Chapter 4
Expression profiling analyses of siRNA knockdowns of the SCL erythroid complex

4.1 Introduction

As discussed in Chapter 1, in order to identify downstream targets of transcription factors, one of the key analyses is to identify gene expression changes which occur when you perturb the function of a transcription factor of interest in a biological system. The siRNA knockdown studies described in Chapter 3 provide a means for perturbing the function of transcription factors of interest. With the characterisation of siRNAs for each transcription factor in the SCL erythroid complex in time-course experiments, the optimal time points for subsequent perturbation studies were determined. Thus, further analyses to identify downstream target genes using microarray gene expression analyses are described in this chapter.

4.1.1 Information generated using expression profiling of perturbation of transcription factors

A. Direct and indirect targets

Studying where transcription factor binds in the genome only allows us to determine the direct target genes they regulate - these are referred to as the primary targets of a particular transcription factor. However, in complex transcriptional pathways or networks, regulation can be achieved at many levels. For example, one transcription factor may regulate another transcription factor, and in turn, this transcription factor may regulate a third, and so on. Studying the direct binding by a transcription factor only reveals the first level of interactions between the transcription factor and its targets. Whole genome gene expression profiling of a transcription factor perturbation, on the other hand, enables us to identify both direct target genes regulated by the transcription factor and as well as other downstream genes regulated at subsequent levels (so-called indirect targets) (Figure 4.1). This is because perturbations at any one point in the network can affect the entire cascade of transcriptional events occurring further down the network of interactions.

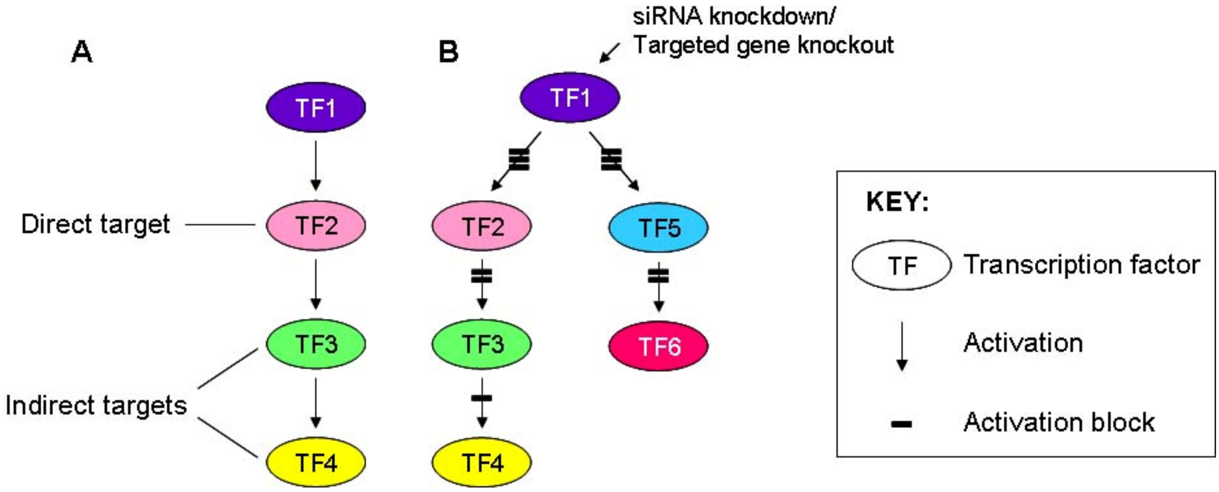


Figure 4.1. Direct and indirect target genes regulated by transcription factors. Panel A: illustration of direct and indirect target genes regulated by transcription factor 1 (TF1). TF2 is the direct target gene of TF1 while TF3 and TF4 are indirect target genes regulated by TF1. Panel B: changes in regulation of direct and indirect target genes when TF1 is silenced by siRNA knockdown or targeted gene knockout. The dash illustrates an activation block and the number of dashes describes the degree of activation block after TF1 is silenced. Typically, the activation block of the direct target genes is the highest (as demonstrated by 3 dashes) whilst that of the indirect target genes is lower (as demonstrated by 1-2 dashes).

B. Mode of regulation

Transcription factor-binding studies such as ChIP-on-chip allow us to study where the transcription factor binds but it does not directly provide information on how this binding event is impinging on the expression of its target gene. However, expression profiling allows one to determine whether a target gene is being activated or repressed by the transcription factor binding event, or whether the binding of the transcription factor has no immediate effect on gene expression. In the case of the latter, the binding of a transcription factor to the regulatory regions of genes may not induce or suppress the expression of a target gene – quite often, the binding of a transcription factor results in a “poised” state of the target for activation or repression later in a developmental programme, when other transcription factors or chromatin-remodelling factors required for regulation are expressed (Chapter 1, Section 1.1.2.5).

4.1.2 Expression profiling studies of the SCL erythroid complex in literature

Several studies have addressed the regulation of SCL and GATA1 target genes in high-throughput assays using expression microarrays. Palomero et al. (2006) delineated downstream targets of SCL in T-cell acute lymphoblastic leukaemia (T-ALL) where SCL is over-expressed due to translocation (Chapter 1, section 1.4.2.1 F) (Palomero et al., 2006). Genome-wide expression profiles of SCL-expressing and non-expressing human T-ALL samples were compared using Affymetrix U133 arrays to identify putative target genes induced by SCL. Lin and Aplan (2007) studied the changes in expression in the mouse genome in thymic tumors from precursor T-cell lymphoblastic lymphoma/leukaemia (pre-T LBL) derived from transgenic mouse overexpressing SCL, LMO1 and NHD13 (Lin and Aplan, 2007). In a very recent study by Landry et al. (2008), a Nimblegen mouse 60-mer oligonucleotide expression microarray platform was used to study the changes in expression after the reintroduction of SCL into SCL$^{-/-}$ mouse yolk sac. This study identified RUNX1, a transcription factor required for definitive haematopoiesis (Landry et al., 2008), as a target of SCL. Welch et al. (2004) studied the expression changes in a sub-set of mouse genes using an Affymetrix GeneChip array before and after the induction of GATA1 expression in the GATA1-null erythroblast cell line G1E-ER4 (Welch et al., 2004). A number of genes were identified which were either up-regulated or down-regulated and both rapid and delayed responses were demonstrated. Affymetrix mouse expression arrays were also used to profile the expression patterns of wild type and GATA1-deficient murine megakaryocytes (Muntean and Crispino, 2005).

While the studies mentioned above described the role of SCL and/or GATA1 in leukaemia, early haematopoiesis or myeloid cells, none of them addressed the role of these transcription factors in regulating genes during erythroid development. In fact, few well characterised downstream targets genes of SCL and GATA1 in erythroid cells have been described in the published literature (Chapter 1, section 1.4.2.1 and 1.4.2.2). Furthermore, downstream targets of E2A, LMO2 and LDB1 in erythroid cells have thus far not been reported. Therefore, genome-wide scale analyses of the five transcription factors in the SCL erythroid complex studied here are necessary in order to have a more complete understanding of their target gene repertoire and roles in gene expression during erythroid development.

4.1.3 The Affymetrix GeneChip expression array

Many methods can be used to study the expression of genes as summarised in Chapter 1, section 1.3.2. Depending on the scale and accuracy required for a particular experimental system, these methods have different strength and weaknesses. For the study of downstream regulation by a particular transcription factor during perturbation, analyses by quantitative PCR or other low-throughput methods can be time-consuming and they often require *a priori* knowledge of the genes of interest. Thus, some important target genes may be excluded in the analyses. Therefore, for identifying targets of transcription factors, genome-wide analyses are desirable because they provide unbiased views of gene expression programmes. To this end, genome-wide profiling by microarrays is a rapid method to study all possible gene expression outputs (depending on the genome coverage of the microarray) - although there can still be biases in the genes represented on such platforms. At the time the project described in this thesis was initiated, whole genome expression microarrays were widely used to analyse gene expression outputs obtained from gene perturbation studies (i.e., the work pre-dates the development of massively parallel sequencing-based methods).

In the work described in this Chapter, GeneChip expression arrays produced by Affymetrix were used. The GeneChip probe arrays generated by Affymetrix use a combination of photolithography and combinatorial chemistry in a series of cycles to construct arrays of oligonucleotides (Singh-Gasson et al., 1999). A glass substrate is coated with linkers containing photolabile protecting groups. This glass substrate is then covered with a mask which exposes selected portions of the probe array to ultraviolet light. Upon illumination, the photolabile protecting groups are removed at the exposed regions enabling selective nucleotide addition to the surface. The nucleotides added at each step also contains light-sensitive protecting group. Different masks are applied and the cycle of illumination and chemical coupling is repeated until the probes reach their full length (25 nucleotides). In the end, a specific set of oligonucleotide probes synthesised at particular known locations on the array are generated.

The GeneChip arrays contain a large number of highly specific probe sets representing each gene (Figure 4.2). Such specificity is very important when measuring the expression of two very similar genes. Within each probe set, a gene is represented by millions of copies of eleven probe pairs (oligos) of 25 bp which are found throughout the mRNA sequence of the gene. The use of multiple probes generates high sensitivity and reproducibility while reducing background noise. A probe pair contains two probes. Probes that are perfectly complementary to the target sequence, called Perfect Matches (PM), are intended to measure mainly specific hybridisation. A second set of probes identical to PM except for a single nucleotide in the centre of the probe sequence (the 13th nucleotide), called Mismatches (MM), are intended to quantify non-specific hybridisation. A PM and its corresponding MM constitutes a probe pair. Such PM and MM probes are essential elements for eliminating effect of non-specific binding.

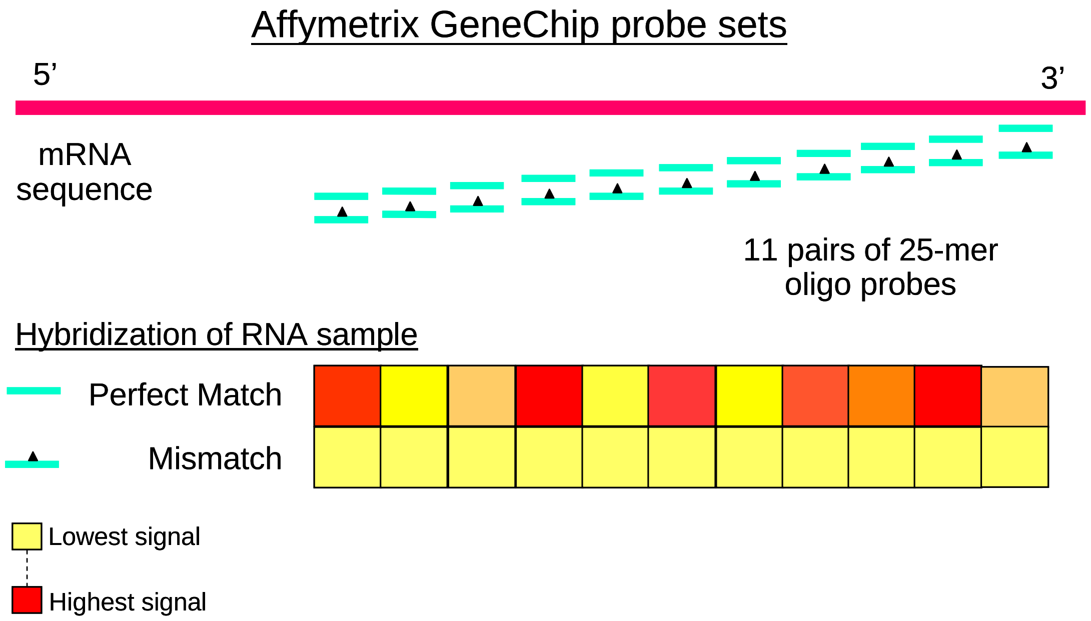

Affymetrix GeneChip probe sets
5'
3'
mRNA sequence
11 pairs of 25-mer oligo probes
Hybridization of RNA sample
Perfect Match
Mismatch
Lowest signal
Highest signal

Figure 4.2. The Affymetrix expression GeneChip probe sets. 11 pairs of 25-mer oligo probes were designed for an mRNA sequence. Each probe pair includes the perfect match probe and the mismatch probe where the middle nucleotide is replaced by a different one. During hybridisation, if the RNA samples contain fragments matching the probe sets, they will generate a signal with the perfect match probes, while no or very low signals will be detected for the mismatch probes.

The Affymetrix GeneChip expression array system is a one-colour microarray system. In a one-colour array, control and experimental samples are hybridised onto different arrays, detected with the same fluorescent dye, and comparisons are made across different hybridisations. The Affymetrix GeneChip has standard and optimised protocols for sample manipulation and hybridisation (Figure 4.3). To perform hybridisation, total RNA or mRNA extracted from the cell or tissues of interest is first reverse-transcribed using a T7-oligo(dT) promoter primer to generate double-stranded cDNA. The cDNA then undergoes an *in vitro* transcription (IVT) reaction in the presence of T7 RNA polymerase and biotinylated ribonucleotides to generate biotin-labelled complementary RNAs (cRNAs). The biotinylated cRNAs are fragmented (to optimise target-probe hybridisation kinetics) and hybridised onto the probe array. The hybridised probe array is stained with a streptavidin phycoerythrin (PE) conjugate and scanned. The PE conjugate is excited by laser and emits fluorescence for detection.

