CIDA: An integrated software for the design, characterisation and global comparison of microarrays

Sabah Khalid[1, 2], Mohsin Khan[1], Alistair Symonds[3], Karl Fraser[2], Ping Wang[3], Xiaohui Liu[2], Suling Li[1, *]

1. Molecular Immunology Group, Microarray Facility, Division of BioSciences, Brunel University, Uxbridge, UB8 3PH, UK.

2. Intelligent Data Analysis Group, Department of Information Systems and Computing, Brunel University, Uxbridge, UB8 3PH, UK

3. Immunology Group, Institute of Cell and Molecular Sciences, Barts and London School of Medicine, London, UK

Abstract

Microarray technology has had a significant impact in the field of systems biology involving the investigation into the biological systems that regulate human life. Identifying genes of significant interest within any given disease on an individual basis is no doubt time consuming and inefficient when considering the complexity of the human genome. Thus, the genetic profiling of the entire human genome in a single experiment has resulted in microarray technology becoming a widely used experimental tool. However, without the use of tools for several aspects of microarray data analysis the technology is limited. To date, no such tool has been developed that allows the integration of numerous microarray results from different research laboratories as well as the design of customised gene chips in a cost-effective manner. In light of this, we have designed the first integrated and automated software called **C**hip **I**ntegration, **D**esign and **A**nnotation (CIDA) for the cross comparison, design and functional annotation of microarray gene chips. The software provides molecular biologists with the control to cross compare the biological signatures generated from multiple microarray studies, design custom microarray gene chips based on their research requirements and lastly characterise microarray data in the context of immunogenomics. Through the relative comparison of related microarray experiments we have identified 258 genes with common gene expression profiles that are not only upregulated in anergic T cells, but also in cells over-expressing the transcription factor Egr2, that has been identified to play a role in T cell anergy. Using the gene chip design aspect of CIDA we have designed and subsequently fabricate immuno-tolerance gene chips consisting of 1758 genes for further research.

The software and database schema is freely available at ftp://ftp.brunel.ac.uk/cspgssk/CIDA/. Additional material is available online at http://www.brunel.ac.uk/about/acad/health/healthres/researchgroups/mi/publications/supplementary/cida

Key words: microarrays, ontology, immunology, gene chip design, immune tolerance, data-fusion.

1 Introduction

Each cell in a multicellular organism carries out its molecular function via the selective expression of genes from the genome. By quantifying this gene expression to reveal a

* Email: Dr Suling Li - Su-ling.li@brunel.ac.uk

transcriptional profile of a cell, scientists can identify activated or repressed genes and begin to understand mechanisms underlying biological processes. Importantly, investigation into the expression profiles can help elucidate the primary transcription events and genetic cascades responsible for healthy and diseased cells. However, the human genome is highly complex consisting of approximately 25,000 genes in which each gene can possess numerous biological functions. Consequently, the resulting transcriptional profile of a cell can range from involving a handful of genes to a few hundred. Either way, how does one identify these genes in the genome? The answer comes in the form of microarray technology, which has the power to distinguish between differentially expressed genes revealing genetic profiles that could be of paramount importance for medical advancement [1, 2, 3]. Once described as being one of the hottest technologies around, microarray technology has now become a standard molecular biology tool utilised by researchers around the world.

However, the technology is rendered futile without the use of software to computationally manage the data in a biologically meaningful manner. In light of this, bioinformaticians are compelled to develop microarray-based analysis tools for statistical manipulation to biological interpretation [4, 5, 6, 7]. Here, we focus on the development and application of the first integrated software called CIDA (**C**hip **I**ntegration, **D**esign and **A**nnotation) offering three unique automated functionalities based on microarray chip annotation, design and comparison. It is imperative to remember that underpinning every microarray experiment is a biological question that a biologist would like to address. In addition, such biological questions are often initiated by researchers working within specialist areas that are interested in annotating their microarray results within the context of their area of expertise. As a result, it is essential for such researchers to have access to software specific for their research area. With a growing interest in functional immunomics following on from functional genomics and proteomics, we have provided immunologists with the ability to functionally annotate their microarray data in the context of immunogenomics in order to understand the integrated behaviour of the immune system. Such a task is currently highly laborious and time-consuming using the Gene Ontology [8, 9], which consists of gene functions relating to multiple disciplines, all of which may not be of immediate interest to the research specific biologist.

Whilst such specialised tools are of great value, it is equally important to have generalised tools that can be used by researchers from all fields. Following the identification of genes of significant interest from a genome wide microarray experiment together with their associated functionalities, researchers often wish to fabricate customised gene chips that relate more to the biological system under investigation. Such chips may not always be commercially available and furthermore, if they are, the sheer cost makes their purchase difficult for academic environments. As a result, it would be highly advantageous if researchers could manipulate and fully utilise their in-house gene sets to extract information from which they could fabricate their own chips, customised to any biological requirement. In light of this, the second set functionality of our software allows the extraction of meaningful data from a given gene set in order to allow researchers design gene chips customised to their research requirements. Information can be extracted from the gene set in several ways depending upon the results that a researcher would like to obtain from the resulting gene chip. Data can be extracted using gene lists related to a particular or several research areas of interest (e.g. oncology, immunology or haematology) collected from external sources (e.g. published literature or commercial chips) for searching within the given gene set containing cDNA clones. This would aim to design a gene chip focussed towards a general area of biology. Alternatively, a researcher may wish to design a custom chip based on several hundred differentially expressed genes as a result of their own microarray experiment. Such a gene chip would relate more to a particular biological system depending upon the question

underlying the microarray experiment. Furthermore, on many occasions, if a laboratory possesses numerous microarray gene sets, genes of interest can be extracted from all the available arrays in order to optimise the resulting gene chip consisting of genes from multiple microarrays. An additional benefit from using customised arrays is that it gives researchers the control to adapt the arrays to their own experimental design and use of controls. Furthermore, since it is not always feasible to purchase array chips for every microarray experiment, this aspect of the software provides an academic environment a cost-effective method for designing multiple chips for several microarray experiments.

Lastly, the flexible nature of microarrays allows them to be exploited in any research field to answer any biological question. Combined with the availability of commercial and custom gene chips, the quantity of microarray experiments that can be performed are endless. As a result, public repositories such as GEO [10, 11] and ArrayExpress [12, 13] have been developed to allow microarray data storage, access and exchange. Ultimately, such databases are of great value to biologists if utilised in the correct manner. If one microarray experiment possesses immense power how much biological knowledge could be gained by combining the results from multiple microarray expression studies? However, without the use of a user-friendly tool to allow biologists to make use of the data via such comparisons, the stored data remains a mine of hidden information. Thus, the final aspect of our software is aimed at meta-profiling via the integration and relative comparison of microarray data from worldwide laboratories. The resulting common genes across related studies can increase the confidence that genes identified as having an important biological role within a disease or in response to a treatment are not by chance alone and furthermore infer a biological relatedness. This in turn provides a more reliable biological insight into the genes and pathways that may be shared in the underlying molecular dysregulation and ultimately common drug targets among related disease states. However, due to the hundreds of interesting genes identified in such microarray studies it is infeasible to locate (by eye) and view expression profiles for the common genes. This process of meta-analysis is further hindered due to the lack of an application for cross comparing microarray studies. In light of this, we have provided an automated solution for molecular biologists within any research field allowing the integration and relative comparison of related microarray studies to generate expression signatures for the common genes.

