statements. Furthermore, for model organisms the functional annotations have accumulated by the studies of many researchers over long periods, while for non-model organisms or new model organisms, there is often only limited detailed knowledge available. Yet, even for these non-model organisms, various resources exist that enable high level analysis of transcriptomics data based on homologies, such as, for example, the Blast2GO suite [13].

Gene lists from transcriptomics experiments are particularly amenable for enrichment analyses of functional annotations. An enrichment of a particular functional annotation implies that it is represented more often among the gene list members than would be expected by chance alone, based on the proportion of the genes in the genome with that annotation. Multiple interfaces and online tools have been developed for this purpose (e.g., DAVID for large gene lists [14] and Catmap for gene lists that are ranked for significance, but without actually applying a significance threshold cutoff [15]). When the conditions or treatments of interest resulted in a coordinated response in the gene interaction network, the likelihood increases of finding genes with changed expression sharing the same annotation. Such enrichments may be informative for identifying different biological processes or protein families that are associated with, or affected by, a response to the condition or treatment of interest. This may also be informative to identify possible costs or trade-offs that are associated with the response. For example, within the gene list for the response to parasite infection [5], we identified a set of genes involved in puparial adhesion. These genes were expressed at lower levels in the infected larvae at 72 hours after infection, and reflect the delay in development these larvae incurred by investing energy and resources in the immune response.

The list of differentially expressed genes can be compared to other gene lists, which could be derived from other transcriptomics studies, known candidate genes for the process of interest, or any other approach that identified a set of genes associated with a particular condition. Venn diagrams can summarize these gene list comparisons (Figure 3). Reporting how many of the genes were shared with other gene list(s), and how many are unique for each gene list, provides a quick overview of the numbers of genes that may be of particular interest. Sometimes it is the genes that are also present in the other gene list(s) that are of particular interest, for example when multiple sources of evidence are combined or to identify cross-talk between gene interaction networks. Alternatively, one could focus on the unique genes to identify novel candidate genes that had not previously been associated with the process of interest.

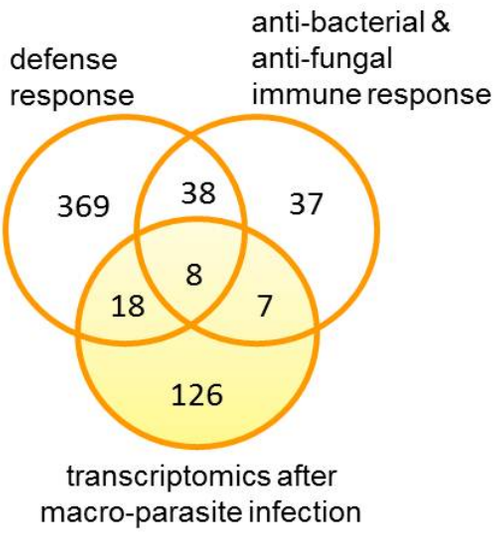


Figure 3. Venn diagram of differentially expressed genes in *Drosophila* larvae after infection by a parasitic wasp, and genes that have been previously implicated in defense responses and anti-microbial immune responses. Infection by a parasitic wasp ('macro-parasite') triggers a cellular immune response that is substantially different from general defense responses and the mostly humoral immune response against bacterial and fungal infections ('micro-parasites'). This is reflected both in the relatively large number of known immunity genes that did not change expression after infection with macro-parasites, and in the large number of differentially expressed genes after macro-parasite infection that had not previously been associated with immunity and defense. *Redrawn with permission after* [5], *first published by BioMed Central.*

When several conditions or time points are included in the experimental design, clustering the genes according to their expression pattern across these conditions or time points allows for identifying groups of genes that responded similarly, and analysing these separately from genes with different behaviour. An enrichment analyses on such groups of genes may identify a common theme to groups with a particular expression profile across the conditions or time points. For example, in our transcriptomics study after infection with macroparasites, we identified groups of genes with a peak in up-regulated expression 1-6 hours after infection, at 6-24 hours after infection, and at 24-72 hours after infection, and groups with down-regulated expression either throughout the time course, or at 72 hours after infection (Figure 4). The first group of genes was enriched for immunity genes (clusters 1 and 2), the second group of genes for proteolysis and serine-type endopeptidases (cluster 12), and the last group in puparial adhesion (cluster 9). These patterns can be used both to get a more detailed profile for the various processes that occur during the response. Additionally, it may serve as a starting point for inferring the functions of unannotated genes. For example, the *Drosophila* genome codes for 201 genes with serine-type endopeptidase activity, which function in development, immunity and various other biological processes. Only 22 of these genes had been functionally annotated with a role in immunity, but unannotated serine-type endopeptidase genes that responded similarly to infection as genes with a functional annotation in immunity or defense could be putatively assigned the same functions [16].