The GeneChip Human Genome U133 Plus 2.0 array provides a comprehensive coverage of protein coding genes the human genome. This chip includes 54 000 probe sets (11 in each set) representing over 47 000 human transcripts and variants, all of which are analysed in a single hybridisation. The sequences from which the probe sets were derived were selected from the GeneBank, dbEST and RefSeq databases and the probe sets themselves have been annotated onto the human genome sequence.

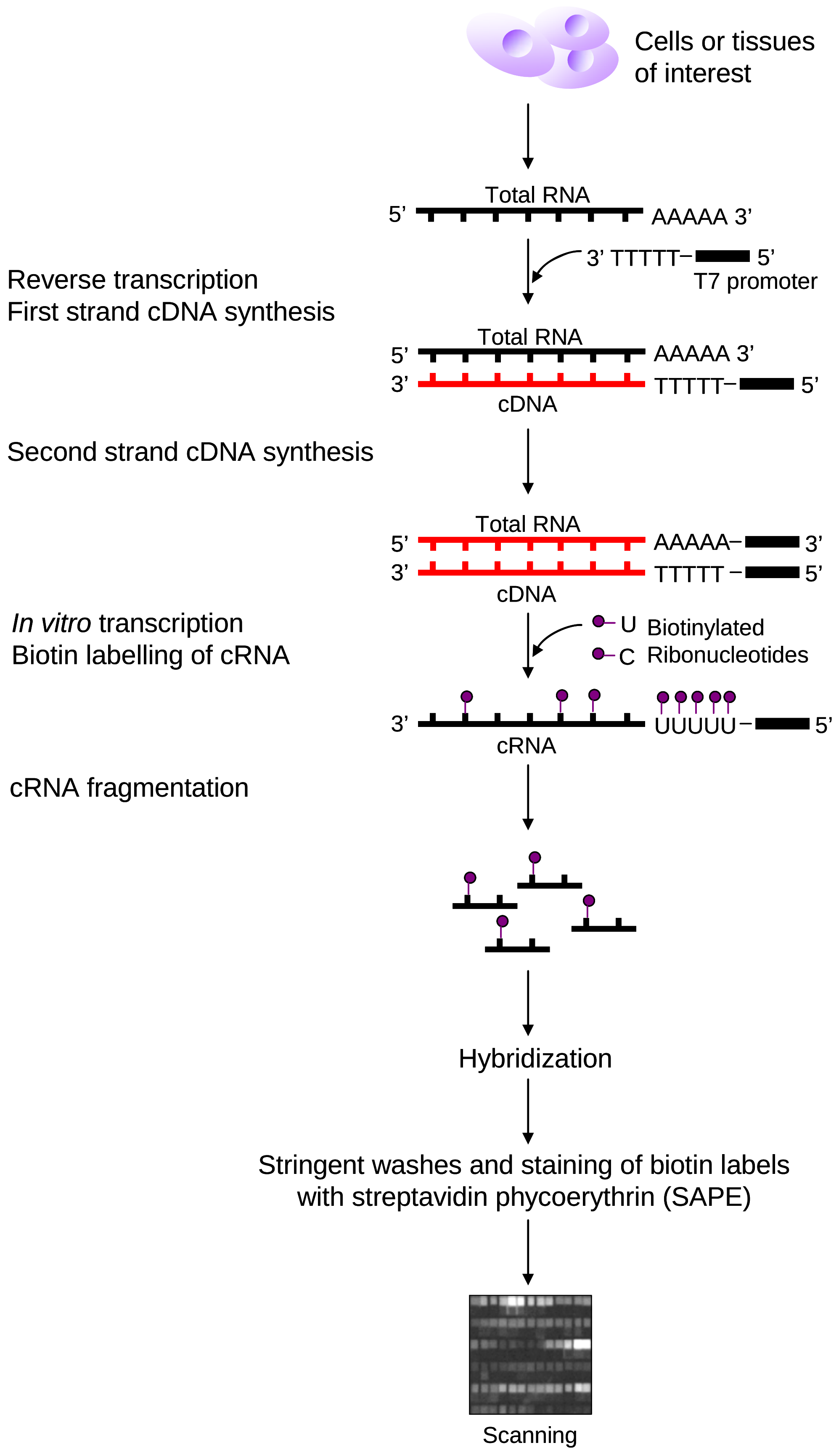


Figure 4.3. Target labelling and hybridisation of Affymetrix GeneChip arrays. Total cellular mRNA samples from the cells or tissues of interest are first reverse transcribed to generate double stranded cDNA with a T7 promoter. Complementary RNAs (cRNA) are generated by *in vitro* transcription with biotin-labelled ribonucleotides. The cRNAs are fragmented and hybridised on the array (please see text for details).

The use of commercial microarrays has many advantages over in-house custom-made microarrays. Firstly, for large-scale genome-wide analysis, generating in-house arrays is very time-consuming and requires a well developed informatics and array manufacture pipeline (which is not always available in academic laboratories). Thus, commercial arrays provide a widely available "off-the-shelf'' alternative. Secondly, commercial arrays are usually tested, validated and quality-controlled by both academic and commercial sources. Thirdly, target preparation and hybridisation protocols are well-established and usually require no further optimisation.

4.1.4 Microarray data analyses

Microarray experiments, regardless of whether they are one-colour or two-colour experiments, involve the measurement of the expression levels of a large number of genes in only a few replicate samples, given that microarrays are expensive and sometimes the biological samples are limiting. Developing appropriate statistical techniques to determine which changes are relevant is thereby very important. Typically, microarray analyses involve five main parts which are discussed below: quantitation, normalisation, inferential statistics, descriptive statistics and data mining.

A. Data processing methods (Quantitation)

Quantitation is the process of measuring the fluorescence intensity of spots or probes on the array while correcting it against the background intensity - which is another source of measurable fluorescence on the image.

Three different ways of processing and measuring probe set intensities on Affymetrix arrays have been developed, namely Affymetrix Microarray Suite v.5 (MAS5) (Affymetrix), robust multichip average (RMA) (Irizarry et al., 2003b) and GC-RMA (Wu and Irizarry, 2005). MAS5 was developed by Affymetrix where the weighted average of the 2% of probes having the lowest intensities was selected as background. It utilises the mismatch probe signals to adjust the perfect match intensity. For RMA analysis, each array is assumed to have a common mean background and the mismatch probes are ignored. GC-RMA is a modified version of RMA which models probe intensity as a function of GC-content. Comparison between the MAS5 and RMA softwares indicated that RMA has better precision to detect low expressing genes and has higher specificity and sensitivity for detecting differential expression (Irizarry et al., 2003a). In addition, GC-RMA was shown to over-correct the G+C content within probe sets whereas RMA introduce less bias than both MAS5 and GC-RMA (Siddiqui et al., 2006).

B. Normalisation

Normalisation is the process of removing systematic bias in the data across different samples while preserving the variation in gene expression that occurs because of biologically relevant changes in transcription. Normalisation is also essential to allow the comparison of gene expression across multiple microarray experiments.

A basic assumption of the normalisation process is that the average gene does not change in an experiment. In the global normalisation procedure, two main steps are involved: scaling and centering. In scaling, the intensity for all the gene expression measurements in one channel for two-colour arrays or one array for single-colour array are multiplied by a constant factor so that the mean measurement equals to one. In centering, the intensity of the measurements is centered to ensure that the mean and the standard deviations of all the distributions are equal. Other normalisation procedures include normalising the measurements to some house-keeping genes e.g. GAPDH and β-actin but this is based on the assumption that the expressions of these genes do not change across samples.

C. Determining Relevant Expression Differences (Inferential statistics)

Determination of the genes which are differentially expressed between two RNA samples is one of the most important yet difficult issues associated with high-throughput microarray analyses. A variety of procedures can be applied to extract the most biologically relevant and significant expression differences. A few examples of ways of determining these significant differences are described below:

- Fold change

The ratios of signal intensity of a gene between the experimental condition and the control conditions are calculated. A ratio is chosen as the threshold or cut-off (usually two fold) to determine genes having a significant change in expression. In otherwords, all genes having a ratio which exceeds the threshold are considered to be *bona fide* gene expression differences between the two samples. However, this method has low specificity and low sensitivity since the fold change chosen is entirely arbitrary and is prone to generate both false positives and false negatives in the analyses.

- Standard deviation

This method assumes the ratios between control and experimental values form a continuous normal distribution. Genes are selected according to their distance from the mean values of the control-to-experimental ratios. Usually the distances are taken to be ±2 or ±3 standard deviations. Two standard deviations from the mean represent a 95.45% confidence level whereas three standard deviations from the mean represent a 99.73% confidence level. In other words, for genes lying more than two standard deviations away from the mean, the probability that the genes selected are differentially expressed is 95.45%. For those genes lying more than three standard deviations away from the mean, the probability that the genes are differentially expressed is 99.73%.

- Univariate statistics

Univariate statistical test such as a *t*-test can be used to assign a probability (P value) to a gene which is being differentially regulated above a given threshold, when the log ratios of the control-to-experiment values follow a normal distribution. A *t*-test is used to determine the difference between the means of two populations. The *t*-test compares the size of the difference between means with the standard error of that difference. From a *t*-test, a *t* statistic is converted to a probability value P. But suppose you are measuring the expression levels of 5,000 genes, instead of applying the standard cut-off for statistical significance of $p<0.05$, it is appropriate to correct the P value estimate by dividing the number of gene expression measurements you are making, i.e. set P to the far more stringent value of $p<(0.05/5,000)$ or $p< 1x10^{-5}$. This is called a Bonferroni correction. However, such correction is sometimes too stringent and no differentially expressed genes may be reported.

This method is particular useful when replicates are present for the microarray analysis. This is a better method than the methods listed above as the variations across replicates can be assessed so that statistically-significant genes across replicates can be chosen. However, this method assumes that the changes in expression level of genes are highly correlated across replicates. This may not be true depending on the manipulation of the samples for hybridisation. Sometimes, large variations in gene expression levels of real differentially-expressed genes may be observed across replicates and they will be missed out when this method is used for analysis.

However, regardless of which analysis being used, false positives may still be identified. The percentage of false positives identified by chance is described as the false discovery rate (FDR). The false discovery rate can range from 10 to 80% depending on the statistical analyses (Tusher et al., 2001). One way to minimise the FDR is to increase the sample size.

D. Descriptive statistics

The patterns or signature of gene expression should be identified in all the gene expression values obtained in an experiment. This type of question is addressed using descriptive statistics or exploratory analysis. Clustering trees can show the relationships between samples (such as normal versus diseased cells), between genes, or both. Hierarchical clustering such as that used in the program Cluster/TreeView (http://rana.lbl.gov/EisenSoftware.htm) (Eisen et al., 1998), is probably the most popular way for making trees with microarray data. This method groups genes and/or samples with similar expression patterns into family trees. Gene expression values are colour coded from bright red (most up-regulated) to bright green (most down-regulated). This allows one to visualize large amounts of data.

Principal components analysis (PCA) is a different exploratory technique used to find patterns in gene expression data from microarray experiments. The central idea behind PCA is to transform a number of variables into a smaller number of uncorrelated variables called principal components. In a typical microarray experiment, the point of PCA is to detect and remove redundancies in the data (such as genes whose expression values do not change) in order to reduce the noise in the data set and to identify outliers (or clusters of outliers) that might be of interest to study.

E. Data mining

Once differentially-expressed genes are identified in the microarray analysis, these data must be interpreted in terms of gene functions and functional relationship between genes using existing biological knowledge. The Gene Ontology Consortium (http://www.geneontology.org/) addressed the need for consistent description of gene products with different databases. The GO project describes functions of gene products in three different categories: cellular components, biological processes and molecular functions. Such effort has made interpretation of differentially-expressed genes a manageable task.

4.1.5 Confirmation and validation of data

Technical and biological variability generated experimentally, and due to data processing and statistical methods affect the results obtained for microarray experiments. Increasing the number of replicates will decrease the false discovery rate (FDR) and thus the chance of getting false positive genes. However, sometimes it is difficult to increase the number of replicates considering that microarray experiments are expensive to perform and samples may often be limiting (e.g. patient samples). Therefore, the results obtain from microarray studies should be verified by other approaches.

- Comparison with existing literature.

The microarray data can be compared with information available in literature and databases. If there is agreement between the microarray analysis and data from other sources, this provides a general confidence level that the data accurately reflects the biological processes involved. Taking the analysis in this Chapter as an example, published target genes have been identified for SCL and GATA1 (Chapter 1, section 1.4.2.1 and 1.4.2.2). If these published target genes are also identified in the expression profiling analysis, this provides evidence that the data is likely to be meaningful.