To summarise, we have provided molecular biologists with solutions to the aforementioned problems during the implementation of our automated software. Each aspect of the software provides a unique benefit for researchers in any area of expertise and furthermore, not only researchers actively involved within microarray research but also those working in biological research in general. More importantly, through the combination of functions aimed at different areas of microarray technology, we have provided a unique software that is powerful, flexible and more crucially, a user-friendly tool for the molecular biologist, ultimate for whom such tools are of immense value.

2 Methods

CIDA has been designed and implemented using the programming language Visual Basic.Net. Processed data from each set functionality procedure using the appropriate underlying database is displayed in the appropriate panel on the graphical user interface (GUI) that is simple and easy to interpret (Figure 1). The front end of the software is a multiple panel interface that is user-friendly interacting with users via menus, mouse clicks and user-input dialogs (Figure 2).

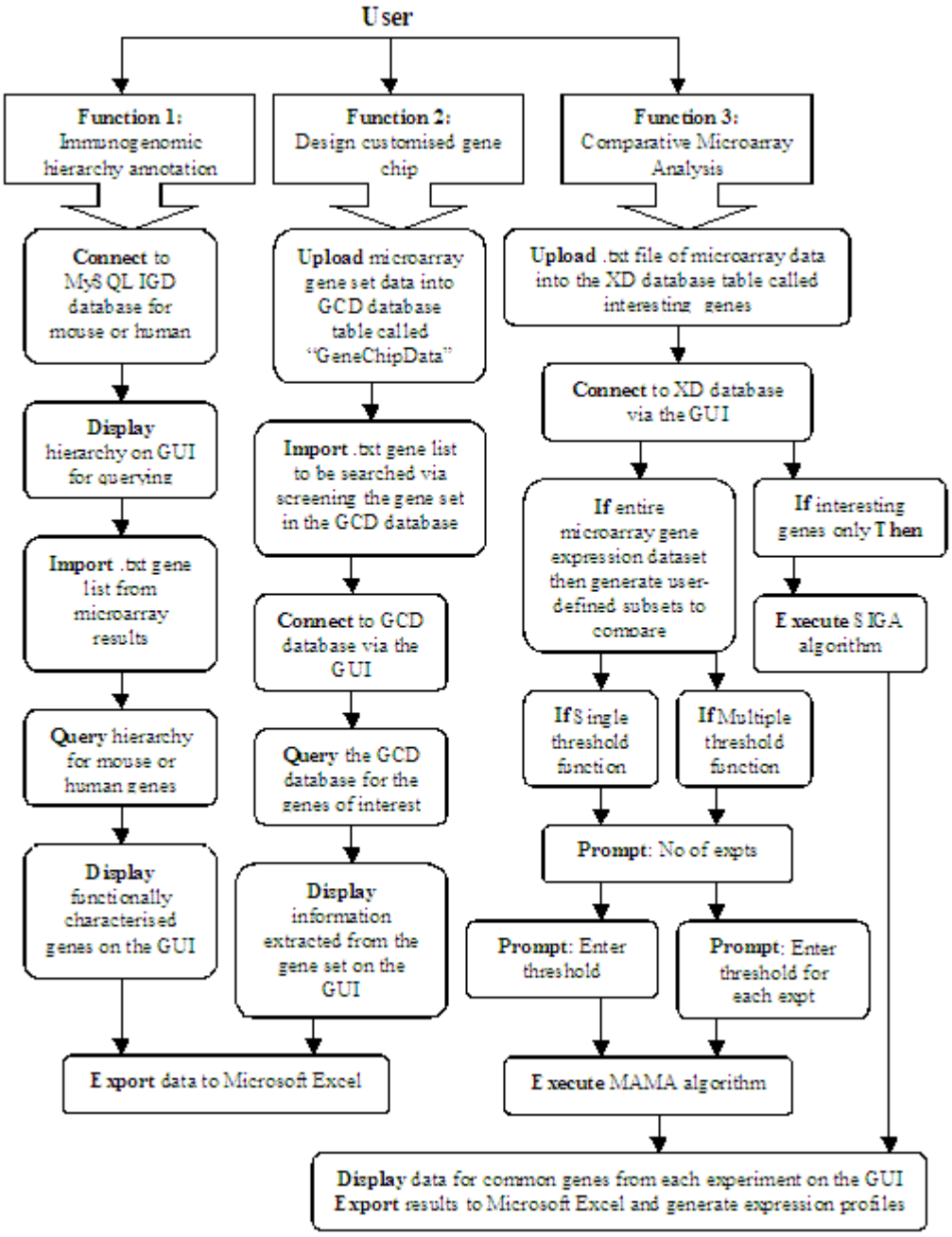


Figure 1: Functionality of CIDA

CIDA carries out several steps during the set functionalities for interpreting, manipulating or analysing microarray gene expression data. Each function of the software requires the microarray data to be imported as a tab delimited text file into the underlying MySQL database before further use. Following data import into the database tables, the databases are queried and results displayed on the GUI and also exported to Microsoft Excel.