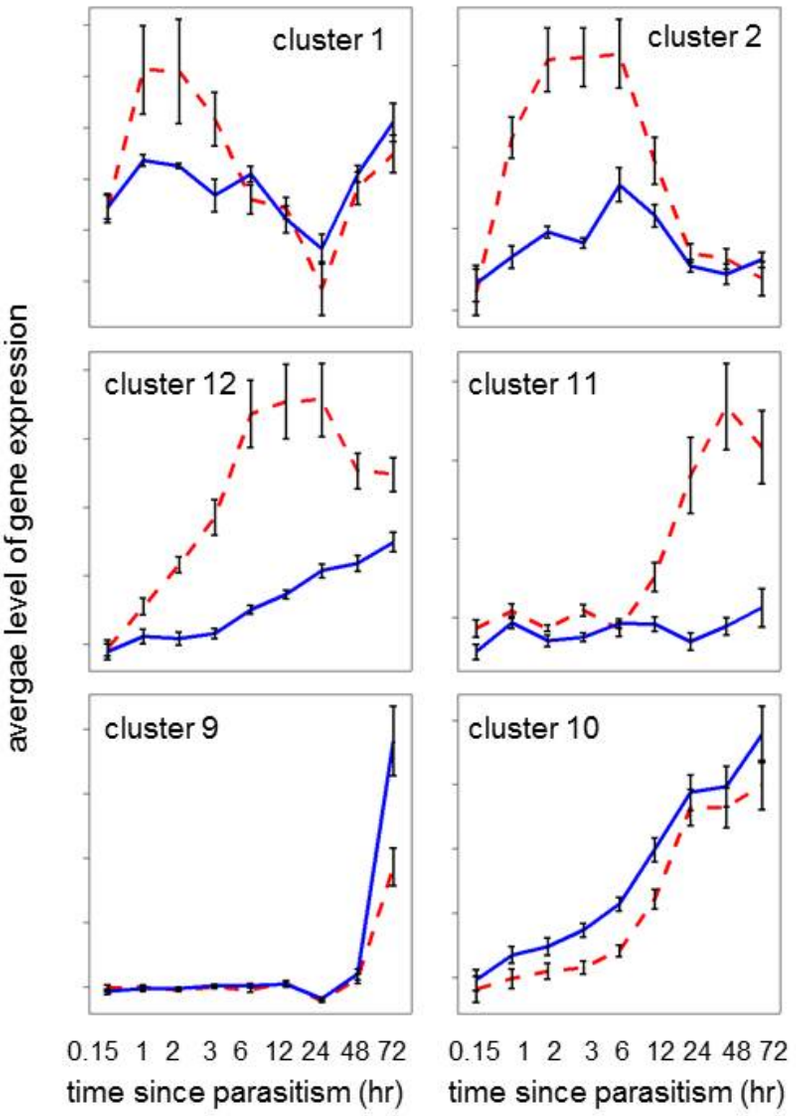


Figure 4. Clustering of genes that show similar expression patterns in *Drosophila* larvae across the 72 hour time course after infection by parasitic wasps. The average expression levels (± standard errors of the mean) for the genes within the clusters (log2 transformed and divided by the median expression level for that gene across all time points) is shown. Dashed red lines represent the gene expression in parasitized larvae and solid blue lines represent the gene expression in control (not parasitized) larvae. *Partially redrawn with permission after* [5] *, first published by BioMed Central.*

In addition to these general approaches for any transcriptomics analysis, regardless of the platform that was used, some additional insights could be gained from using tiling arrays or transcriptome sequencing. Not only the expression level could be determined for each gene, but also alternative isoforms of transcripts, including splice variants and sequence variations (either in the coding regions or in the untranslated regions of the transcripts). In humans, transcriptome sequencing revealed that splicing isoforms from various tissues showed systematic differences, including exon skipping, alternative 3′ or 5′ splice sites, mutually exclusive exons and alternative first or last exons [17]. New methods allow for the quantification of gene expression levels for the individual isoforms, which can improve the

accuracy of expression measures and provide details on the role of the untranslated regions in gene expression regulation [18].

4. Beyond the gene list

The descriptions of the analyses so far have centred on querying repositories containing the functional annotations for genes, to explore what is known on the genes in the gene list and what additional light this may shed on sub-processes, the unannotated genes and associated responses. Yet, many additional resources and genomic databases are available that may be cross-referenced and combined with the gene list, to obtain additional information on these genes and their interactions. Rather than focussing on individual members of the gene list and what is known, these approaches search for emergent properties of the gene list. Especially when the organism that is studied is a model organism for which many sources of additional information are publicly available, there is a large array of possibilities for further analyses.

In addition to searching in specific repositories for functional annotations of genes, the extraction of information on genes and proteins from text documents (e.g., scientific papers) can leap across the boundaries of scientific disciplines. Text mining is the automated extraction of information on proteins or genes from a large literature collection (such as PubMed). It searches for associations between proteins and functional descriptors in the text. These descriptors can be of molecular origin to describe the annotations of the protein (as in the repositories), but also of a physiological, phenotypic or pathological origin to describe the inferences for the organism, or of phylogenomics origin related to the evolution of a gene. Through this additional dimension of information, text mining can help, for instance, to identify associations of the protein to rare mutations that are implicated in diseases, or to protein-protein interactions and regulatory pathways [19]. Text mining is different from a typical literature search, in that it not simply lists the hits, but parses the retrieved information according to further specifications (Figure 5). Various tools are available online (see for example www.ebi.ac.uk/Rebholz/resources.html for an overview).