- Other gene expression assays

Experimental approaches should be used to further confirm the results obtained from microarray analyses. qPCR and other assays mentioned in Chapter 1 (Chapter 1, section 1.3.2) are useful and sensitive assays to confirm the changes in gene expression obtained from microarray analyses.

4.2 Aims of this chapter

The overall aim of work presented in this Chapter was to identify putative target genes of the SCL erythroid as follows:

1. To study the changes in global gene expression patterns identified by siRNA knockdown of each of five members of the SCL erythroid complex. This would be accomplished by using the Affymetrix expression GeneChips.
2. To validate the gene expression differences obtained in the microarray analyses by q-PCR.
3. To identify differentially-expressed genes which are common to the 5 knockdown states. Such co-regulated genes would be considered to be putative targets of the SCL erythroid complex.
4. To search for common motifs in the regulatory regions of these co-regulated genes as a means of identifying the locations where these transcription factors bind in order to regulate them.

4.3 Overall strategy

In Chapter 3, siRNAs with knockdown efficiencies that satisfied specific criteria were selected for each transcription factor in the SCL erythroid complex. In this Chapter, the changes in expression of other genes in the genome which were a consequence of these knockdowns were studied by Affymetrix expression arrays. Three biological replicates for each transcription factor knockdown and the luciferase negative controls were performed. The qualities of the hybridised arrays were monitored and only the arrays passed the quality control were used in statistical analyses to identify differentially expressed genes by comparing the data from luciferase siRNA transfected cells and the cells transfected with specific siRNAs against transcription factors in the SCL erythroid complex. Confirmation of differentially expressed genes was addressed by performing quantitative PCR from the knockdown condition mRNA samples. Comparisons of the differentially expressed gene sets for different transcription factor knockdowns were used to identify co-regulated genes which were considered to be putative targets of the SCL erythroid complex. Computational analyses of common DNA binding motifs were also performed using the NestedMICA programme (Down and Hubbard, 2005) as a means of determining the binding location of transcription factors in the complex. The overall strategy of this expression study was summarised in Figure 4.4.

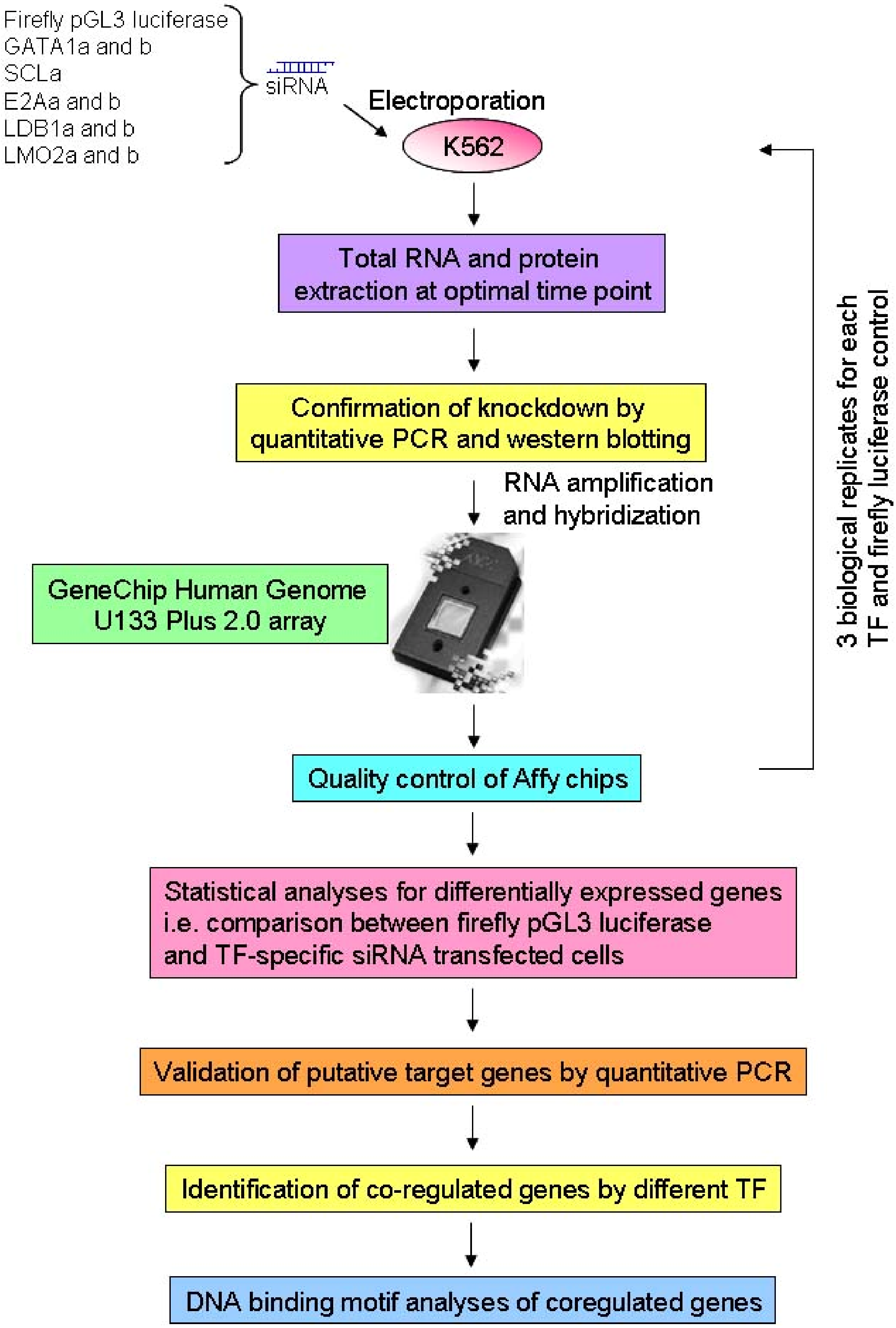

Firefly pGL3 luciferase
GATA1a and b
SCLa
E2Aa and b
LDB1a and b
LMO2a and b
siRNA
Electroporation
K562
Total RNA and protein extraction at optimal time point
Confirmation of knockdown by quantitative PCR and western blotting
RNA amplification and hybridization
GeneChip Human Genome U133 Plus 2.0 array
Quality control of Affy chips
Statistical analyses for differentially expressed genes i.e. comparison between firefly pGL3 luciferase and TF-specific siRNA transfected cells
Validation of putative target genes by quantitative PCR
Identification of co-regulated genes by different TF
DNA binding motif analyses of coregulated genes
3 biological replicates for each TF and firefly luciferase control

Figure 4.4. Overall strategy of expression profiling study of the silencing of the SCL erythroid complex. The effect of silencing of 5 members of the SCL erythroid complex on the expression of genes in the human genome was studied using Affymetrix expression GeneChip arrays. High quality hybridised GeneChips for 3 biological replicates for each transcription factor were used for statistical analyses to identify differentially expressed genes. Validation of differentially expressed genes was performed using SYBR green qPCR. Identification of co-regulated genes for each transcription factor was done by comparing the differentially-expressed genes of the 5 transcription factors. DNA motif analysis was also performed using NestedMICA.

4.4 Results

4.4.1 Preparation and quality control of samples

In order to minimise variation in the data obtained from three independent bioreplicates of each transcription factor knockdown, a number of parameter were controlled for the preparation of samples to be used on the Affymetrix GeneChips as follows:

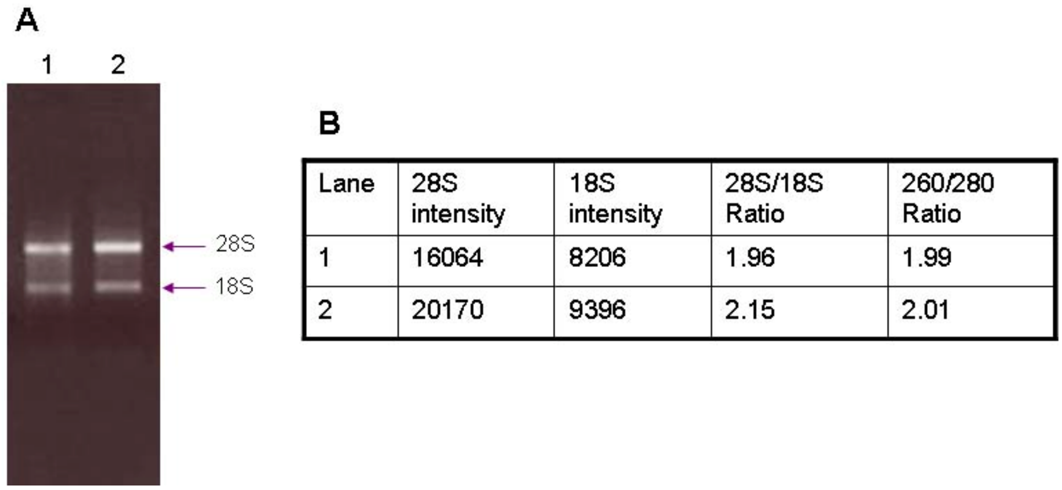
A. Culturing of cells

K562 cells were cultured and maintained at a concentration of 0.5 to 1 million/ml according to the ATCC specification. To ensure that transfections performed for individual bioreplicates behaved in a consistent manner, K562 cells were cultured for no more than a week before siRNA transfections were performed. K562 cells were split and fresh media were added one day before transfections.

B. RNA quality

RNA can be easily degraded by RNases and this can affect the quality of the RNA samples in subsequent manipulations and analyses. To control the quality of total cellular RNA samples used in the Affymetrix experiments, electrophoresis of the total RNA samples was performed to check if there were any signs of degradations. In the total cellular RNA, mRNA only comprises 1-3% of the total amount whereas ribosomal RNA (rRNA) makes up over 80% of the sample. After electrophoresis, only the rRNAs (28S, 18S, 5.8S and 5S) are visualised on the gel and can be used as a reference to monitor the overall RNA quality. For intact RNA samples, the ratio of 28S and 18S should be approximately 2:1 - this is traditionally used as the benchmark to monitor RNA degradation. In the Affymetrix experiments, all the RNA samples were checked for degradation. Figure 4.5 shows examples of the RNA samples extracted from E2Aa siRNA transfected and untransfected K562 cells, and the quantification of the 28S and 18S rRNA subunits. The bands for 28S and 18S rRNA subunits were quantified and yielded a 28S/18S ratio of approximately 2.

In addition to the 28S/18S ratio, the contamination of RNA samples by organic solvents and protein was also monitored. The ratio of absorbance at 260 nm and 280 nm indicates the purity of the RNA sample and should fall into the range between 1.8 and 2.1. Ratios of more than 2.1 indicate RNA degradation while ratios below 1.8 indicate protein contamination. The 260/280 absorbance readings of all the RNA samples were measured and fell between the range of 1.8 to 2.1. The readings in Figure 4.5 showed the 260/280 ratios of the two RNA samples described above - both samples showed ratios between 1.8 and 2.1.



Lane	28S intensity	18S intensity	28S/18S Ratio	260/280 Ratio
1	16064	8206	1.96	1.99
2	20170	9396	2.15	2.01

Figure 4.5. Agarose gel electrophoresis and 260/280 nm absorbance check for total RNA samples. Total cellular RNAs were visualised by ethidium bromide staining of a 1% denaturing TBE agarose gel. Panel A: Gel picture shows subunits of rRNA. Lane 1, total RNA sample extracted from E2A siRNA transfection; lane 2, total RNA sample extracted from untransfected K562 cells. Purple arrows on the right indicate the positions of the 28S and 18S rRNA subunits. The 5S rRNA subunit could not be detected on the gel. Panel B: Table shows the quantification of 28S and 18S bands on the gel by Labworks and the 260/280 absorbance ratios. The ratios of 28S to 18S were close to 2 while the 260/280 ratios were between 1.8 and 2.1 in both instances.