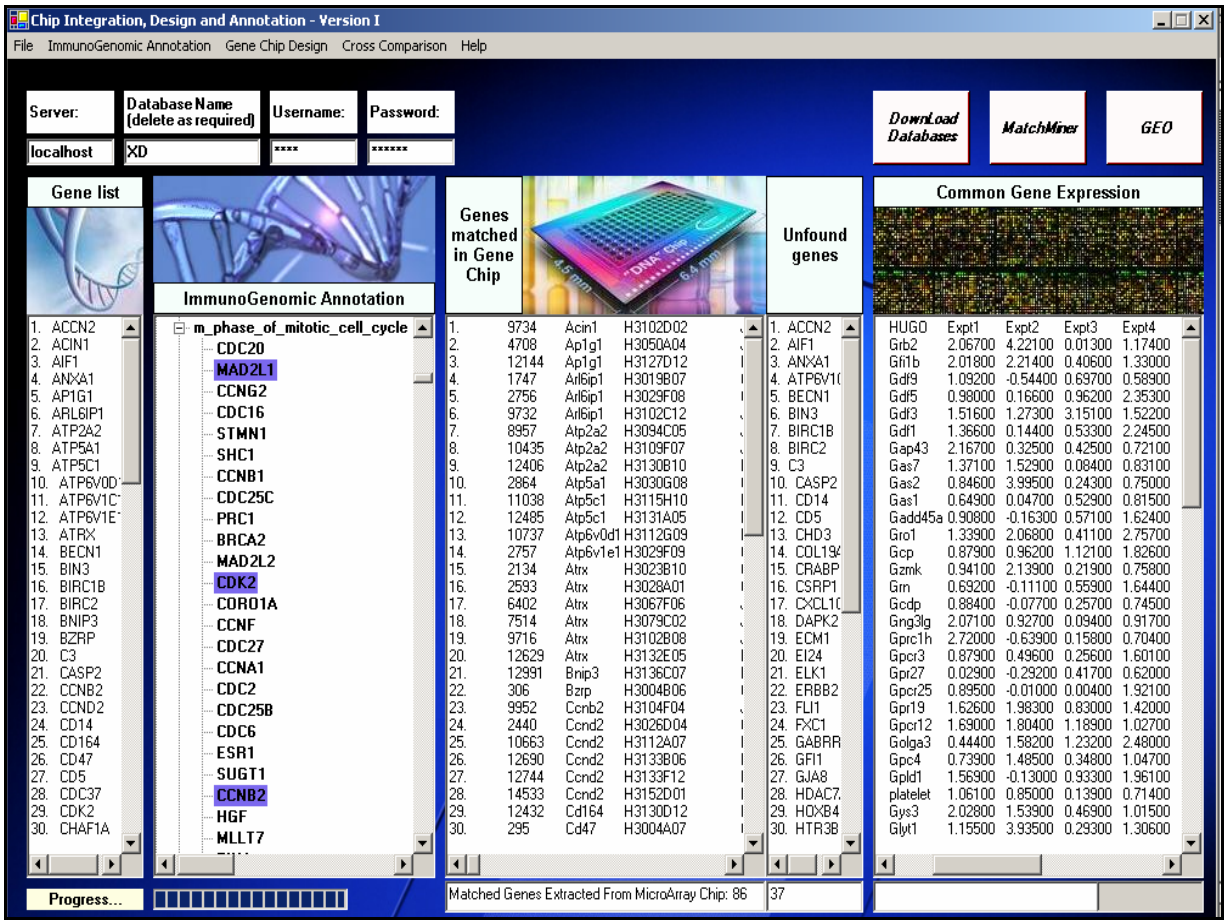


Figure 2: The graphical user interface of CIDA

The multi-panelled user interface of the CIDA software allows microarray data interpretation and integration for any gene expression dataset. The immunogenomic annotation panel displays a hierarchical treeview detailing gene functionality for the interesting genes within a biological system under investigation. The third panel utilises an underlying GCD database to store data from an entire clone gene set and aids the design of in-house gene chips for specific research requirements. Upon providing the software with a gene list of interest, CIDA returns the necessary information required for fabricating subsequent gene chips of interest. The final panel integrates multiple related microarray studies displaying common genes possessing potential common roles within the underlying diseases together with their gene expression profiles.

2.1 System Architecture

The back end of our software uses relational databases implemented in MySQL [14] containing tables to store the required data for the correct functioning of the software. The database schema is highly optimised for rapid querying and extraction of information for large lists of genes and compatible with a MySQL database server no later than 3.23.58. The Visual Basic.Net application and MySQL database communicate with one another via an ODBC MySQL connector available from [14]. The immunogenomic annotation functionality utilises an underlying "ImmunoGenomicDatabase" (IGD) database to generate the hierarchy for subsequent querying. The aspect of the software facilitating the design of a gene chip utilises a database called "GeneChipDesign" (GCD) to store the relevant information pertaining to the user's gene chip. The last set functionality enabling the cross comparison and integration of microarray data uses a database called XD ("ComparisonDatabase") to store microarray data that is to be compared using CIDA.

2.2 Implementation

The IGD consists of a parent, child, gene and relation table storing the ontology data and the relationships that are required to construct a hierarchy allowing functional annotation in the context of immunogenomics (Figure 3a). The database currently consists of 1926 human genes assigned to a total of 1630 biological GO-terms and 2043 mouse genes assigned to 2113 biological GO-terms. The genes were collated from commercial immunology-related gene chips and publications [15, 16, 17], merged into one master list and converted into HUGO id's using the online application called MatchMiner [18, 19]. Their subsequent biological functions were obtained using the Gene Ontology [8, 9] and clustered into 27 functional groups, which in turn are presented as parent nodes in the hierarchy (Supplemental Table 1). The biological functions themselves are represented as child nodes together with the corresponding genes. Unlike the Gene Ontology [8, 9], our customised immunogenomic hierarchy provides researchers with a view of the immunology-related processes underlying their microarray data at the click of a button in a user-friendly manner. Our GCD-database that is used to assist in the design of a tailor-made gene chips contains a table called "GeneChipData" and is structured to hold any in-house microarray gene set provided by the researcher (Figure 3b). The microarray dataset provided must be in a text file format, with a blank first column that is automatically filled with an "id" number when imported into the database table. The second column must represent the gene identities that are to be searched when the user imports a gene list of interest. Subsequent columns can consist of any necessary data required by the user, e.g. gene descriptions, clone positions within respective plate numbers. Once the data is read into the GeneChipData table the information can be used by the software for further manipulation.

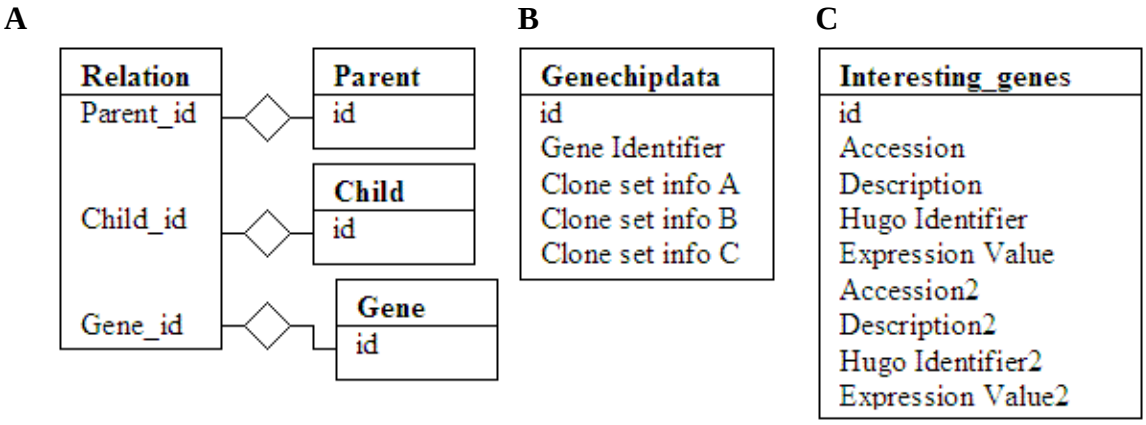


Figure 3: Database schemas for IGD, GCD and XD underlying CIDA

(A) The ImmunoGenomicDatabase (IGD) consists of a parent, child and gene table, which are connected via their primary "id" keys to their respective foreign keys in the relation table. The tables are used to generate a multi-level hierarchy on the GUI. The relationship between the parent and child is one-to-many and many-to-many between child and gene. (B) The GeneChipDesign (GCD) database consists of a single table called Genechipdata to store the information from any given clone gene set. The field "id" is automatically populated when data is imported into the database. The second field must contain the gene identifiers within the clone set. Subsequent fields (maximum of 60) are for storing relevant information for each gene as decided by the user. (C) The ComparisonDatabase (XD) contains a table called interesting_genes allowing the user to import data multiple microarray experiments in the above sequence. The database is integrated with VisualBasic.Net and Microsoft Excel allowing manipulation of the stored data and extraction of biologically meaningful information.

The XD database is employed for the final set functionality of CIDA, allowing the integration and relative comparison of gene expression data from related microarray studies. The XD database comprises a table called "interesting_genes" to store the microarray data that is to be analysed by the software (Figure 3c). The data uploaded into the database must be in a text file format in which the first column is blank, as this will automatically be populated with an "id". The second, third, fourth and fifth columns must contain the information for the first microarray experiment and include the gene identifiers, gene description, Hugo id and expression value, respectively. However, for flexibility the third and fourth columns may contain different information. Subsequent columns for additional microarray experiments must conform to the same format. The gene identifiers that are to be compared across each experiment must be of the same type (e.g. GenBank accession number). In addition, the expression values associated with the genes should represent a ratio of the test to control samples (e.g. median of ratio) and should be consistent throughout each experiment. Most often the genes that are of significant interest to a researchers underlying biological question following a microarray experiment, will be published in the full text or supplementary information of the corresponding article. Thus, this aspect of the software is designed for the relative comparison of such biologically significant genes, identifying the common genes that are of increased biological relevance to each researcher's investigation. Due to the use of the most interesting gene lists that have already been published in the full text of an article the data will have been pre-processed and the gene expression levels averaged for duplicate genes. Thus, following the import of the significantly expressed genes (excluding duplicates) into the underlying XD database, the software executes an underlying SIGA (**si**gnificant **g**ene **a**nalysis) algorithm to analyse the commonality between the gene lists. The algorithm continues until all the gene lists have been compared, presenting one final dataset together with the gene expression values corresponding to each gene list provided. The pseudo-code for SIGA is as follows:

```
Inputs
Interesting gene set 1, Interesting gene set 2 (Maximum of 10)
Outputs
Gene list A
Algorithm
1.    Function: Interesting gene similarity (Interesting gene set 1,
      Interesting gene set 2: Gene list A)
2.    For each item (n) in Interesting gene set 1
3.          If Interesting gene set 1 (n) = Interesting gene set 2 (j)
4.                Gene list A (n) = Interesting gene set 1 (n)
5.          End if
6.    Next Item

7.    Continue until function completed for the total number of interesting
      gene sets imported
8.    End Function
```