Physiological responses or the focal tissue of a response to the treatment or condition of interest may also be explored through analysis of the gene list. For some model organisms, a tissue atlas is publicly available that specifies the level of expression for each gene in all tissues and/or developmental stages (e.g., FlyAtlas, Human Atlas Suite and eMouse Atlas). A large fraction of genes in the genome are not expressed homogeneously throughout the body, but show high specificity for particular tissues [21]. Using this information provides a means to screen for tissues that may contribute disproportionally to the response. For example, when the gene list is enriched for genes that are primarily expressed in a particular tissue (e.g. testes, brain, liver or salivary glands), this could indicate that these tissues are most severely affected or responding to the treatment of interest. Additionally, the atlases have raised an awareness for experimental design in transcriptomics studies: when the transcriptomics responses are localized in a particular (minor) tissue, it is difficult to detect

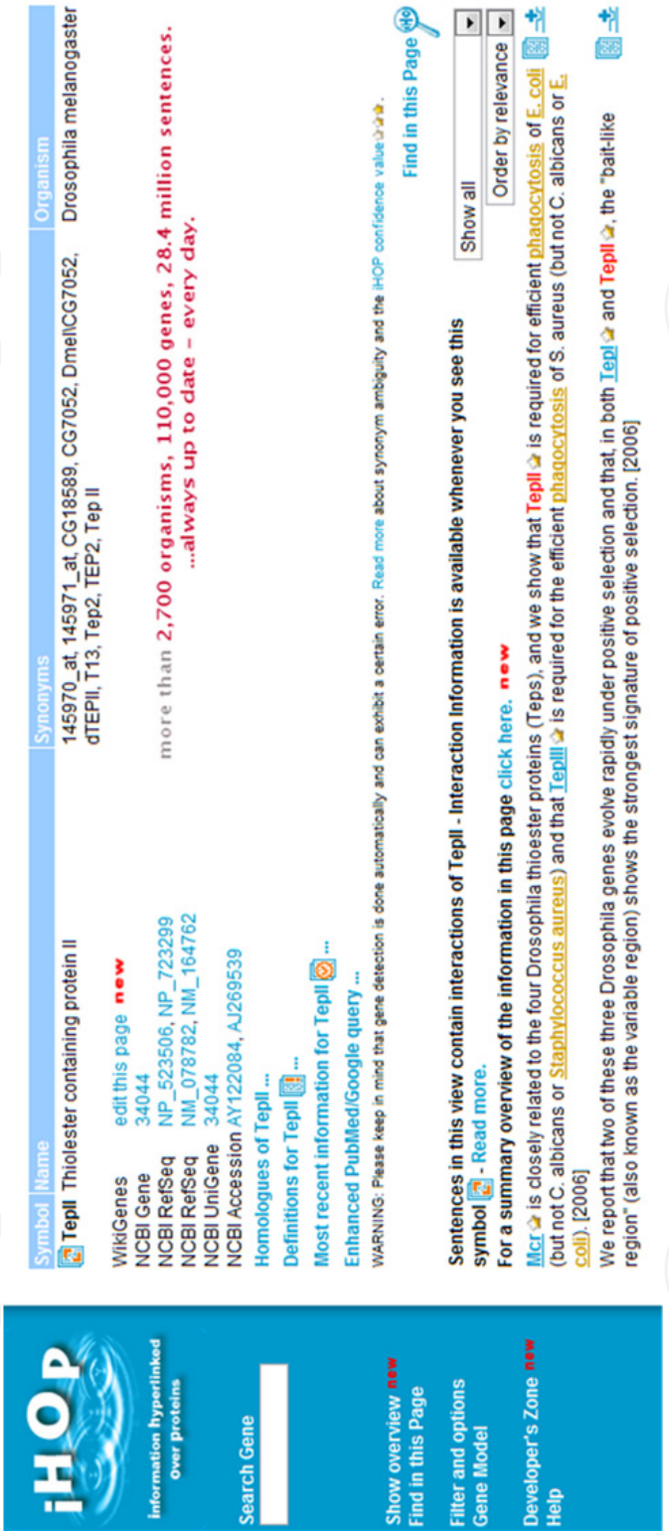
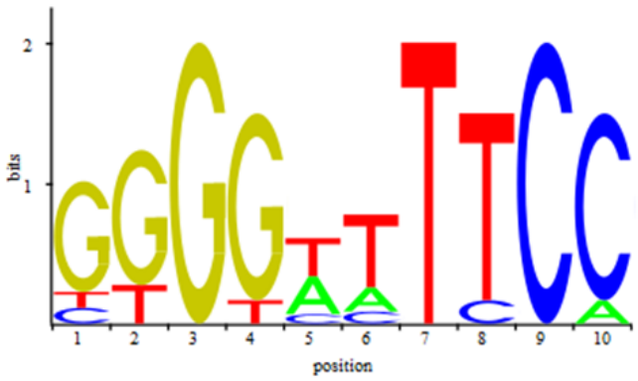


Figure 5. Example of the output from a text mining tool, iHOP [20], for one of the genes that was differentially expressed in *Drosophila* larvae after parasite infection. The functional annotations for the same gene, TepII, are summarized in Table 1. The text mining tool provided additional information on the evolution of the gene through information on related genes (paralogs) and domains of the gene that show signs of positive selection. *Screenshot retrieved from "iHOP - http://www.ihop-net.org/".*

accurate expression differences when the tissue is not studied in isolation. The chances of missing or underestimating the change in gene expression in mixed-tissue comparisons, or inappropriate tissues, are substantial.