C. Amplification rate

Sample preparation for hybridisation to the Affymetrix GeneChip expression array required only a small amount of starting total RNA. This was because the RNA was reverse-transcribed to generate double-stranded cDNA containing a T7 promoter. *In vitro* transcription (IVT) of the cDNA was carried out under the control of the T7 promoter and large amounts of complementary RNA (cRNA) were generated. This amplification process allowed the synthesis of sufficient amounts of cRNA for hybridisation onto the array when the initial RNA sample was limiting (Figure 4.3). During the amplification process, the RNA could have become degraded which may have resulted in low amplification rates. Therefore, it was crucial to assess the amplification process before hybridisation. With a starting total RNA quantity of 5µg, an adjusted complimentary RNA (cRNA) amount of over 60 µg was expected according to the manufacturer’s protocol (adjusted cRNA amount was the amount of cRNA measured after IVT minus the starting amount of total RNA). If the yield of cRNA was substantially lower, there may have been RNA degradation during the amplification or the amplification may have been inefficient due to RNA purity or degradation of the starting material. In all the biological replicates of Affymetrix array hybridisation performed for this study, the amount of amplified cRNA was over 60 µg (Table 4.1).

siRNA	Selected optimal time point	Amount of adjusted amplified cRNA (µg)		
		Biorep 1	Biorep 2	Biorep 3
Luciferase	24 hr	---	---	121
Luciferase	36 hr	191	165	138
GATA1a	24 hr	---	140	127
GATA1b	24 hr	60	98	87
SCLa	24 hr	127	101	194
E2Aa	24 hr	62	68	130
E2Ab	24 hr	102	110	74
E12	24 hr	---	104	83
E47	24 hr	90	111	115
LDB1a	36 hr	128	119	109
LDB1b	36 hr	159	122	159
LMO2a	24 hr	---	80	97
LMO2b	24 hr	---	66	103

Table 4.1. Amount of adjusted amplified cRNA in all biological replicates. Adjusted cRNA amount is the amount of cRNA measured after IVT minus the starting amount of total RNA. Note: the quantities of cRNAs were not available for samples indicated by a ---. cRNAs for these samples were generated by others in the lab (Amanda Hall, Sanger Institute). However, these samples gave a cRNA yield greater than the 60 µg threshold.

D. Amplified RNA quality

In addition to checking the amplification rate, the quality of the amplified RNA was also monitored. This was because the resultant cRNAs may have been degraded during the amplification procedure. cRNA quality was assessed by agarose gel electrophoresis and by using a Bioanalyzer. The Bioanalyzer was not used in this study as it was not available in the lab when this project was carried out. Using the former method, the typical size distribution of the unfragmented cRNA below 1 kb was observed as shown in Figure 4.6. All the amplified cRNA samples used in this study were visualised by agarose gel electrophoresis before hybridisation to the Affymetrix GeneChips. Any samples which did not show the expected size distribution were discarded and the process was repeated until high quality cRNA was obtained.

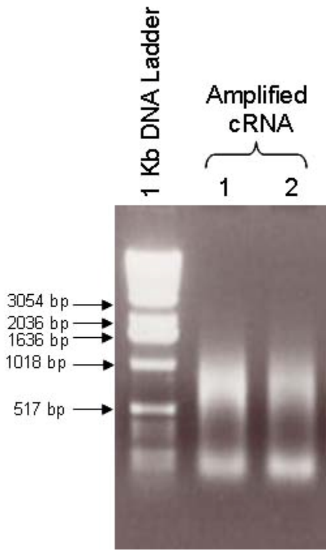


Figure 4.6. Agarose gel electrophoresis of amplified cRNAs. The purified and unfragmented cRNAs obtained from IVT were visualised by ethidium bromide staining of a 1% denaturing TBE agarose gel. The left lane shows the 1 kb DNA ladder and the corresponding sizes of the bands are labelled. Lane 1, luciferase control siRNA cRNA sample; lane 2, GATA1a siRNA knockdown cRNA sample.

4.4.2 Quality control of Affymetrix GeneChips

Once the cRNA samples were checked for quality control, they were hybridised onto the Affymetrix GeneChip expression arrays. However, the hybridisations themselves were subjected to strict quality control criteria even before any of the arrays were analysed with respect to the biological study being performed. Any GeneChips which do not pass the criteria for quality controls were discarded. The following criteria were used to assess the quality of Affymetrix Gene Chip hybridisations.

A. Probe array image inspection

One of the first criteria to be checked was the scanned GeneChip image. This was done to determine the overall quality of the hybridisation. The presence of observable image artifacts such as scratches, uneven signal intensity across array etc. was inspected by eye. Each probe cell was also visualised by zooming in. None of the GeneChips hybridised for this study showed obvious and visible artifacts. An example of a high quality GeneChip hybridisation from this study is shown in Figure 4.7.

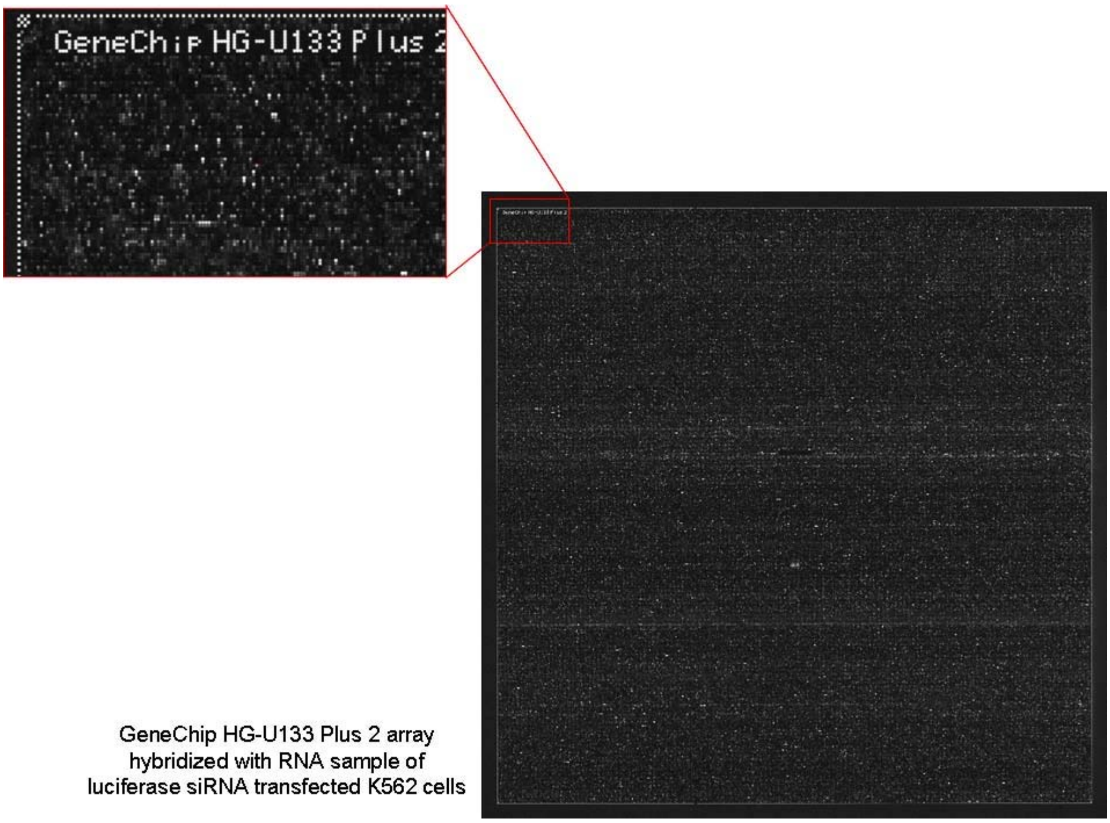

GeneChip HG-U133 Plus 2
GeneChip HG-U133 Plus 2 array
hybridized with RNA sample of
luciferase siRNA transfected K562 cells

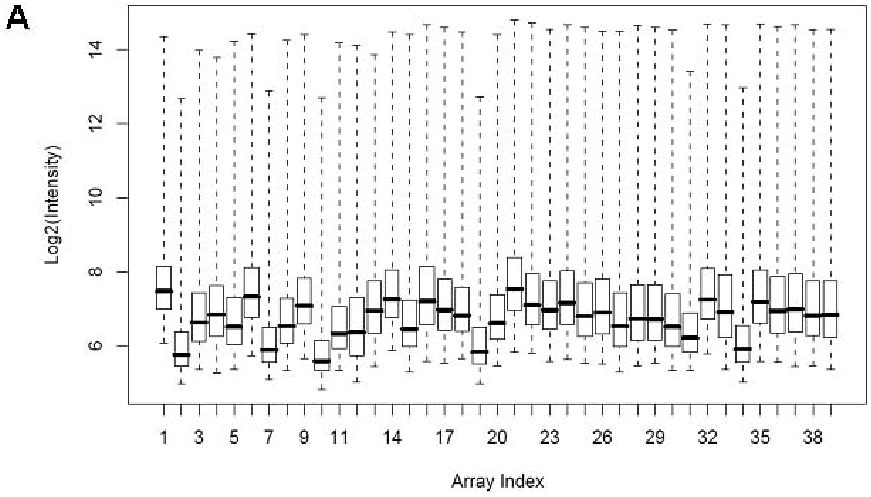
Figure 4.7. Image of hybridised Affymetrix GeneChip. GeneChip HG-U133 Plus 2 GeneChip arrays hybridised with cRNA derived from K562 cells transfected with luciferase siRNA at the 36 hour time point is shown. Image on the right shows the entire scanned GeneChip. Image on the left in the red box shows the zoomed image of the top left hand corner of the GeneChip. This is an example of a high quality GeneChip image with no visible artifacts.

B. Intensity correlation

Further quality control of the Affymetrix GeneChips was carried out by analysing the signal intensity and control gene profiles. This was done using the AffyQC Report package of Bioconductor. The signal intensity of all arrays included in the data analyses was assessed. The AffyQC Report package generated the log2 intensity of all the perfect match probes in various GeneChips and the density plots of these intensity values. Regardless of the samples being hybridised, the overall signal intensity of all the GeneChips should be similar since the majority of probe signals (i.e., gene expression levels) are not changing amongst the samples. All the GeneChips hybridised for this project showed similar patterns in the density of intensity values with the 50% of probes having a log2 intensity value between 5 to 8 (Figure 4.8). This indicated that the GeneChips all showed similar hybridisation characteristics and passed the signal intensity quality control criteria.

siRNA	GeneChip Index		
	Biorep 1	Biorep 2	Biorep 3
Luciferase (24 hr)	1	2	3
Luciferase (36 hr)	37	38	39
GATA1a	4	5	6
GATA1b	7	8	9
SCLa	10	11	12
E2Aa	19	20	21
E2Ab	22	23	24
E12	13	14	15
E47	16	17	18
LDB1a	25	26	27
LDB1b	28	29	30
LMO2a	31	32	33
LMO2b	34	35	36

Table 4.2. GeneChip (array) index used in the affyQCReport.



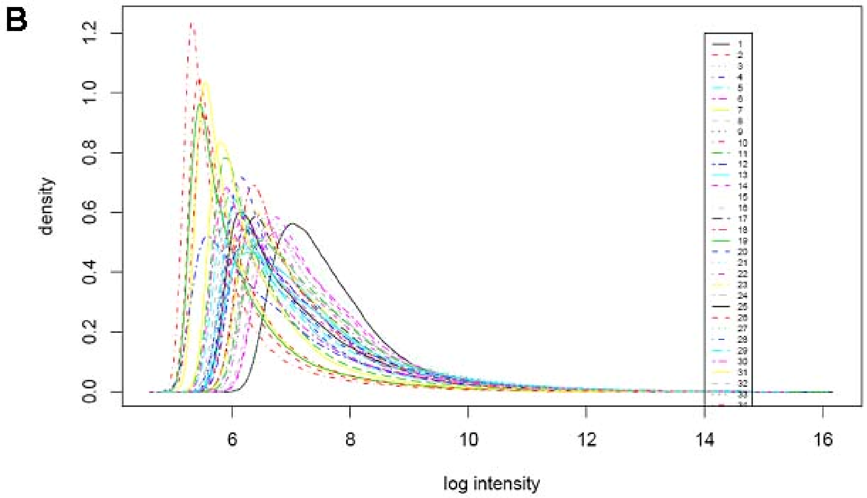


Figure 4.8. Perfect match intensity of all GeneChips studied in this project. Perfect match probe log2 intensities were calculated in the AffyQC Report package of Bioconductor. Panel A: box plot of log intensity of all perfect match probes across all GeneChips used in this project. The boxes contain the median centred 50% of the datapoints for each GeneChip hybridisation. The x-axis shows the GeneChip index (Table 4.2) while the y-axis shows the log2 intensity of probes. Panel B: density plot of log intensity of all perfect match probes across all GeneChips used in this project. The x-axis shows the log intensity of probes while the y-axis shows the kernel density of probes having a particular log intensity. The numbering of GeneChips and the corresponding samples are shown in Table 4.2.

C. Housekeeping gene profiles

The intensity signals of housekeeping genes on the GeneChips were also used as a means of assessing hybridisation quality. The signal intensity of the 3’ probe sets for the house-keeping genes β-actin and GAPDH were compared to the signal intensity of the corresponding 5’ probe sets. For good quality hybridisation samples, the 3’ to 5’ ratio should be less than 3 as degradation usually occurs from the 5’ end of mRNA, resulting in an accumulation of 3’ fragments. A high 3’ to 5’ ratio may also indicate inefficient transcription of double-stranded cDNA (ds cDNA) or biotinylated cRNA as the antisense cRNA was transcribed from the sense strand of the ds cDNA via the T7 promoter at the 3’ end of the sense strand. The 3’ to 5’ ratios for β-actin and GAPDH for all the GeneChip hybridisations for this project were below 3 as shown in Figure 4.9.