In addition to comparing significantly expressed genes from the literature, the software is able to relatively compare expression data for entire gene sets from array chips from different laboratories where the user has the control to decide whether to compare the datasets in their

totality or a subset of the gene expression data. Comparisons involving the entire datasets could help infer potentially valuable insights into related diseases that may have been unnoticeable during the analysis of individual microarray experiments. Following the selection of the conditions from each microarray experiment to compare, the user is required to import the data from a microarray chip (including duplicate genes) into the XD database. During the analysis, CIDA prompts the user for a gene expression significance threshold in order to define the differentially expressed genes from the selected analyses. This allows the user to chose a subset of the expression data for comparison, as genes above the chosen threshold will be further utilised. Subsequently the software computes the average expression levels for duplicate genes and generates a final set of common genes across the biological experiments together with the gene expression profiles presented in a graphical format. Furthermore for added flexibility, the software allows the input of a single gene expression threshold across multiple experiments or different thresholds to account for biological and technical differences between the experiments. In the latter case, the expression thresholds for generating interesting gene sets can be based on those published within the corresponding literature. The process for identifying the common genes and relatively comparing the expression levels to generate common biological signatures is carried out using CIDA's underlying microarray meta-analysis (MAMA) algorithm. The pseudo-code for this algorithm is described below:

Inputs

```
Gene set 1, Gene set 2, threshold (Maximum of 10)
```

Outputs

```
GeneList A, GeneList B, Common 1, FinalAverage 1, FinalAverage 2, Counter,
Calc, Store
```

Algorithm

```
1.    Function: Entire gene set similarity based on a cut-off gene
      expression threshold (Gene set 1, Gene set 2: GeneList A, GeneList B,
      GeneList C, Common 1, Common 2, FinalAverage 1, FinalAverage 2,
      FinalAverage 3)
2.    For each item (n) in Gene set 1
3.    If Gene set 1(n) = > threshold
4.    GeneList A (n) = Gene set 1 (Gene Identifier)
5.    End if
6.    Next

7.    For each item (n) in Gene set 2
8.    If Gene set 2 (n) = > threshold
9.    GeneList B (n) = Gene set 2 (Gene Identifier)
10.   End if
11.   Next

12.   For each item (n) in GeneList A
13.   For each item (n) in GeneList B
14.   If GeneList A (n) = GeneList B (n)
15.   Common 1 (n)=GeneList A (n)
```

```
16.   End if
17.   Next
18.   Next

19.   For each item (n) in Common 1
20.   For each item (n) in Gene set 1
21.   If Common 1(n) = Gene set 1(n)
22.   Counter = Counter +1
23.   Store = Gene set 1 (expression value)
24.   Calc = Calc + Store
25.   FinalAverage 1 = Calc/Counter
26.   End if
27.   Next
28.   Next

29.   For each item (n) in Common 1
30.   For each item (n) in Gene set 2
31.   If Common 1(n) = Gene set 2(n)
32.   Counter = Counter +1
33.   Store = Gene set 2 (expression value)
34.   Calc = Calc + Store
35.   FinalAverage 2= Calc/Counter
36.   End if
37.   Next
38.   Next

39.   Continue until function completed for the total number of gene sets
      imported
40.   End Function
```

2.3 Data accumulation for the immuno-tolerance gene chip

In order to demonstrate the functionality of the gene chip design aspect of the software, we aimed to design a custom in-house immuno-tolerance gene chip for further research purposes by exploiting our NIA 15K Mouse cDNA Clone Gene Set. Genes to be contained on the chip were collated from multiple sources for manipulation using our software. We obtained our initial gene list from an immunology related human gene list from [15] consisting of 2028 human gene symbols. Using the online MatchMiner application [18, 19] we converted the human symbols to mouse gene symbols, resulting in 1875 genes with a known mouse gene symbol output. Furthermore and more importantly, having performed a microarray experiment to understand the biological processes and molecular mechanisms regulating immune tolerance, we identified several sets of differentially expressed genes each with a different gene expression profile that could potentially distinguish between immune tolerance and normal conditions (Supplemental Table 2). For a comprehensive experimental protocol,

array layout and normalised data see the public repository, ArrayExpress [12, 13] accession number E-MEXP-283. This resulted in the addition of a further 900 genes to be contained within the gene chip. Lastly, to further enhance the contents of our immuno-tolerance gene chip we exploited the immunogenomic annotation aspect of our software, to characterise our entire NIA 15K Mouse cDNA Clone Gene Set. This revealed 605 genes with immunology related functions, thus increasing the comprehensiveness of our gene chip. Merging all gene lists and eliminating duplicate genes resulted in a final master list consisting of 2766 genes for further manipulation via exploitation of the gene chip design aspect of our software.

2.4 Preparation of Egr2 transduced cells and in vitro stimulation for microarray analysis

The early growth response transcription factor 2 (Egr2) was inserted into the retroviral pBabe-GFP vector [20], which was subsequently transfected into the Phoenix™ Eco packaging cell line [21] using FuGENE reagent [22], facilitating retroviral production. Transduction into the MF2 T cell line was carried out using RetroNectin (1mg/ml) [23]. In order to compare the gene expression profile elicited as a result of Egr2 over-expression, we carried out a microarray experiment using MF2 cells transduced with or without the Egr2 gene, activated with 25ng/ml PMA and 1120ng/ml Ionomycin for 3 hours as the test sample and normal un-stimulated corresponding cells as the control sample. The subsequent microarray experiment was conducted using a 10K known mouse oligo gene chip (ArrayExpress id: A-MEXP-185) [13] using the RNA extracted from the Egr2-transduced activated, normal activated and normal un-stimulated MF2 cells.

3 Results and discussion

Our Chip integration, design and annotation software is a multi-panelled, multi-functional application for use by any researcher with an interest in microarray research. The GUI displays several panels that can be separated into three functional aspects. The first functionality enables microarray gene expression annotation to further understand the biology of the immune system. The second allows the production of gene chips according to a biologists research requirements based on the manipulation of in-house gene sets and the final function facilitates the cross comparison of related microarray datasets.

The software can be utilised in multiple ways depending upon the results that a researcher would like to obtain. To demonstrate the various functionalities of our software we have designed a novel immuno-tolerance gene chip for future research purposes, based on our in-house NIA 15K Mouse cDNA Clone Gene Set [24]. Secondly, we have characterised the interesting gene expression data that has been generated as a result of our in-house microarray experiment investigating T cell anergy using our immunogenomic hierarchy. Lastly, we have relatively compared that results this microarray experiment with related microarray experiments in order to make further inferences regarding the molecular biology underlying T cell anergy.

3.1 Designing customised gene sets for the development of microarray chips

To develop our immuno-tolerance gene chips, we exploited the “chip design” aspect of our software, which requires the user to connect to the underlying GCD-database and upload their available microarray gene set. Hence, prior to the development of our gene chips we uploaded the entire dataset from our 15K Mouse Clone Gene Set into the GCD-database. The user is then required to import a list of genes for containment within their novel gene chip.