To gain insight in the regulatory control of the response to the treatment or condition, a screen for *cis*-regulatory elements in the upstream regions of genes with differential expression may reveal transcription factors and/or co-factors that are involved. These *cis*-regulatory elements can consist of Transcription Factor Binding Motifs (TFBM), promotors, enhancers, silencers and other sequence motifs that regulate the genes [22]. To identify (putative) *cis*-regulatory elements, one could search for known sequence motifs (e.g., TFBMs or promotors) within a specified region upstream of the start codon and in the first intron. Several databases exist (for example, TRANSFAC, RegTransBase and JASPAR) that contain the published TFBMs and promotors. As the binding sites are often relatively short (often 4-12, but up to 30 bases long), and not all positions in the sequence are interacting (strongly) with the transcription factor, some sequence variation in the motif is common. Therefore, the TFBM are usually provided as positional weight matrices, which describe the relative occurrences of each base for each position. This can be converted into a graphical representation, or sequence logo, where the size and order of the stacked letters (A,C,G,T) represents the relative occurrence of the base at that position (Figure 6). These motifs may be investigated for particular genes of interest to obtain a prediction on the Transcription Factor(s) that regulate their expression.



POSITIONAL FREQUENCY MATRIX										
A	0	0	0	0	4	2	0	0	0	1
C	1	0	0	0	1	1	0	1	9	8
G	7	7	9	8	0	0	0	0	0	0
T	1	2	0	1	4	6	9	8	0	0

Figure 6. The Transcription Factor Binding Motif for the NF-ΚB transcription factor *Relish / dorsal* of *Drosophila melanogaster*, depicted as sequence logo and Positional Frequency Matrix. The variation that is commonly found in the binding motif for a transcription factor is incorporated by specifying for each position in the motif the frequency at which each base is recorded. The size of the stacked letters for each position represent the relative occurrence of the respective bases on each position.

Apart from investigating the *cis*-regulatory elements for particular genes of interest, transcriptomics data is also highly suitable to test for over-represented *cis*-regulatory elements across (clusters of) co-expressed genes. This approach can identify groups of genes that are possibly co-regulated by the same Transcription Factor(s). Programs have been specifically developed to screen whether certain known motifs occur more often than you would expect by chance (for example, Clover [23]). These programs can also be extended with custom-made libraries of motifs, to include sequence motifs that could contain yet unidentified *cis*-regulatory elements. These novel motifs could be derived from aligning the upstream sequences of orthologs to identify conserved sequences among related taxa, or through the use of *de novo* motif discovery programs. Alternatively, MotifRegressor searches for any motif that is shared among genes that responded similarly in an expression study [24].

Analyzing the *cis*-regulatory elements in co-expressed genes can be used to start unravelling the genetic architecture of a trait. In our study for the response to parasite infection, we identified seven *cis*-regulatory elements that were over-represented among the differentially expressed genes, using a combination of MotifRegressor and Clover. Three of these motifs were TFBMs for transcription factors that were known to be involved in the immune responses (the GATA-factor *serpent*, the NF-ĸB *Relish/dorsal* and the Janus kinase *Stat92E*), while three others novel motifs were identified that have not yet been associated with any regulatory function. The expression levels of the transcription factor *serpent* was not changed after parasitation, which may appear counter-intuitive as the TFBM was over-represented in differentially expressed genes. Analysing the expression patterns of the clusters of co-expressed genes with the enriched TFBMs, however, and linking these to functional annotations for these groups of genes, suggested that this transcription factor was drawn away from it regular functions in development and metabolism (co-regulated genes with lower expression levels), towards the activation of the immune response (co-regulated genes higher expression levels) [5]. Additionally, we could hypothesize that the novel motifs may also be involved in coordinating the immune response to parasite infection. Using the cisRED database [25] as a first exploration of these novel motifs, two of these motifs were retrieved as a predicted regulatory element in the human genome sequences, including a hit in the upstream region of a known trans-activator of the MHC II complex (ZXDA). Although the functional characterization of the novel motif is still awaiting, these examples illustrate the complex genetic interactions that may coordinate the regulation of a trait.

Not only transcription is regulated through regulatory sequences associated with genes, translation into proteins is also partially coordinated by regulatory sequences. A rich world of small non-coding RNA molecules have been discovered since the start of the genomic era, which added a completely new dimension to the regulation of gene interaction networks [26]. One large class of these non-coding RNAs, the microRNAs, bind to the 3′ untranslated regions (3′ UTRs) of mRNAs, inhibiting their translation by polymerases and targeting the mRNAs for degradation. Several databases exist that link target genes or sequence motifs in the 3′ UTR to specific microRNAs. These tools are accessible through the microRNA database miRBase [27]. Associating microRNAs to the genes in a gene list could be achieved in an analogous manner as the association to the transcription factors: either by searching

for known microRNA binding motifs within the 3'UTRs of differentially expressed genes, or by searching for any over-represented or conserved motifs in the 3'UTRs among the genes in the gene list and trying to associate those to microRNAs.