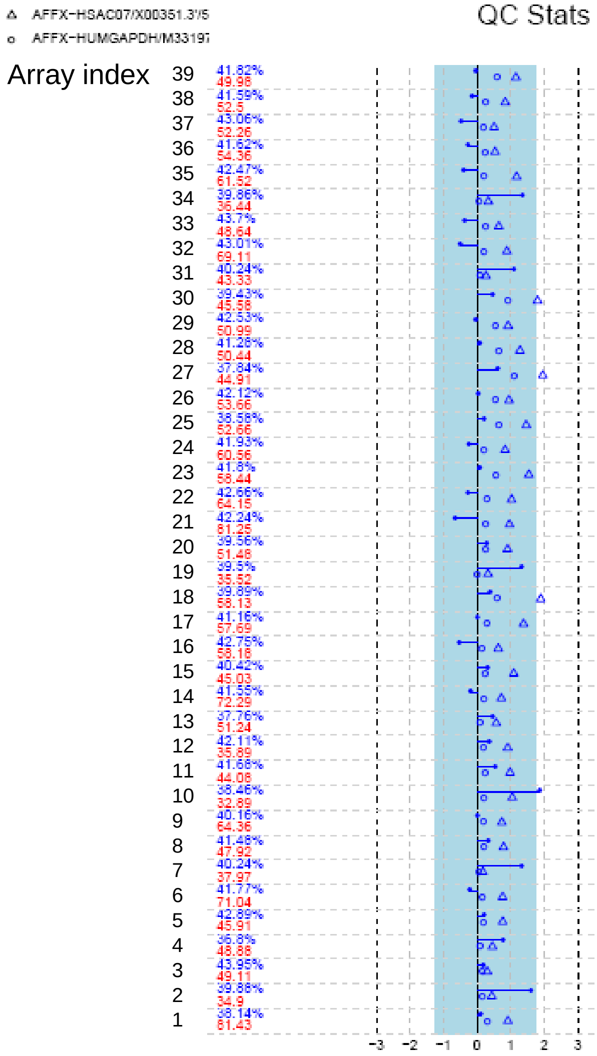


Figure 4.9. Internal house-keeping control gene profile and present call profile. The black numbers on the left are indicative of the GeneChip index (see Table 4.2). The blue numbers and red numbers next to the GeneChip index show the percentage of present call probes and the average background intensity respectively. The dotted vertical lines delineate the scale of -3 to 3 for the 3' to 5' ratios of the house-keeping genes. The triangles show the 3' to 5' ratios for β-actin while the circles showed the 3' to 5' ratios for GAPDH. When the circles and triangles are coloured in blue, they were within the acceptable quality control ratios, otherwise they are coloured in red.

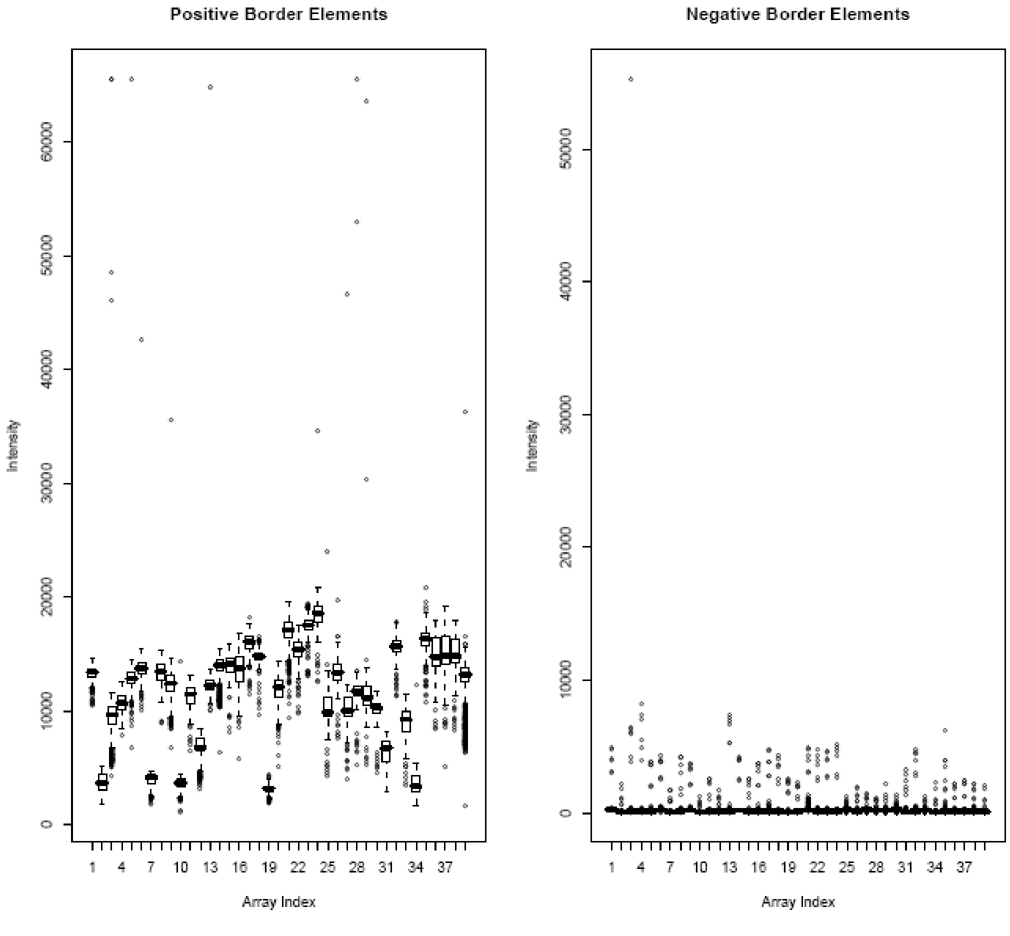
D. Average background intensity and percentage of present genes

The background intensity of the hybridisation signals on the GeneChips has a great impact of quantification of probe intensity and therefore it was also monitored as one of the quality control criteria. According to the documentation in the Affymetrix manual, the typical average background values should range from 20 to 100. The GeneChip hybridisations obtained for this project all had average background intensities falling within this range (see Figure 4.9; red numbers).

The number of probe sets called 'present' relative to the total number of probe sets on the GeneChip is described as the percentage of present genes. This percentage is an indication of sample quality and is dependent on the cell type and biological or environmental stimuli. Low percentage values imply poor sample quality whilst replicates are expected to have similar percentage values. All the arrays hybridised have percentage of present calls of approximately 40% indicating that the hybridisations and the sample qualities were similar for the GeneChips analysed for this project (see Figure 4.9; blue numbers).

E. Border elements intensity correlation

During the hybridisation of samples onto the Affymetrix array, control “spikes” were included. The control oligo B2 was spiked into the hybridisation mix and it hybridised to features along the outer edges and corners of the GeneChip (so-called “border elements”). These hybridisation controls were independent of RNA sample quality and amplification and were used as positive controls for even hybridisation characteristics across the GeneChip and were also used by the software for automatic grid alignment over the image during quantitation of signals. To assess for even hybridisation of the GeneChips, the intensities for all border elements were collected. Elements with an intensity of 1.2 times above the mean were regarded as “signal” controls (positive controls). Elements with a signal less that 0.8 of the mean were regarded as “background” controls (negative controls). The intensities of positive and negative border elements for each GeneChip should be similar. Large variations in the positive control elements indicate non-uniform hybridisation or gridding problems. Variations in the negative controls indicate background fluctuations. At least 50% of positive border elements of the GeneChip hybridisations had intensity values below 20000 with an average of 11774 of all the positive border elements in all arrays. The average intensity of all negative border elements in all of the GeneChips was 141 with 50% of them close to 0 (Figure 4.10).


Positive Border Elements

Negative Border Elements

Figure 4.10. Boxplot of intensity of positive and negative border elements. Left panel shows the intensity of the positive border elements and right panel shows the intensity of the negative border elements. The boxes contain the median centred 50% of the datapoints for each GeneChip hybridisation. In both panels, the y-axis is the signal intensity and the x-axis is the GeneChip (array) index.

F. Hybridisation and Poly-A controls

In addition to the housekeeping control genes discussed above, the quality of the entire amplification and labelling process, and the sensitivity of the GeneChips was assessed using exogenous positive control poly-A mRNA "spikes". These "spikes" were poly-A mRNAs derived from *in vitro* synthesised, polyadenylated transcripts for several *B. subtilis* genes (*lys*, *phe*, *thr* and *dap*). Probe sets for these genes were represented on the GeneChips. These control mRNAs were added to the starting RNAs at different concentrations and were amplified and labelled together with the RNA samples. Assessing the signal intensity generated for these controls helped monitor the amplification and labelling process independent of the RNA sample. The median of signal intensities of these spike controls in the 39 GeneChips in this experiment is shown in Figure 4.11. The "spike" controls showed increasing and linear signal intensities with increasing concentrations in the starting RNA samples. The lowest concentration of these controls allowed messages which were represented at 1 copy in 50 000 mRNAs to be detected on the GeneChip. All the GeneChips hybridised for this Chapter showed similar patterns for the poly-A spike controls.

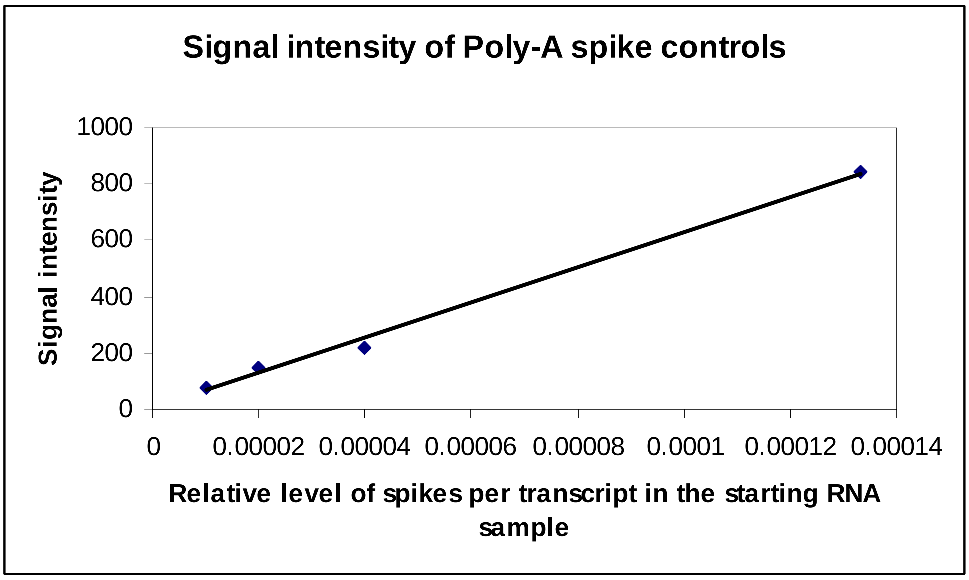


Figure 4.11. Signal intensity of poly-A RNA "spike" controls of the Affymetrix GeneChip hybridised with cRNA derived from K562 cells transfected with the various siRNAs. The x-axis is the relative level of each of the *B. subtilis* "spike" control transcripts per transcript in the starting RNA sample. Blue dots indicate each of the four *B. subtilis* spike control transcripts with increasing mRNA concentration from left to right: *lys*, *phe*, *thr* and *dap*. The y-axis is the median values of signal intensity of these transcripts on the 39 GeneChips hybridised in this experiment.

Additional controls were also included in the hybridisation to the GeneChips. These controls were used to evaluate the hybridisation efficiency independent of the RNA preparation and amplification procedure. These mRNA transcript controls were derived from genes in the biotin synthesis pathway of *E. coli* (*Cre*, *BioB*, *BioC* and *BioD*). Probe sets for these genes were represented on the GeneChips. Like the poly-A RNA controls, the hybridisation controls were added at different concentrations (1.5 pM, 5 pM, 25 pM and 100 pM for *BioB*, *BioC*, *BioD* and *Cre* respectively*)*. However, unlike the poly-A RNA controls, these mRNAs were labelled separately from the starting RNA samples and were added directly into the hybridisation mixture. The signal intensity of these genes should increase according to their relative concentrations if the hybridisation was performed according to manufacturer's standard. The median of signal intensities of these hybridisation controls in the 39 GeneChips in this experiment is shown in Figure 4.12. The hybridisation controls showed increasing and linear signal intensities with increasing concentrations in the hybridisation mixture. All the GeneChips hybridised for this Chapter showed similar patterns for the hybridisation controls.