We imported our 2766 immunology-related gene list into the software for screening the gene set. Through this comparison, a total of 1758 genes were identified and displayed on the GUI together with the gene identity, plate number and clone position (Supplemental Table 3).

3.2 Fabrication of immuno-tolerance array chips

The 1758 cDNA clones were collected into nineteen 96-well microtitre plates and subjected to amplification by polymerase chain reaction (PCR). The resulting PCR products are required for printing our immuno-tolerance gene chip ensuring that only good quality and clear positive cDNAs are selected (Figure 4). This resulted in a more specialised version of the 15K Gene Set and the subsequent production of an in-house immuno-tolerance chip containing genes that largely play a functional role within T cell anergy and the immune response. Our high performance benchtop array-spotting robot, Omni Grid Accent TM Microarrayer supplied by Gene Machines [25] within our Microarray Facility was used to print the cDNA PCR products during the fabrication of the gene chips.

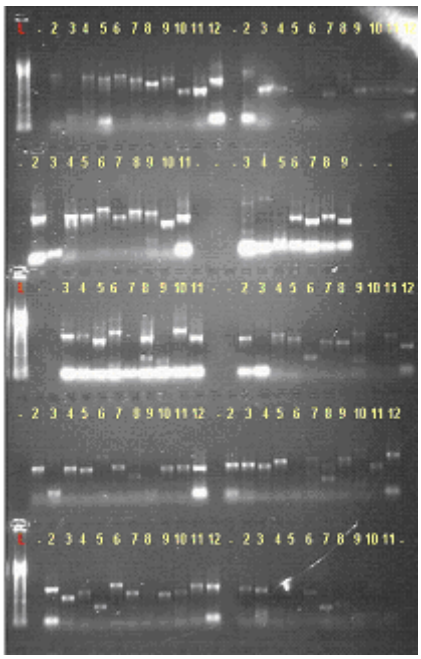


Figure 4: Gene amplification by polymerase chain reaction.

This plate represents a selection of the resulting DNA bands generated after amplifying selected genes from the 15K Mouse Gene Set via PCR. The first column shows a DNA ladder used to allow size estimation of the individual genes.

3.3 Biological processes underlying our designed immuno-tolerance gene chip

In order to determine the biological functions that could potentially be detected following the use of our immuno-tolerance gene chip, we used the functional annotation aspect of our software, to gain further insight into the gene chip. The immunogenomic hierarchy highlights genes of interest that have been imported into the software and matched in the underlying MySQL IGD. A count is displayed alongside each functional category immediately drawing attention to the more significant functions within the chip. The advantage of using this hierarchy is that not only does it display the genes and their respective functions for those provided by the user, but also displays additional genes known to belong to each functional cluster. This gives researchers further knowledge and the opportunity to enhance their genechip with such genes. Using this aspect to analyse our immuno-tolerance gene set

revealed several biological processes consisting of numerous genes within their sub-functions including the cell cycle, cell activation, the central nervous system, cytokine production, cell death, the immune response and signal transduction. (Table 1; Supplementary Table 4). With the intention of designing an immuno-tolerance gene chip for further use in immunological related studies, such function are anticipated and thus increase the confidence for the use of our chip in future research.

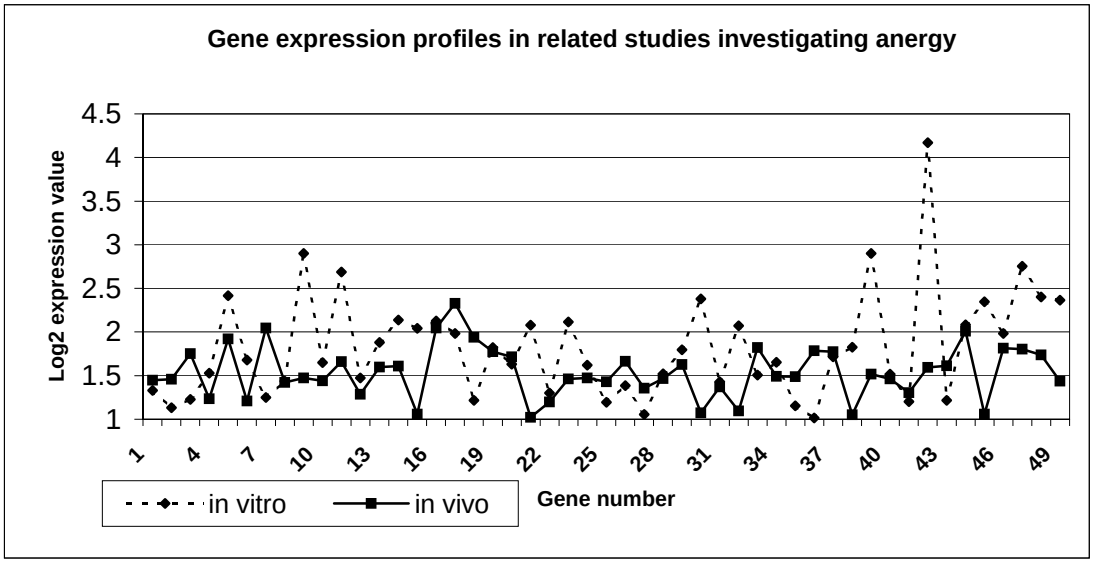
3.4 Identifying genes significantly up-regulated through Egr2 over-expression

The Egr2 transcription factor is known to play a role in the myelination of nerves within the central nervous system [9]. However dysregulation in the myelination process due to a breakdown in the anergic mechanism can result in T cells targeting the myelin basic protein of nerve cells causing their demyelination and the subsequent development of autoimmune diseases such as multiple sclerosis. More importantly, Egr2 has been identified to negatively regulate transcription and is necessary for the full induction of T cell anergy [31, 32]. Thus, further characterisation of this transcription factor is imperative for understanding the mechanisms that regulate T cell anergy. In light of this, our microarray experiment aimed to identify and investigate the genes significantly expressed as a result of Egr2 over-expression. The microarray experiment was performed using an oligonucleotide gene chip (ArrayExpress id: A-MEXP-185) consisting of approximately 10,000 known mouse genes [13], which was scanned and normalised using GenePix Pro [4] for further analysis. The microarray experiment performed using activated MF2 cells over-expressing Egr2 versus normal un-stimulated MF2 cells (Egr2Act) was then combined with the resulting dataset from the microarray experiment using normal activated MF2 cells versus normal un-stimulated MF2 cells (MF2Act). This generated a single combined dataset for further manipulation. Using Acuity 3.1 [5] the gene expression values were averaged for duplicate genes and gene expression values considered unreliable for subsequent analysis were eliminated. This resulted in a dataset consisting of 7229 genes represented by the median of ratio for analysis. To identify the genes significantly expressed as a result of Egr2 over-expression we compared Egr2Act with MF2Act. An initial threshold of =>1.5 was selected in Egr2Act to identify genes significantly expressed when compared with the control resulting in 931 genes. When compared with the corresponding expression values in MF2Act to identify differential gene expression, this gave rise to 19 genes significantly upregulated more than =>1.0 fold in Egr2Act (Supplemental Figure 1). Among these are the genes, Nr1i3 (nuclear receptor subfamily 1) and Dmap1 (DNA methyltransferase 1-associated protein) identified as being negative regulators of DNA transcription [9]. Nr1i3 is a tumour necrosis factor related protein of the nuclear receptor superfamily and more importantly contains a region strikingly similar to the death domains of Tnfr1 and Fas [26]. These death domains are important for the ability of these receptors to induce apoptosis and for allowing Fas to more specifically regulate peripheral tolerance [27]. Alongside Nr1i3, Dmap1 is known to possess DNA transcription repressive activity binding to the transcriptional co-repressor tumour susceptibility gene 101 (Tsg101) [28]. In turn, Tsg101 has been identified to play a role in cell cycle arrest, negative regulation of cell proliferation, and negative regulation of transcription [9]. Furthermore, Dmap1 has been identified to interact with Daxx (Fas death domain associated protein) establishing a repressive transcription complex in the nucleus [29] and in turn Daxx has been identified to bind to Fas initiating the apoptotic pathway. Interestingly, Egr2 has been reported to regulate Fas-l expression [30]. Fas in turn, triggers programmed cell death and more importantly, is essential for T cell homeostasis [30] by eliminating non-functional or autoreactive T cells. Thus, the structural similarity of Nr1i3 with Fas combined with the ability of Dmap1 to interact with Daxx and subsequently Fas, suggests that both Nr1i3 and Dmap1 play a role in apoptosis and the negative regulation of T