Another approach to analyse the genetic architecture for a trait is to make use of protein-protein interaction (PPI) network databases. These databases contain the known and predicted protein-protein association network, based on experimental approaches (e.g., two hybrid assays, purification of protein complexes, Chromatin immunoprecipitation (ChIP), etc.) and/or computational methods for predicting protein interactions. A large collection of these PPI databases is publicly available (see for example the Jena Protein-Protein Interaction website ppi.fli-leibniz.de/jcb_ppi_databases.html for an extensive overview). Several web-based tools can be used to analyse and visualize the PPI networks (e.g., STRING [28] and VisANT [29]). Gene lists submitted to these tools are being assembled into inter-connected networks of proteins, based on the PPI databases. The submitted proteins, as well as the proteins that it is known (or predicted) to interact with, form the 'nodes' in the network. All connections between any of these proteins (directly, or through an intermediary protein) are depicted by lines or 'edges' (Figure 7). The topology of these networks describe the frequency distributions of edges per node, and this can reveal whether the network resembles a random assembly of proteins or not [30].

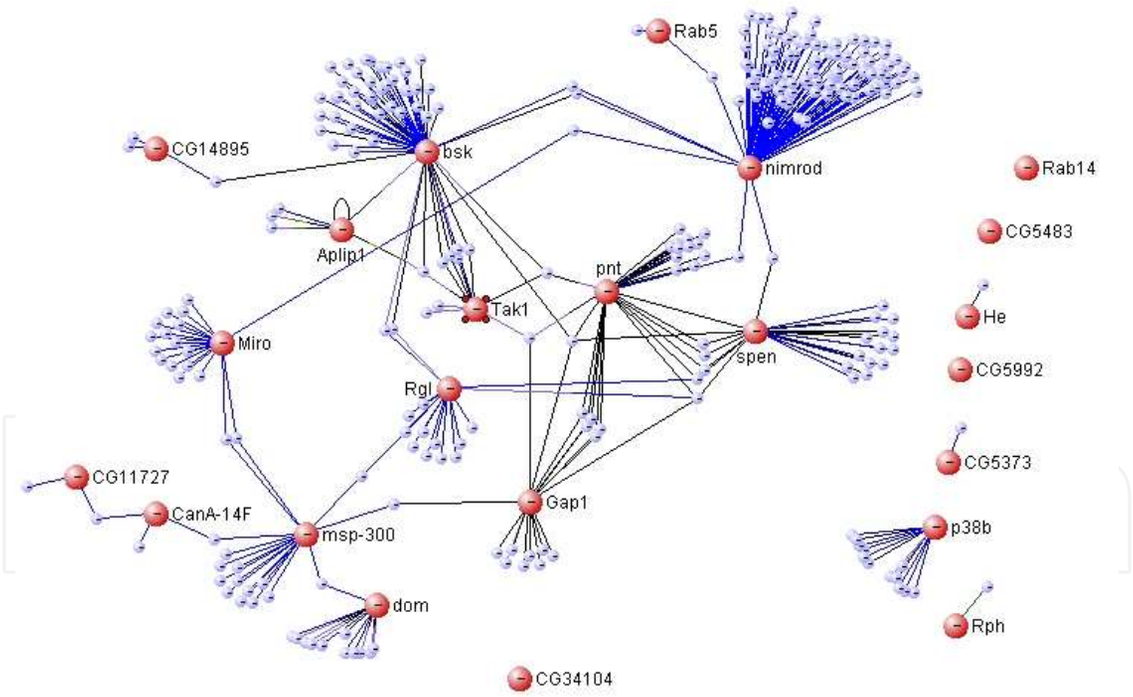


Figure 7. A Protein-Protein Interaction (PPI) network for a subset of the genes involved in the regulation of blood cell proliferation and differentiation in *Drosophila*. The proteins (or 'nodes') are depicted by red or blue circles. The red symbols represent genes with changed expressed in a *Drosophila* strain with an increased immunological resistance against parasites [6]. The known PPIs among these proteins are depicted by lines (or 'edges') between nodes, mostly based on two-hybrid data. Some of the proteins are highly interconnected to other modules of proteins (e.g., pnt, bsk), and these genes can be considered 'hubs' or key coordinators of the changes in expression.

Constructing a PPI network for genes that changed expression in a transcriptomics study may reveal modules of genes that are associated through functional processes, or identify key regulators/modulators to the treatment or condition of interest. Different than with the clustering of genes based on similarity of expression patterns for various conditions or time points, a PPI network will also group genes together that behaved very different transcriptionally, yet may participate in the same signal transduction cascade. We assembled a PPI network for the genes that changed expression between two *Drosophila* lines from the same genetic background, but differing genetically in their resistance to parasites after only five generations of artificial selection. Approximately a third of the nearly 900 genes with changed expression were inter-connected in several modules through an intricate and non-random PPI network [6]. Some genes could be identified within the network that had a central position with a high level of interconnectedness, and these genes may function as a 'hub', as they have the potential to influence the activity of a large number of genes. These 'hub' genes, or their regulators, could be hypothesized to provide targets for selection for increased resistance to parasites, in regulating and coordinating a multitude of phenotypic responses.