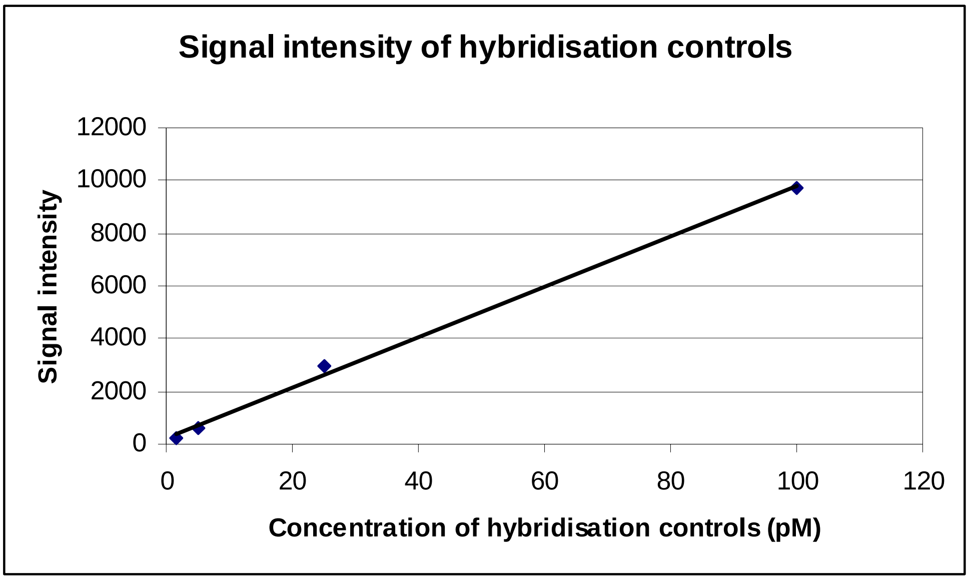


Figure 4.12. Signal intensity of the hybridisation controls of the Affymetrix GeneChip hybridised cRNA derived from K562 cells transfected with the various siRNAs. The x-axis is the concentration of the hybridisation controls of *E. coli* genes. Blue dots indicate the spike control genes with increasing concentration from left to right: *BioB*, *BioC*, *BioD* and *Cre*. The y-axis is the signal intensity of these genes in the 39 GeneChips hybridised in this experiment.

4.4.3 Data analysis of Affymetrix GeneChips

Once the hybridisations onto Affymetrix GeneChips passed the quality control criteria, the data derived from the biological study was analysed by statistical methods in order to determine differentially expressed genes between the relevant luciferase control (time points 24 or 36 hrs) and its corresponding transcription factor-specific siRNA knockdown conditions. Many methods of quantification of probe sets have been developed e.g. MAS5, RMA and GC-RMA (Section 4.1.4). In the analyses described in this Chapter, the RMA method was used as it was shown to be more sensitive and more specific while introducing less bias to G+C content of probes (Irizarry et al., 2003a; Siddiqui et al., 2006). To handle such large data sets involving large numbers of probe sets and transcript information from across entire human genome, the GeneSpring GX 7.3.1 data analysis software was used.

4.4.3.1 Normalisation and statistical analyses of Affymetrix GeneChip data

Figure 4.14 outlined the strategy used to determine the genes that were differentially expressed between the control and experimental conditions. Signal intensities of the probe sets in all the 39 scanned GeneChips were imported into the GeneSpring analysis suite and quantitated by RMA. Experiments were created in GeneSpring to include all three biological replicates of the luciferase controls, the three biological replicates of the siRNAa transfections and the three biological replicates of the siRNAb transfections (except for SCL, where only the siRNAa assay was used). The signal intensities of all probes/genes were normalised in the following ways:

1. values of lower than 0.01 were set at 0.01,
2. to the median probe intensity of all measurements per hybridisation,
3. to the median of all gene intensities in all the samples in the experiment.

These normalised intensity values of all genes were then exported to Microsoft Excel and statistical analyses of differentially expressed genes were performed. The analyses of the two siRNAs for each transcription factor were done separately. The mean signal intensities of each gene were derived for the 3 biological replicates for luciferase, siRNAa and siRNAb respectively. Comparisons were made between luciferase and siRNAa as well as between luciferase and siRNAb in order to derive ratios of differences in gene expression. For statistical purposes, it was assumed that the ratios of gene expression for any one experiment occur as a normal distribution centered around the mean (Figure 4.13). Genes which were differentially expressed between the luciferase and experimental siRNAs by more than 2 standard deviations away from the mean ratio of the entire dataset were chosen for further analyses. Two standard deviations were used as a cut-off as it represented a 95.45% confidence level – in other words, the genes identified were statistically significant in terms of differential expression.

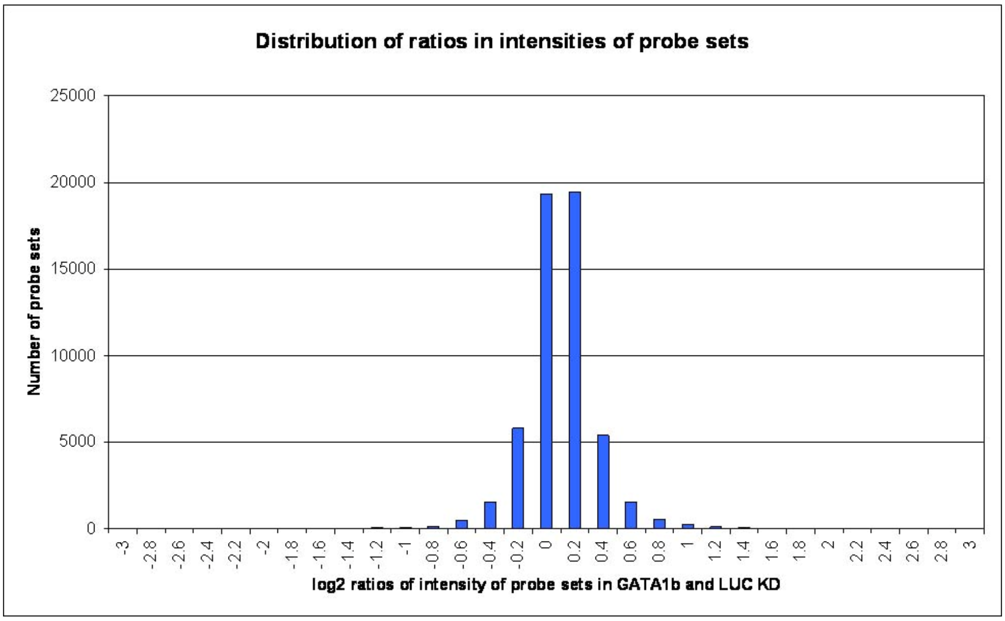


Figure 4.13. Distribution of log2 ratios of intensity of probe sets in GATA1b knockdown against luciferase knockdown. Y-axis: number of probe sets; x-axis: log2 ratios of the intensity of probe sets in the GATA1b knockdown against luciferase knockdown. The probe sets are centred around the mean in a normal distribution.

For each transcription factor knockdown experiment, four gene lists were obtained: genes down-regulated in siRNAa, genes down-regulated in siRNAb, genes up-regulated in siRNAa and genes up-regulated in siRNAb. Gene lists for the two siRNAs (a and b) for each transcription factor were treated independently up to this point because different siRNAs for the same gene can generate different off-targeting effects (Chapter 1, section 1.3.1.3 B). To filter away these off-target genes from further analyses, the down-regulated gene lists of the two siRNAs for each transcription factor were compared while the up-regulated gene lists of the two siRNAs for each transcription factor were also compared. These comparisons were performed using Venn diagrams and would allow for the identification of genes which were differentially expressed by both siRNAs. The gene lists identified by each siRNA are shown in the Venn diagrams of Figure 4.15. The genes found in the overlaps of the Venn circles (either up- or down-regulated) were considered as putative target genes of each transcription factor. Three points should be noted when interpreting the data from the Venn diagrams:

1. The numbers shown in the Venn diagrams are number of probe sets rather than numbers of genes. On the Affymetrix GeneChip, a gene can be represented by more than one probe set. Thus the actual numbers of genes found to be up- or down-regulated by each transcription factor are less than the numbers shown in Figure 4.15.

2. Genes/probe sets which were down-regulated by the knockdown of a transcription factor were considered to be putative target genes which were activated by the transcription factor.

3. Gene/probe sets which were up-regulated by the knockdown of a transcription factor were considered to be putative target genes which were repressed by the transcription factor.

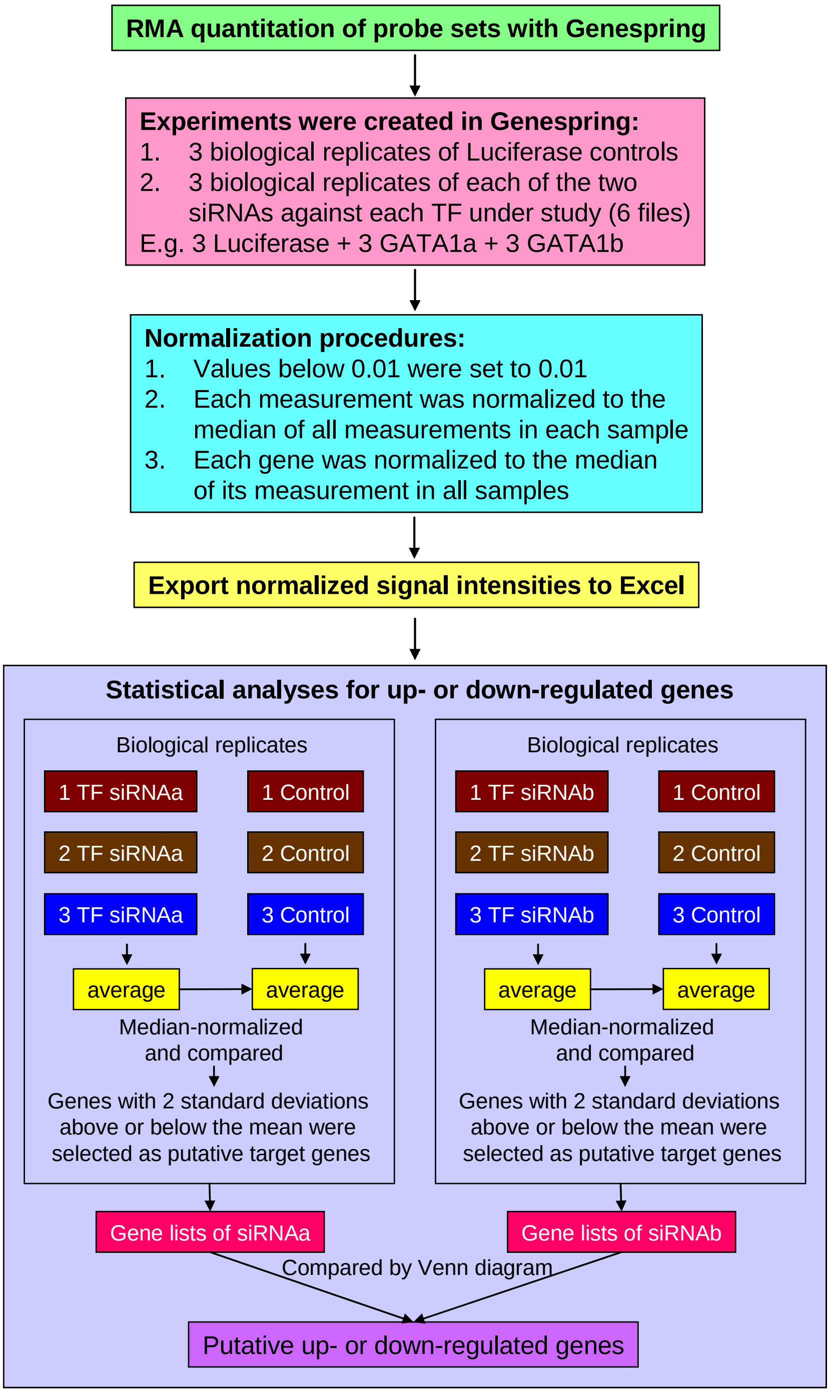
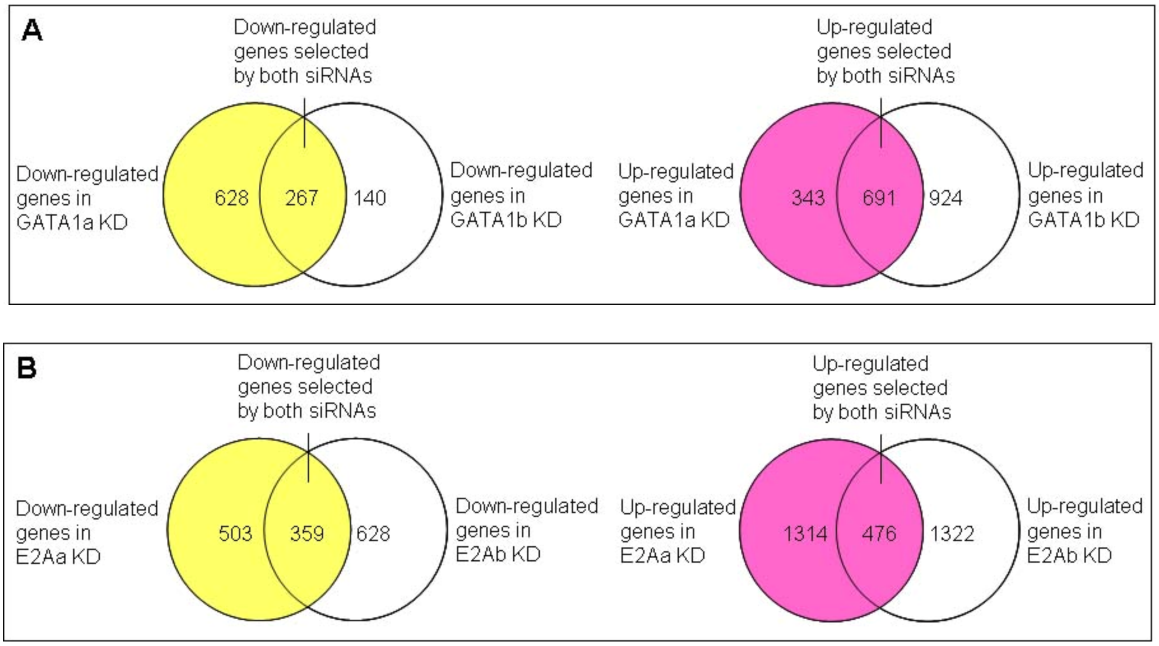