cell proliferation. Furthermore, since Nr1i3 and Dmap1 have been identified as a result of the over-expression of the transcription factor Egr2, which is known to play a role in T cell tolerance [31, 32], this suggests that these genes are worthy of further investigation as they may participate in T cell tolerance via regulation by Egr2.

3.5 Common gene expression profiles through the relative comparison of T cell anergy related microarray experiments

Gene expression profiling is particularly valuable within the field of immunology and whilst much research has been conducted to study the gene expression patterns in autoimmune diseases [33, 34, 35], not many microarray studies have been carried out to systemically investigate the underlying biological process called anergy that becomes dysregulated, resulting in autoimmune diseases. By determining the transcriptional mechanisms underlying anergy, which have not yet been fully established we can understand the biological dysregulation in autoimmune diseases. This understanding can be further improved through the combinatorial analysis of microarray experiments investigating similar research areas. Although this is possible for a few genes via literature mining methods it is not a practical solution for genes that are derived via microarray methods where the genes of interest can be numerous. Thus, using the functionality of our software facilitating the comparative analysis of multiple studies we compared our microarray experiment using *in-vitro* Egr2 transduced cells with a previously conducted microarray experiment carried out using *in-vivo* mouse models (Array Express id: E-MEXP-283) [36] investigating the transcriptional profile of T cells in a state of anergy. Through the relative comparison of the entire gene expression arrays our software was able to identify 258 genes commonly upregulated more than 1.0 fold in both *in-vitro* and *in-vivo* experiments (Supplemental Figure 2). Furthermore, a high-level fingerprinting (HLF) algorithm was created using the Matlab language to render the overall relationships between the nodes as a matching percentage. The resulting HLF value of 51% revealed that the two streams react in a similar fashion at least half of the time irrespective of the underlying experimental strategy. These include Ap1s1, which form part of the adaptor protein complex, the SH2 domain-containing adapter protein D (Shd) involved in the intracellular signalling cascade, and Ifnar1, which have been further confirmed by reverse transcriptase-PCR (Figure 5) [36]. Interestingly, Ifnar1 contains a binding site for the transcription factor Stat3 [37, 38], which has been reported to play a critical role in immune tolerance [39]. Furthermore our software has also recognised the common over-expression of both Cacnb3 involved in the T cell receptor signalling pathway [9] and Axl receptor tyrosine kinase (Axl). More significantly, activation of Axl has recently been identified to suppress TNF-alpha production and subsequent inhibition of NF-kappaB-dependent transcription required for T cell proliferation [40]. The upregulation of these genes in both Egr2 transduced cells *in-vitro* and anergic T cells *in-vivo* suggests that they may not only play a significant role in T cell anergy but may also be regulated by Egr2. In addition, the transcription factors Irf3 and Irf7 were also identified to be significantly upregulated in both studies and whilst these genes have been studied in the context of antiviral responses [41] and anti-tumour properties [42] they have not been scrutinised with respect to T cell anergy. The common over-expression of the aforementioned genes identified by our software across the studies suggests these genes are worthy of further investigation with an increased confidence of their involvement and contributory role towards T cell anergy.



Gene No	Hugo ID	Gene No	Hugo ID	Gene No	Hugo ID	Gene No	Hugo ID
1	Taf1a	**14**	Vdac3	**27**	Zfp42	**40**	Ifnab
2	Zfp112	**15**	Tcfap2a	**28**	Zfp2	**41**	Hoxa13
3	Ltb4r2	**16**	Spna1	**29**	Axl	**42**	H2-M9
4	Cited4	**17**	Scya1	**30**	Tnfip6	**43**	Gstt1
5	Grip1	**18**	Mnt	**31**	Tnfaip1	**44**	Fgf7
6	Tnfsf14	**19**	Krtap8-2	**32**	Surf6	**45**	Egr1
7	Irf7	**20**	Ifnar	**33**	Shd	**46**	E2f1
8	Irf3	**21**	Dusp2	**34**	Rgs5	**47**	Ctsh
9	Mapk12	**22**	Cebpb	**35**	Rgs4	**48**	Cacnb3
10	Ifrd1	**23**	Camk2a	**36**	Notch2l	**49**	Ap1s1
11	Hoxd10	**24**	Ap1m2	**37**	Kcnb1		
12	Sitpec-pending	**25**	Anxa7	**38**	Junb		
13	Map2k7	**26**	Zfp67	**39**	Igfbp1		

Figure 5: Genes upregulated in both *in-vitro* and *in-vivo* studies investigating the molecular events underlying T cell anergy

The results from both microarray investigations showed significant expression of numerous common genes. Amongst them we identified Egr1 alongside the transcription factors Irf3 and Irf7 as well as Axl showing potential for further investigation. The corresponding table shows the information for each gene within the graph. For the entire set of common genes, refer to supplemental Figure 2.

Each aspect of our software offers a unique advantage to the molecular biologist incorporating both specificity and universality in terms of microarray data interpretation. Whilst the Gene Ontology and related applications such as FatiGO [43], GoMiner [44] and GenMAPP [45] and is available for the identification of genes across multiple disciplines it is not user-friendly for the molecular biologist concentrating on a specific research area. Although we have offered functional annotation more applicable for the immunologist, this strategy can be adopted for any research field. More importantly, our immunogenomic annotation hierarchy is highly advantageous for the design of further research specific gene chips, as it not only annotates and highlights those genes provided by the user, but also displays other genes known to be involved within the biological functions for the genes of interest, which can subsequently be incorporated into the gene chip. However, this aspect can also be used to annotate gene expression data from any given microarray experiment in the

context of the immunological system. With respect to the customised gene chip design aspect of our software, whilst many commercial microarray services are available (Affymetrix Custom Array Design Program [46], the Keck DNA Microarray Resource [47], Combimatrix [48] and SuperArray [17] they do not provide a public software for the creation and will not be able to process your design without a purchase order for their array chips. Furthermore, due to the immense cost of DNA gene chips it is not feasible to purchase microarrays for every biological question. In addition, unlike Onto-Design [49], our application does not require the biologist to acquire precise biological or molecular gene ontology GO terms for hundreds of genes. Using gene lists from research areas of interest we have provided a method to maximise the benefits from in-house microarray gene sets used by laboratories, which would suit academic environments and non-profit organisations as well as proving cost-effective for many commercial bodies if available chips are not tailored to their scientific requirements. Focussing on the meta-analysis of microarray data whilst methods have been described [50, 51] there is no automated tool that has been implemented and made available for the molecular biologist enabling the relative comparison of gene expression data for both cDNA and Affymetrix microarray chips. We have provided an automated solution for meta-analysis not only for entire array datasets but also for a selection of the dataset based on user-defined gene expression thresholds, allowing the potential uncovering of new information that may not have been identified during the original analyses. The identification of the common genes between related studies suggests their potential common role within the underlying biology of the diseases under scrutiny. Whilst we have provided one method for the meta-analysis of microarray data, the process can be carried out with increased efficiency if the probe names across microarray platforms and different versions of the same microarray gene chip were standardised (i.e. naming the same probe on different gene chips with the same identifier). Furthermore, the use of common tools for generating the expression values and a consensus for representing gene expression (e.g. log ratio) will ensure fewer manipulations before meta-analysis and thus generate more reliable relative gene expression patterns. The MicroArray Quality Control (MAQC) consortium [52] has discussed these issues, however continued investigation into such problem areas and possible collaboration with public repositories to develop more stringent guidelines when accepting microarray data shall ease the process of generating common gene expression profiles.