Another aspect of the genetic architecture of a trait is its relation to the genome architecture. The genes in a gene list can be mapped to chromosomal positions to search for chromosomal 'hotspots' of differential expression. Transcriptional activity varies for chromosomal domains or regions, and characterizing these patterns may indicate regulatory mechanisms that act on these genes. For example, some chromosomal domains are highly transcribed due to epigenetic mechanisms (e.g., chromatin architectures) that maintain a high activity state, as is seen for heat-shock genes [31]. Such domains under epigenetic control may be recognized by mapping multiple highly expressed genes, or conversely, a complete lack of expression, in the same chromosomal region. Such genomic domains may evolve at a different rate. For instance, the regions around heat-shock genes are more susceptible to insertion by Transposable Elements (i.e., mobile DNA sequences that can translocate themselves within the genome) due to their chromatin architecture, which may lead to a faster accumulation of mutations [32]. Furthermore, some chromosomal domains are highly transcribed in particular tissues only, and the gene arrangements within these domains are highly conserved across taxa [33]. Moreover, chromosomal regions show different expression patterns in healthy tissues compared to cancers [34]. These examples indicate that the physical arrangement of genes within the genome may be a target of evolution, likely due to epigenetic and other regulatory mechanisms that control gene expression of sections of the genome.

Additionally, examining the genomic positions of differentially expressed genes may reveal evolutionary processes that acted on the genes. Strong selection for a particular allele or genomic variant leaves a detectable pattern in the genome, which may be represented by a genomic clustering of genes with changed expression levels. When a particular allele provides an selective advantages to the individual, this locus may be swept through the population. Any allelic variation that is physically linked to this locus (i.e., resides in the nearby chromosomal domain) would be swept through the population as well. One of the

best examples of a strong selective sweep is a mutation in the lactase gene in humans, that confers lactose resistance and is highly common among Europeans. Yet, not only this mutation has spread through the European population, but a region spanning approximately a million bases was swept along as well [35]. Such a sweep can also be detectable in expression assays. In our own studies, we imposed a strong selective sweep for immunological resistance in *Drosophila* against parasites, and mapped the genes with changed expression to the chromosomes. This revealed a part of one chromosome bearing a signature of positive selection [6].

Especially when information is available on sequence variations (i.e., different genotypes, or alleles) among the different biological samples in the experiment, genome-wide association mapping (GWAS) is another option to start unravelling the genetic architecture of a trait. In this approach, the individual variation in sequence is related to the variation in expression by statistical modelling. Using a multiple regression approach, the allelic states at various loci (e.g., whether it has an A, T, C or G at locus x, an insertion or deletion (indel), or inversion) is related to the expression level of each gene. This approach can be applied both when the sequence variation is independently acquired, for example through independent genotyping assays on the same samples, or from the more detailed information that can be extracted from tiling arrays or transcriptome sequencing data. This approach requires large sample sizes to obtain sufficient power and resolution for the statistical modelling, and has been used in a medical context to associate rare mutations with diseases. Causally linking sequence variants to diseases, however, has proved to be daunting [36]. Yet, this approach has been useful in obtaining more basal knowledge on genome functioning, and the relative importance of various sequence variants (e.g., copy number variants (CNVs), Single Nucleotide Polymorphisms (SNPs), small insertions and deletions (indels)) on gene expression variation [37].

5. Conclusions

Transcriptomics analysis has been hugely popular to explore the unknown players in a wide range of biological processes, diseases, traits and responses to stimuli. The technique is extremely powerful as a first step to implicate novel genes and pathways that may be involved or associated with a particular condition. It should be emphasized, however, that a difference in expression *per se* is not sufficient evidence to infer a direct involvement of the gene in the particular process or trait. This is a limitation of the technology, and it urgently requires the development of high-throughput empirical approaches to validate and functionally characterize the large numbers of genes that are putatively of interest. The availability of genome-wide libraries of RNAi stocks to knock down any gene of interest [38], or reference panels of genetic variants with fully sequenced genomes [39] are prime examples of the resources that are needed to follow up on transcriptomics studies. At the same time, the list of genes with potential involvement is certainly not the only information that can be derived from a transcriptomics study. It is especially the information on all the differentially expressed genes, including those that are not directly involved, that provides an exceptional source of information on regulation, correlated responses and the genetic architecture of a trait.

A large number of databases and bio-informatic tools are publically available to explore and annotate the individual genes on the gene list, and more importantly, to analyse the gene list collectively. The latter provides both additional power and a more comprehensive insight in the mechanisms and genetic architecture of a trait. Most traits, diseases and responses to environmental stimuli are highly complex, with environmental factors and genetic networks of interactions that contribute to the trait, disease or response. The factors and genetic network underlying a trait may be elucidated by a combination of bioinformatics approaches, and the emergent properties of such approaches may be more revealing than the search for individual candidates for a trait or process.