RMA quantitation of probe sets with Genespring
Experiments were created in Genespring:
1. 3 biological replicates of Luciferase controls
2. 3 biological replicates of each of the two siRNAs against each TF under study (6 files)
E.g. 3 Luciferase + 3 GATA1a + 3 GATA1b
Normalization procedures:
1. Values below 0.01 were set to 0.01
2. Each measurement was normalized to the median of all measurements in each sample
3. Each gene was normalized to the median of its measurement in all samples
Export normalized signal intensities to Excel
Statistical analyses for up- or down-regulated genes
Biological replicates
1 TF siRNAa
1 Control
2 TF siRNAa
2 Control
3 TF siRNAa
3 Control
average
average
Median-normalized and compared
Genes with 2 standard deviations above or below the mean were selected as putative target genes
Biological replicates
1 TF siRNAb
1 Control
2 TF siRNAb
2 Control
3 TF siRNAb
3 Control
average
average
Median-normalized and compared
Genes with 2 standard deviations above or below the mean were selected as putative target genes
Gene lists of siRNAa
Gene lists of siRNAb
Compared by Venn diagram
Putative up- or down-regulated genes

Figure 4.14. Flow diagram of statistical analyses of differentially-expressed genes in Affymetrix GeneChip. Signal intensities of the probe sets were quantitated by RMA. Experiments were created in GeneSpring to include all three biological replicates of the luciferase controls, the three biological replicates of the siRNAa transfections and the three biological replicates of the siRNAb transfections. The signal intensities of all probes/genes were normalised at three levels. The statistical analyses of the two siRNAs for each transcription factor were done separately. The mean signal intensities of each gene were derived for the 3 biological replicates for luciferase, siRNAa and siRNAb respectively. Comparisons were made between luciferase and siRNAa as well as between luciferase and siRNAb in order to derive ratios of differences in gene expression. Genes which were differentially expressed between the luciferase and

experimental siRNAs by more than 2 standard deviations away from the mean ratio of the entire dataset were chosen for further analyses.

4.4.3.2 Differentially-expressed genes and comparison of two siRNAs

Based on the Affymetrix studies described above, the general functional roles of GATA1, SCL, E2A, LDB1 and LMO2 were determined with respect to how they affected gene expression patterns across the entire human genome in K562 cells. These roles were based on the numbers of up-regulated and down-regulated genes (probe sets) which were identified to be common to both the siRNA a and b knockdowns for each transcription factor (Figure 4.15). For the GATA1 knockdowns, 267 probe sets were found to be down-regulated (activated by the transcription factor) while 691 probe sets were found to be up-regulated (repressed by the transcription factor) for both siRNAs. This suggests that GATA1 is primarily a repressor in K562 cells. 486 and 359 probe sets were shown to be up- and down-regulated by E2A respectively. This suggests that E2A acts as both a repressor and an activator in K562 cells. Similarly, for the LDB1 knockdowns 716 probe sets were up-regulated, while 822 probe sets were down-regulated, suggesting that LDB1 acts as both an activator and a repressor in K562 cells. LMO2 was seen to act mainly as a repressor - 1063 probe sets were up-regulated by LMO2 while 54 probe sets were down-regulated in the knockdown experiments. As only one siRNA was found to be effective in silencing the expression of SCL in K562 cells, it was difficult to determine its general role in regulating gene expression in K562 cells. 897 probe sets were found to be down-regulated by SCL while 1811 probe sets were found to be up-regulated. This suggests that SCL is more likely to be a repressor. However, as only one siRNA knockdown was used in the Affymetrix expression study, some of these genes might be off-target genes. Therefore, it was not possible to draw any firm conclusions of how SCL normally affects gene expression in this cell line.


A
Down-regulated genes selected by both siRNAs
Down-regulated genes in GATA1a KD
628
267
140
Down-regulated genes in GATA1b KD
Up-regulated genes selected by both siRNAs
Up-regulated genes in GATA1a KD
343
691
924
Up-regulated genes in GATA1b KD
B
Down-regulated genes selected by both siRNAs
Down-regulated genes in E2Aa KD
503
359
628
Down-regulated genes in E2Ab KD
Up-regulated genes selected by both siRNAs
Up-regulated genes in E2Aa KD
1314
476
1322
Up-regulated genes in E2Ab KD

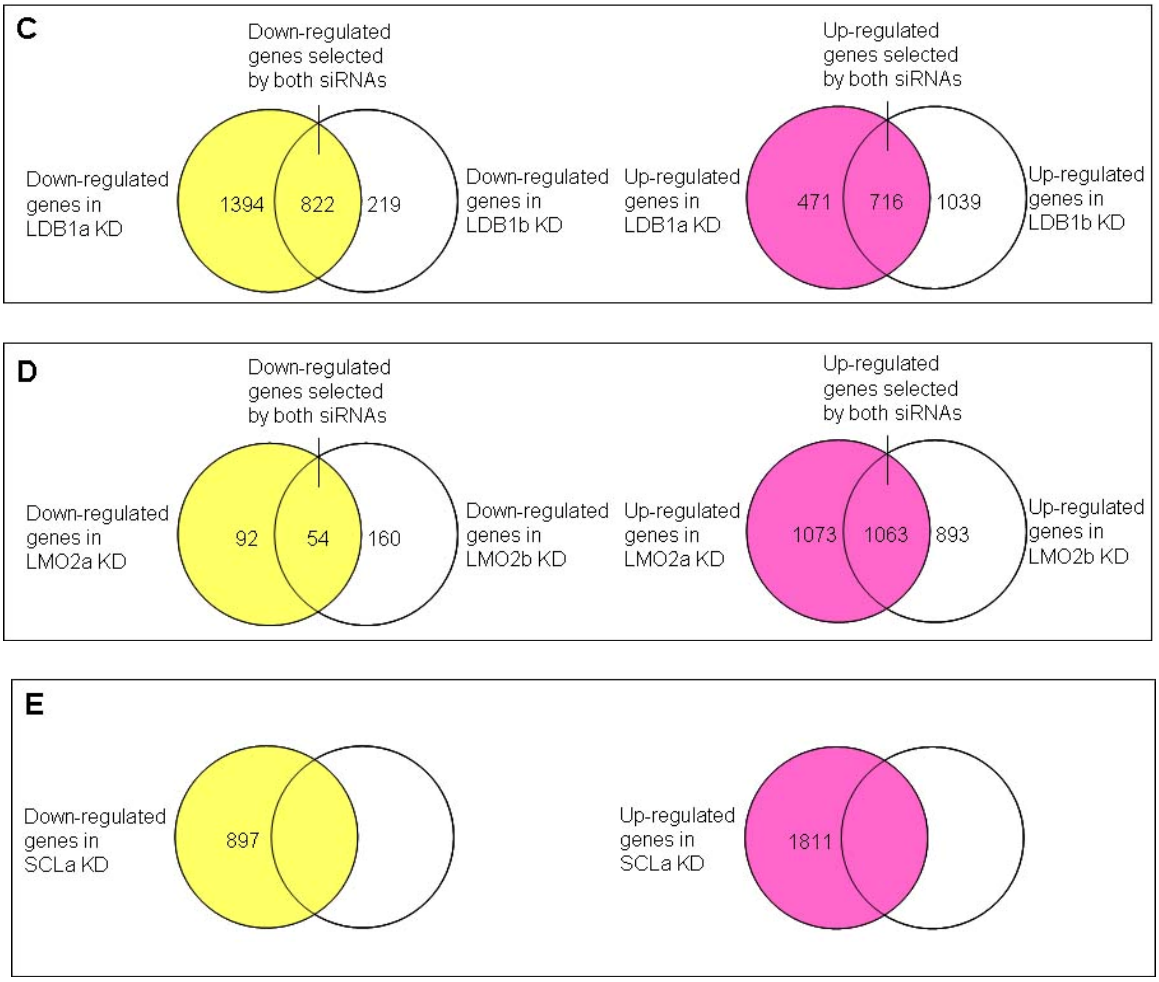

C
Down-regulated genes selected by both siRNAs
Down-regulated genes in LDB1a KD
1394
822
219
Down-regulated genes in LDB1b KD
Up-regulated genes selected by both siRNAs
Up-regulated genes in LDB1a KD
471
716
1039
Up-regulated genes in LDB1b KD
D
Down-regulated genes selected by both siRNAs
Down-regulated genes in LMO2a KD
92
54
160
Down-regulated genes in LMO2b KD
Up-regulated genes selected by both siRNAs
Up-regulated genes in LMO2a KD
1073
1063
893
Up-regulated genes in LMO2b KD
E
Down-regulated genes in SCLa KD
897
Up-regulated genes in SCLa KD
1811

Figure 4.15. Venn diagram comparison of genes (probe sets) identified by siRNA knockdown experiments in K562 cells for each of five transcription factors found in the SCL erythroid complex. Down-regulated genes are shown in the Venn diagrams on the left (yellow) while up-regulated genes are shown in the Venn diagrams on the right (pink). Numbers shown in the Venn circles are numbers of probe sets for genes in the human genome identified in the relevant siRNA knockdown studies. The numbers shown in the overlap of the Venn circles denote those probe sets found in both the siRNA a and b knockdown conditions. Panel A: Venn diagram of GATA1 knockdowns; panel B: Venn diagram of E2A knockdown; panel C: Venn diagram of LDB1 knockdowns; panel D: Venn diagram of LMO2 knockdowns; panel E: Venn diagram of SCL knockdown study.

4.4.3.3 Validation of selected differentially-expressed genes by quantitative PCR

False positive expression differences are common in microarray analyses (Tusher et al., 2001). To determine whether the data obtained from the Affymetrix GeneChip analyses represented *bona fide* expression differences between the control luciferase and transcription factor knockdown conditions, a subset of up-regulated and down-regulated genes were further studied by quantitative PCR. Such validation allows us to determine and evaluate the Affymetrix GeneChip technology as a means of studying differential expression. For this purpose, the differentially-expressed genes of GATA1, SCL and E2A were investigated. Up-regulated and down-regulated genes which were transcription factors were chosen in the validation as they are the key components of a transcription network and are thus excellent genes to study in the context of understanding transcriptional cascades downstream of the SCL erythroid complex in future work (see Chapter 7).

Table 4.3 lists the transcription factors that were up- or down-regulated in the knockdown studies that were chosen for validation. The majority of the genes selected were implicated in transcriptional regulation in various developmental processes including haematopoietic development. Within these genes, some of them are published targets of the corresponding transcription factors (MYC, EKLF, NFE2 and GFI1B). They were included in the validation to evaluate the sensitivity of the qPCR assays.