4 Conclusions

Overall, we believe that CIDA is an attractive application for the scientific community involved in microarray research allowing gene expression data annotation to microarray integration. The capability to firstly, annotate microarray gene expression data or entire array chips encourages focus towards active biological categories containing genes of interest. Secondly, facilitating the generation of customised gene chips for genes of interest involved in targeted pathways and biological processes from in-house microarrays provide researchers with a fast and cost-effective methodology for gene chip designing. Thirdly, the software's ability to make biological inferences through the comparison of global gene expression datasets and relative gene expression analysis for related microarray studies gives biologists immense power to gain further insights into their research question under investigation. As a result, further microarray experiments could be conducted through the in-house design and subsequent fabrication of gene chips using CIDA. To demonstrate the functionality of our software, we have developed an immuno-tolerance microarray chip to explore mechanisms involved in the regulation of autoimmune diseases and cross-compared anergy related microarray experiments. Importantly, however, the flexibility of the software ensures that it can be exploited to design gene chips applicable for any research purpose and furthermore, compare any given related microarray datasets from research laboratories in any field.

5 Availability and requirements

Project name: CIDA

Project Home Page: Databases including the software executable can be accessed from ftp://ftp.brunel.ac.uk/cspgssk/CIDA

Operating system: Tested on Windows 2000 Workstation (SP4) and Windows XP (SP24)

Programming language: Microsoft Visual Basic.Net and MySQL

Other requirements: Microsoft .NET Framework version 2.0 Software Development Kit (SDK) min, MySQL database server no later than 3.23.58, MySQL Connector/ODBC 3.51 and Microsoft Office 2000

6 Acknowledgements

This study was partly supported by grants from the UK Medical Research Council (MRC) (Grant number: G0300520) and the Brunel University Studentship. We thank Daniela Grazio for performing the polymerase chain reaction experiments used to fabricate the immuno-tolerance gene chips.

7 References

[1] Brown, P.O. and Botstein, D. (1999) Exploring the new world of the genome with DNA microarrays. Nat Genetics. 21, 33-7

[2] Lockhart, D.J. and Winzeler, E.A. (2000) Genomics, gene expression and DNA arrays. Nature. 405, 827-836.

[3] Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (2005) Quantitative monitoring of gene expression patterns with complementary DNA microarray. Science. 270, 467-470.

[4] GenePix: http://www.moleculardevices.com/pages/software/gn_genepix_pro.html

[5] Acuity: http://www.moleculardevices.com/pages/software/gn_acuity.html

[6] GoMiner: http://discover.nci.nih.gov/gominer/

[7] Ingenuity Pathways Analysis: http://www.ingenuity.com/

[8] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics, 25, 25-9.

[9] Gene Ontology: http://www.geneontology.org

[10] Edgar, R., Domrachev, M., Lash, A.E. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research, 30, 207-210.

[11] Gene Expression Omnibus: http://www.ncbi.nlm.nih.gov/geo/

[12] Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G.G., Oezcimen, A., Rocca-Serra, P. and Sansone, S.A. (2003) ArrayExpress-a public repository for microarray gene expression data at the EBI. Nucleic Acids Research. 31, 68-71.

[13] ArrayExpress: http://www.ebi.ac.uk/arrayexpress/

[14] MySQL: http://www.mysql.com

[15] Waukau, J., Jailwala, P., Wang, Y., Khoo, H.J., Ghosh, S., Wang, X. and Hessner, M.J. (2003) The design of a gene chip for functional immunological studies on a high-quality control platform. Annals of the New York Academy of Sciences, 1005, 284-287.

[16] BD Biosciences Clontech: http://www.clontech.com/clontech

[17] SuperArray Bioscience Corporation: http://www.superarray.com/home.php

[18] Bussey, K.J., Kane, D., Sunshine, M., Narasimhan, S., Nishizuka, S., Reinhold, W.C., Zeeberg, B., Ajay, W. and Weinstein, J.N. (2003) MatchMiner: a tool for batch navigation among gene and gene product identifiers. Genome Biology. 4, R27.

[19] MatchMiner: http://discover.nci.nih.gov/matchminer

[20] Addgene: http://www.addgene.org/pgvec1

[21] Orbigen: http://www.orbigen.com

[22] Biocompare: http://www.biocompare.com

[23] Takara: http://www.takarabiousa.com

[24] National Institute on Aging: http://www.nia.nih.gov

[25] Genomic Solutions: http://www.genomicsolutions.com/showPage.php

[26] Brojatsch, J., J. Naughton, M. M. Rolls, K. Zingler, and J. A. Young. (1996) CAR1, a TNFR-related protein, is a cellular receptor for cytopathic avian leukosis-sarcoma viruses and mediates apoptosis. Cell 87, 845-855

[27] Chang H.Y., Yang, X. and Baltimore, D. (1999) Dissecting Fas signaling with an altered-specificity death-domain mutant: requirement of FADD binding for apoptosis but not Jun N-terminal kinase activation. Proc Natl Acad Sci U S A, 96, 1252-6.

[28] Rountree, M,R., Bachman, K.E. and Baylin, S.B. (2000) DMAP1 has intrinsic transcription repressive activity, and binds to the transcriptional co-repressor TSG101: Nat Genetics, 25, 269-77.

[29] Muromoto, R., Sugiyama, K., Takachi, A., Imoto, S., Sato, N., Yamamoto, T., Oritani, K., Shimoda, K. and Matsuda, T.J. (2004) Physical and functional interactions between Daxx and DNA methyltransferase 1-associated protein, DMAP1. Immunol, 172, 2985-93.

[30] Droin, N.M., Pinkoski, M.J., Dejardin, E., Green, D.R. (2003) Egr Family Members Regulate Nonlymphoid Expression of Fas Ligand, TRAIL, and Tumor Necrosis Factor during Immune Responses. Molecular and Cellular Biology 23, 7638-7647

[31] Harris, J.E., Bishop, K.E., Phillips, N.E., Mordes, J.P., Greiner, D.L., Rossini, A.A. and Czech, M.P. (2004) Early Growth Response Gene-2, a Zinc-Finger Transcription Factor, Is Required for Full Induction of Clonal Anergy in CD4+ T Cells. The Journal of Immunology, 173, 7331-7338.

[32] Safford, M., Collins, S, Lutz., M.A, Allen., A., Huang, C., Kowalski, J., Blackford, A., Horton, M.R., Drake, C., Schwartz, R.H. and Powell, J.D. (2005) Egr-2 and Egr-3 are negative regulators of T cell activation. Nature Immunology, 6, 472-480.