Many of the bio-informatic tools that can be applied for these analyses have been made accessible to the research community through the Bioconductor platform (www.bioconductor.org) [40]. This platform is based primarily on the open-source R programming language and runs on all operating systems. A good introduction into this versatile bio-informatic environment has been made available by the Girke lab at the University of California, Riverside through a combination of online manuals (http://manuals.bioinformatics.ucr.edu/). Other freely available, online suites for the analysis of transcriptomics data include Babelomics (http://www.babelomics.org) [41] and Galaxy (http://galaxy.psu.edu/, especially for transcriptome sequencing) [42-44].

The latest development in high-throughput sequencing are opening up new possibilities for the analysis of transcriptomics data. More detailed characterization of transcripts is achievable, and the power of transcriptomics analysis can now also be fully harnessed for organisms without a sequenced genome. Many of the approaches that have been developed for transcriptomics data with microarrays are equally applicable to data from transcriptome sequencing. In that sense, the knowledge-base that has accumulated in the research community in transcriptomics analysis over the past decade will largely remain a valuable resource. The experience and expertise that has been developed in dealing with the limitations and possibilities of analysing microarray data will also be of use while exploring the specific limitations and opportunities that are associated with this new platform. Robust and accurate methods need to be developed fast for the pre-processing, normalizing and analysing of transcriptome sequencing data. This will ensure that the full potential of this new technology can be made accessible to the wide research community.

Author details

Bregje Wertheim
Evolutionary Genetics, Centre for Ecological and Evolutionary Studies,
University of Groningen, Groningen, The Netherlands

Acknowledgement

I thank Eric Blanc and Eugene Schuster for their advice and our valuable discussions on the various bioinformatics approaches in gene expression studies. BW was supported by funding from the Netherlands Organization for Scientific Research (NWO) (Vidi grant no. 864.08.008).

6. References

[1] Hey Y, Pepper SD. (2009) Interesting times for microarray expression profiling. Brief Funct.Genomic Proteomic. 8(3):170-173.

[2] Mockler TC, Chan S, Sundaresan A, Chen H, Jacobsen SE, Ecker JR. (2005) Applications of DNA tiling arrays for whole-genome analysis. Genomics. 85(1):1-15.

[3] Wang Z, Gerstein M, Snyder M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat.Rev.Genet. 10(1):57-63.

[4] Ozsolak F, Milos PM. (2011) RNA sequencing: advances, challenges and opportunities. Nat.Rev.Genet. 12(2):87-98.

[5] Wertheim B, Kraaijeveld AR, Schuster E, Blanc E, Hopkins M, Pletcher SD, Strand MR, Partridge L, Godfray HCJ. (2005) Genome-Wide Gene Expression in Response to Parasitoid Attack in *Drosophila*. Genome Biology. 11(6):R94.

[6] Wertheim B, Kraaijeveld AR, Hopkins MG, Walther Boer M, Godfray HC. (2011) Functional genomics of the evolution of increased resistance to parasitism in Drosophila. Mol.Ecol. 20(5):932-949.

[7] Lemaitre B, Hoffmann J. (2007) The host defense of *Drosophila melanogaster*. Annual Review of Immunology. 25:697-743.

[8] Allison DB, Cui X, Page GP, Sabripour M. (2006) Microarray data analysis: from disarray to consolidation and consensus. Nat.Rev.Genet. 7(1):55-65.

[9] Hansen KD, Irizarry RA, Wu Z. (2012) Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics. 13(2):204-216.

[10] Tsai CA, Wang SJ, Chen DT, Chen JJ. (2005) Sample size for gene expression microarray experiments. Bioinformatics. 21(8):1502-1508.

[11] Storey JD, Tibshirani R. (2003) Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences USA. 100(16):9440-9445.

[12] Lathe W, Williams J, Mangan M, Karolchik D. (2008) Genomic Data Resources: Challenges and Promises. Nature Education. 1(3).

[13] Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talon M, Dopazo J, Conesa A. (2008) High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 36(10):3420-3435.

[14] Huang da W, Sherman BT, Lempicki RA. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37(1):1-13.

[15] Breslin T, Eden P, Krogh M. (2004) Comparing functional annotation analyses with Catmap. BMC Bioinformatics. 5:193.

[16] Shah PK, Tripathi LP, Jensen LJ, Gahnim M, Mason C, Furlong EE, Rodrigues V, White KP, Bork P, Sowdhamini R. (2008) Enhanced function annotations for Drosophila serine proteases: A case study for systematic annotation of multi-member gene families. Gene. 407(1-2):199.

[17] Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. (2008) Alternative isoform regulation in human tissue transcriptomes. Nature. 456(7221):470-476.
[18] Haas BJ, Zody MC. (2010) Advancing RNA-Seq analysis. Nat.Biotechnol. 28(5):421-423.
[19] Krallinger M, Valencia A, Hirschman L. (2008) Linking genes to literature: text mining, information extraction, and retrieval applications for biology. Genome Biol. 9 Suppl 2:S8.
[20] Hoffmann R, Valencia A. (2004) A gene network for navigating the literature. Nat.Genet. 36(7):664-664.
[21] Chintapalli VR, Wang J, Dow JAT. (2007) Using FlyAtlas to identify better Drosophila melanogaster models of human disease. Nat Genet. 39(6):715.
[22] Maston GA, Evans SK, Green MR. (2006) Transcriptional regulatory elements in the human genome. Annu.Rev.Genomics Hum.Genet. 7:29-59.
[23] Frith MC, Fu Y, Yu L, Chen J-, Hansen U, Weng Z. (2004) Detection of functional DNA motifs via statistical over-representation. Nucleic Acids Research. 32.(4):1372-1381.
[24] Conlon EM, Liu XS, Lieb JD, Liu JS. (2003) Integrating regulatory motif discovery and genome-wide expression analysis. PNAS. 100(6):3339–3344.
[25] Robertson G, Bilenky M, Lin K, He A, Yuen W, Dagpinar M, Varhol R, Teague K, Griffith OL, Zhang X, Pan Y, Hassel M, Sleumer MC, Pan W, Pleasance ED, Chuang M, Hao H, Li YY, Robertson N, Fjell C, Li B, Montgomery SB, Astakhova T, Zhou J, Sander J, Siddiqui AS, Jones SJ. (2006) cisRED: a database system for genome-scale computational discovery of regulatory elements. Nucleic Acids Res. 34(Database issue):D68-73.
[26] Zamore P, Haley B. (2005) Ribo-gnome: The big world of small RNAs. Science. 309(5740):1519-1524.
[27] Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. (2008) miRBase: tools for microRNA genomics. Nucleic Acids Res. 36(Database issue):D154-8.
[28] Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ, von Mering C. (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res. 39(Database issue):D561-8.
[29] Hu Z, Hung JH, Wang Y, Chang YC, Huang CL, Huyck M, DeLisi C. (2009) VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. Nucleic Acids Res. 37(Web Server issue):W115-21.
[30] Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL. (2000) The large-scale organization of metabolic networks. Nature. 407(6804):651-654.
[31] Farkas G, Leibovitch BA, Elgin SC. (2000) Chromatin organization and transcriptional control of gene expression in Drosophila. Gene. 253(2):117-136.
[32] Walser JC, Chen B, Feder ME. (2006) Heat-shock promoters: targets for evolution by P transposable elements in Drosophila. PLoS Genet. 2(10):e165.

[33] Yamashita T, Honda M, Takatori H, Nishino R, Hoshino N, Kaneko S. (2004) Genome-wide transcriptome mapping analysis identifies organ-specific gene expression patterns along human chromosomes. Genomics. 84(5):867-875.

[34] Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P, Hermus MC, van Asperen R, Boon K, Voute PA, Heisterkamp S, van Kampen A, Versteeg R. (2001) The human transcriptome map: clustering of highly expressed genes in chromosomal domains. Science. 291(5507):1289-1292.

[35] Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN. (2004) Genetic signatures of strong recent positive selection at the lactase gene. Am.J.Hum.Genet. 74(6):1111-1120.

[36] Marian AJ. (2012) Molecular genetic studies of complex phenotypes. Translational Research. 159(2):64-79.

[37] Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavaré S, Deloukas P, Hurles ME, Dermitzakis ET. (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science. 315(5813):848-53.

[38] Dietzl G, Chen D, Schnorrer F, Su KC, Barinova Y, Fellner M, Gasser B, Kinsey K, Oppel S, Scheiblauer S, Couto A, Marra V, Keleman K, Dickson BJ. (2007) A genome-wide transgenic RNAi library for conditional gene inactivation in Drosophila. Nature. 448(7150):151-156.

[39] Mackay TF, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM, Richardson MF, Anholt RR, Barron M, Bess C, Blankenburg KP, Carbone MA, Castellano D, Chaboub L, Duncan L, Harris Z, Javaid M, Jayaseelan JC, Jhangiani SN, Jordan KW, Lara F, Lawrence F, Lee SL, Librado P, Linheiro RS, Lyman RF, Mackey AJ, Munidasa M, Muzny DM, Nazareth L, Newsham I, Perales L, Pu LL, Qu C, Ramia M, Reid JG, Rollmann SM, Rozas J, Saada N, Turlapati L, Worley KC, Wu YQ, Yamamoto A, Zhu Y, Bergman CM, Thornton KR, Mittelman D, Gibbs RA. (2012) The Drosophila melanogaster Genetic Reference Panel. Nature. 482(7384):173-178.

[40] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, Zhang J. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biology. 5:R80.

[41] Medina I, Carbonell J, Pulido L, Madeira SC, Goetz S, Conesa A, Tarraga J, Pascual-Montano A, Nogales-Cadenas R, Santoyo J, Garcia F, Marba M, Montaner D, Dopazo J. (2010) Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. Nucleic Acids Res. 38:W210-W213.

[42] Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J. (2010) Galaxy: a web-based genome analysis tool for

experimentalists. Current protocols in molecular biology / edited by Frederick M.Ausubel ...[et al.]. Chapter 19:21.

[43] Goecks J, Nekrutenko A, Taylor J, Galaxy Team. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 11(8):R86.

[44] Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A. (2005) Galaxy: A platform for interactive large-scale genome analysis. Genome Res. 15(10):1451-1455.