[33] Lock, C., Herman, S.G., Pedotti, R., Brendolan, A., Schadt, E., Garren, H., Langer-Gould, A., Strober, S., Cannella, B., Allard, J., Klonowski, P., Austi, A., Lad, N., Kaminski, N., Galli, S.J., Oksenberg. J.R., Raine, C.S., Heller, R. and Steinman, L. (2002) Gene-microarray analysis of multiple sclerosis lesions yields new targets validated in autoimmune encephalomyelitis. Nature medicine, 8, 500- 8.

[34] Baechler, E.C., Batliwalla, F.M., Karypis, G., Gaffney, P.M., Ortmann, W.A., Espe, K.J., Shark, K.B., Grande, W.J., Hughes. K.M., Kapur, V., Gregersen, P.K. and Behrens, T.W. (2003) Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus. Proceedings of the National Academy of Sciences of the United States of America, 100, 2610-2615.

[35] Wilson, K.H., Eckenrode, S.E., Li, Q.Z., Ruan, Q.G., Yang, P., Shi, J.D., Davoodi-Semiromi, A., McIndoe, R.A., Croker, B.P. and She, J.X. (2003) Microarray analysis of gene expression in the kidneys of new- and post-onset diabetic NOD mice. Diabetes, 52, 2151-9.

[36] Anderson, P.O., Manzo, B.A., Sundstedt, A., Minaee, S., Symonds, A., Khalid, S., Rodriguez-Cabezas, M.E., Nicolson, K., Li, S., Wraith, D.C. and Wang, P. (2006) Persistent antigenic stimulation alters the transcription program in T cells, resulting in antigen-specific tolerance. European Journal of Immunology, 36, 1374-85.

[37] Pfeffer, L.M., Mullersman, J.E., Pfeffer, S.R., Murti, A., Shi, W. and Yang, C.H. (1997) STAT3 as an adapter to couple phosphatidylinositol 3-kinase to the IFNAR1 chain of the type I interferon receptor. Science, 276, 1418-20.

[38] Benkhart, E.M., Siedlar, M., Wedel, A., Werner, T., Ziegler-Heitbrock, H.W. (2000) Role of Stat3 in lipopolysaccharide-induced IL-10 gene expression. Journal of Immunology, 165, 1612-7.

[39] Cheng, F., Wang, H.W., Cuenca, A., Huang, M., Ghansah, T., Brayer, J., Kerr, W.G., Takeda, K., Akira, S., Schoenberger, S.P., Yu, H., Jove, R. and Sotomayor, E.M. 2003. A critical role for Stat3 signaling in immune tolerance. Immunity, 19, 425-36.

[40] Sharif, M.N, Sosic, D., Rothlin, C.V., Kelly, E., Lemke, G., Olson, E.N. and Ivashkiv, L.B.J (2006) Twist mediates suppression of inflammation by type I IFNs and Axl. Experimental Medicine, 203, 1891-901.

[41] Doyle, S., Vaidya, S., O"Connell, R., Dadgostar, H., Dempsey, P., Wu, T., Rao, G., Sun, R., Haberland, M. and Modlin, R. (2002) IRF3 Mediates a TLR3/TLR4-Specific Antiviral Gene Program. Immunity, 17, 251-263

[42] Romieu-Mourez, R., Solis, M., Nardin, A., Goubau, D., Baron-Bodo, V., Lin, R., Massie, B., Salcedo, M., Hiscott, J. (2006) Distinct roles for IFN regulatory factor (IRF)-3 and IRF-7 in the activation of antitumor properties of human macrophages. Cancer Research, 66, 10576-85.

[43] Al-Shahrour, F., Diaz-Uriarte R. and Dopazo, R. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics 20, 578-580.

[44] Zeeberg, B.R., Feng, W., Wang, G., Wang, M.D., Fojo, A.T., Sunshine, M., Narasimhan, S., Kane, D.W., Reinhold, W.C., Lababidi S., Bussey, K.J., Riss, J., Barrett, J.C. and Weinstein, J.N. (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biology 4, R28.

[45] Doniger, S.W., Salomonis, N., Dahlquist, K.D., Vranizan, K., Lawlor, S.C. and Conklin, B.R. (2003) MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. Genome Biology, 4, R7.

[46] Affymetrix Custom Array Design: http://www.affymetrix.com/products/arrays/custom_design/index.affx

[47] Keck DNA Microarray Resource: http://keck.med.yale.edu/dnaarrays/custom.htm

[48] Combimatrix: http://combimatrix.com

[49] Draghici, S., Khatri, P., Bhavsar, P., Shah, A., Krawetz, S.A. and Tainsky, M.A. (2003) Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. Nucleic Acids Research, 31, 3775-81.

[50] Rhodes, D.R., Barrette, T.R., Rubin, M.A., Ghosh, D. and Chinnaiyan, A.M. (2002) Meta-analysis of microarrays: Interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. Cancer Research, 62, 4427-4433.

[51] Choi, J.K., Choi, J.Y., Kim, D.G., Choi, D.W., Kim, B.Y., Lee, K.H., Yeom, Y.I., Yoo, H.S., Yoo, O.J. and Kim, S. (2004) Integrative analysis of multiple gene expression profiles applied to liver cancer study. FEBS Letters, 565, 93-100.

[52] Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, de Longueville F, Kawasaki ES, Lee KY, Luo Y, Sun YA, Willey JC, Setterquist RA, Fischer GM, Tong W, Dragan YP, Dix DJ, Frueh FW, Goodsaid FM, Herman D, Jensen RV, Johnson CD, Lobenhofer EK, Puri RK, Schrf U, Thierry-Mieg J, Wang C, Wilson M, Wolber PK, Zhang L, Amur S, Bao W, Barbacioru CC, Lucas AB, Bertholet V, Boysen C, Bromley B, Brown D, Brunner A, Canales R, Cao XM, Cebula, T.A., Chen, J.J., Cheng, J., Chu, T.M., Chudin, E., Corson, J., Corton, J.C., Croner, L.J., Davies, C., Davison, T.S., Delenstarr, G., Deng, X., Dorris, D., Eklund, A.C., Fan, X.H., Fang, H., Fulmer-Smentek, S., Fuscoe, J.C., Gallagher, K., Ge, W., Guo, L., Guo, X., Hager, J., Haje, P.K., Han, J., Han, T., Harbottle, H.C., Harris, S.C., Hatchwell, E., Hauser, C.A., Hester, S., Hong, H., Hurban, P., Jackson, S.A., Ji, H., Knight, C.R., Kuo, W.P., LeClerc, JE., Levy, S., Li, Q.Z., Liu, C., Liu, Y., Lombardi, M.J., Ma, Y., Magnuson, S.R., Maqsodi, B., McDaniel, T., Mei, N., Myklebost, O., Ning, B., Novoradovskaya, N., Orr, M.S., Osborn, T.W., Papallo, A., Patterson, T.A., Perkins, R.G., Peters, E.H., Peterson, R., Philips, K.L., Pine, P.S., Pusztai, L., Qian, F., Ren, H., Rosen, M., Rosenzweig, B.A., Samaha, R.R., Schena, M., Schroth, G.P., Shchegrova, S., Smith, D.D., Staedtler, F., Su, Z., Sun, H., Szallasi, Z., Tezak, Z., Thierry-Mieg, D., Thompson, K.L., Tikhonova, I., Turpaz, Y., Vallanat, B., Van, C., Walker, S.J., Wang, S.J., Wang, Y., Wolfinger, R., Wong, A., Wu, J., Xiao, C., Xie, Q., Xu, J., Yang, W., Zhang, L., Zhong, S., Zong, Y. and Slikker, W. Jr. (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nature Biotechnology, 24, 1151-61.