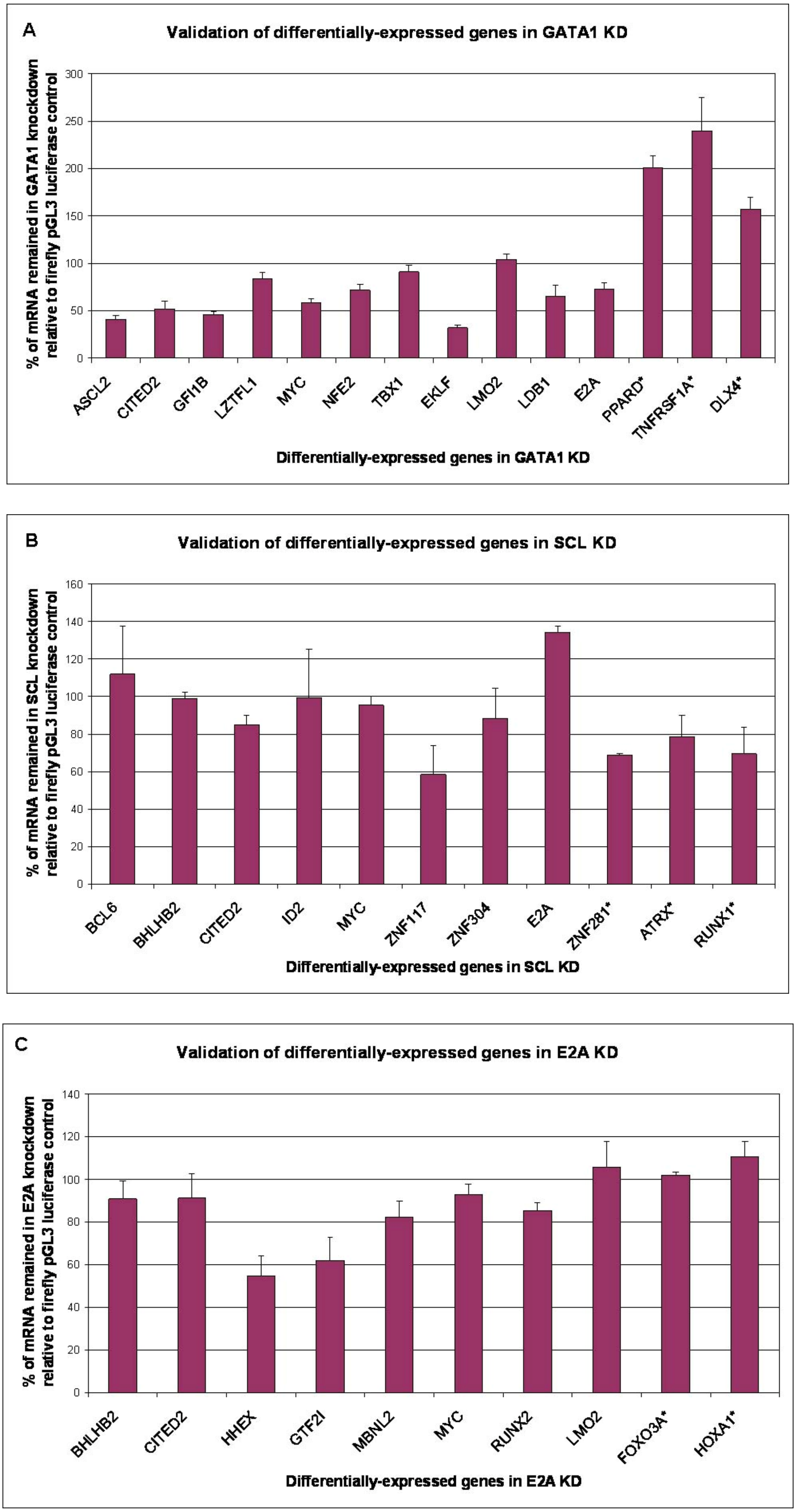
The RNAs used for the quantitative PCR validation were those used in the Affymetrix GeneChip analyses. They represented samples obtained from two independent biological replicates of the siRNA knockdown studies of GATA1, SCL and E2A. In each case, the mRNA levels of the putative target genes were compared between the knockdown sample and the control firefly luciferase sample. The mRNA levels remaining after knockdown relative to the luciferase control are shown in the bar charts in Figure 4.16. As described previously for the knockdown time-course experiments in Chapter 3 (Section 3.4.5), four house-keeping genes, β-actin, GAPDH, β-tubulin and RPL19, were used as controls for normalisation to minimise variations of RNA quality and concentrations.

The cut-off in fold change used to determine whether a differentially-expressed gene showed a *bona fide* expression difference between the transcription factor knockdown and the luciferase control was different for each transcription factor. This cut-off was determined following the fold change observed for two standard deviations from the mean ratios in the Affymetrix GeneChip experiment. The fold change cut-offs used for GATA1, SCL and E2A were 1.39, 1.41 and 1.39 respectively. Ten out of fourteen of the differentially-expressed genes in GATA1 knockdown were found to have a fold change above the cut-off except for E2A, LMO2, LZTFL1 and TBX1 (Figure 4.16 A). For SCL and E2A, the validation rates were substantially lower (Figure 4.16 B and C). Only 1 gene (ZNF117) out of 11 genes and 2 genes (GTF2I and HHEX) out of 10 genes were validated for SCL and E2A respectively (Figure 4.16 D). Overall, a validation rate of 37% (13/35 assays) was achieved for the target genes studied. This data would suggest that the Affymetrix GeneChips identified a relatively high proportion of false positives in the knockdown studies described here.

Name of putative target gene	Transcription factor regulating target	Mode of regulation	Protein subunit/ family	Functions of putative target gene
ASCL2	GATA1	Activation	bHLH family	lineage-specific transcription factors essential for development of the trophectoderm
CITED2	GATA1, SCL, E2A	Activation	C-terminal domain binds to CBP/p300 CH1 domain	Transactivates transcription factor AP2, an important regulator of neural and cardiac development
GFI1B	GATA1	Activation	Zinc finger protein	Represses transcription by recruiting corepressors and histone modifiers such as histone deacetylases (HDACs). Plays important roles in erythropoiesis.
LZTFL1	GATA1	Activation	Leucine zipper family	
MYC	GATA1, SCL, E2A	Activation	MYC family bHLH/Leucine Zipper domain	Oncogene of leukemia. Activates transcription of growth-promoting genes and represses growth-arrest genes by dimerizing MAX. Induces epigenetic reprogramming of human cells to pluripotency.
NFE2	GATA1	Activation	Leucine zipper family	Activates β-globin gene expression. Required for megakaryocytes maturation and platelet production.
TBX1	GATA1	Activation	T-box DNA binding domain family	Required for the development of epithelial cells and auditory organs
EKLF	GATA1	Activation	Krüppel-like factor family Zinc finger protein	Expressed in erythroid lineage. Activates β-globin gene expression.
LMO2	GATA1, E2A	Activation	LIM domain protein	Regulates erythroipoietic and endothelial development. Member of the SCL erythroid complex.
LDB1	GATA1	Activation	LIM-domain interacting protein	Regulates developmental processes. Member of the SCL erythroid complex.
E2A	GATA1, SCL	Activation	bHLH family	Activates transcription of B-cell specific genes. Regulates B-cell lineage development. Member of the SCL erythroid complex.
PPARD	GATA1	Repression	Peroxisome proliferator-activated receptor (PPAR) superfamily	Represses transcription of adipogenesis.
TNFRSF1A	GATA1	Repression	Tumor necrosis factor receptor superfamily	Required for inflammatory response
DLX4	GATA1	Repression	Homeobox family	Represses β-globin gene expression by binding to two silencer elements.
BCL6	SCL	Activation	Zinc finger protein	Acts as a sequence-specific transcriptional repressor by recruiting histone deacetylases.

				Chromosomal translocation results in B-cell lymphomas.
BHLHB2	SCL, E2A	Activation	bHLH family	Regulates chondrocyte differentiation via cAMP pathway.
ID2	SCL	Activation	ID family HLH protein domain	Inhibits function of bHLH transcription factors by heterodimerisation in a dominant ne manner. Regulate cell proliferation and differentiation.
ZNF117	SCL	Activation	Zinc finger protein	Unknown
ZNF304	SCL	Activation	Zinc finger protein	Unknown
ZNF281	SCL	Repression	Zinc finger protein	Unknown
ATRX	SCL	Repression	Zinc finger protein (PHD finger)	Mutations in the XH2 gene cause the alpha-thalassemia/mental retardation syndrome. Represses α-globin expression. Interacts with EZH2, a chromatin regulator.
RUNX1	SCL	Repression	RUNX family Runt domain protein	Chromosome translocations of RUNX1 are associated with leukaemia. Fusion partner of ETO in acute leukaemia. Required for definitive haematopoiesis and bone cell development.
HHEX	E2A	Activation	Homeobox family	Functions as a transcriptional repressor in liver cells and may be involved in the differe and/or maintenance of the differentiated state in hepatocytes. Implicated in haematopoietic and endothelial development. Regulatory region contains the SCL stem cell enhancer.
GTF2I	E2A	Activation	Zipper-like motif Helix-loop/span-helix motif	General transcription factor. Subunit of a chromatin-modifying complex.
MBNL2	E2A	Activation	Zinc finger protein	Implicated in myotonic dystrophy, a neuromuscular disorder.
RUNX2	E2A	Activation	RUNX family Runt domain protein	Master regulator of bone development. Transcriptional regulator of bone lineage specific genes.
FOXO3A	E2A	Repression	Forkhead domain	Chromosomal translocation involved in acute lymphoblastic leukaemia. Triggers apoptosis by inducing the expression of genes that are critical for cell death. Regulates erythroid development. Implicated in haematopoietic cell renewal.
HOXA1	E2A	Repression	Homeobox family	Implicated in neural, inner ear and cardiovascular development.

Table 4.3. Differentially-expressed genes of GATA1, SCL and E2A selected for validation by quantitative PCR. The transcription factor regulating the putative target genes, mode of regulation, protein family and functions of putative target genes are listed in each column.


A
Validation of differentially-expressed genes in GATA1 KD
% of mRNA remained in GATA1 knockdown relative to firefly pGL3 luciferase control
300
250
200
150
100
50
0
ASCL2
CITED2
GFI1B
LZTFL1
MYC
NFE2
TBX1
EKLF
LMO2
LDB1
E2A
PPARD*
TNFRSF1A*
DLX4*
Differentially-expressed genes in GATA1 KD
B
Validation of differentially-expressed genes in SCL KD
% of mRNA remained in SCL knockdown relative to firefly pGL3 luciferase control
160
140
120
100
80
60
40
20
0
BCL6
BHLHB2
CITED2
ID2
MYC
ZNF117
ZNF304
E2A
ZNF281*
ATRX*
RUNX1*
Differentially-expressed genes in SCL KD
C
Validation of differentially-expressed genes in E2A KD
% of mRNA remained in E2A knockdown relative to firefly pGL3 luciferase control
140
120
100
80
60
40
20
0
BHLHB2
CITED2
HHEX
GTF2I
MBNL2
MYC
RUNX2
LMO2
FOXO3A*
HOXA1*
Differentially-expressed genes in E2A KD

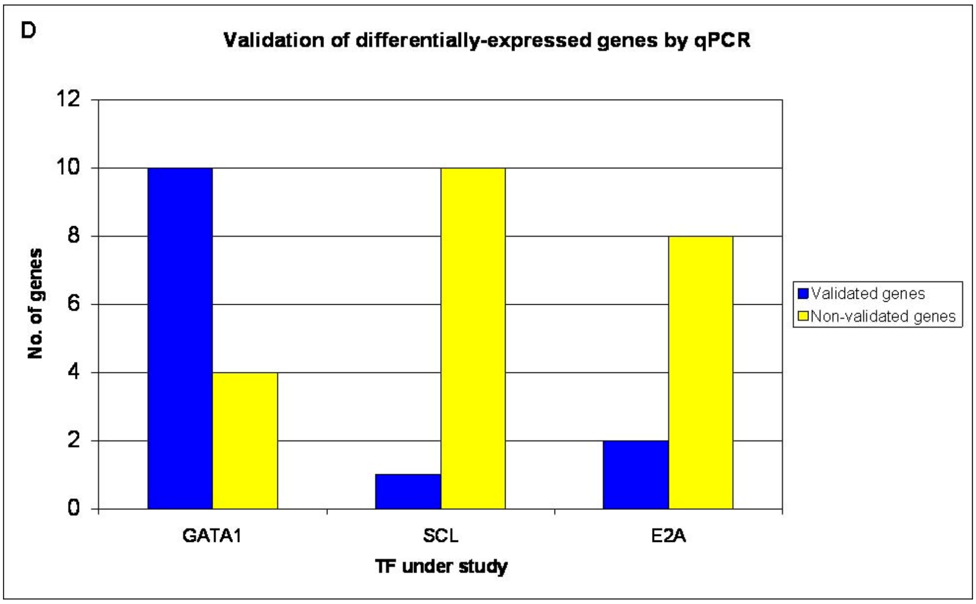


Figure 4.16. Validation of differentially-expressed genes by quantitative PCR. Expression of differentially-expressed transcription factor of GATA1, SCL and E2A were studied by quantitative PCR. mRNA of differentially-expressed genes was compared between the specific knockdown and the luciferase control in two independent biological replicates. In panels A, B and C, the bar charts shows the mRNA level of differentially-expressed genes in the knockdown compared to the control in the qPCR analyses. The error bars show the standard error between the two independent biological replicates. The genes marked with an asterisk are genes identified to be up-regulated in the original Affymetrix experiments. Panel A, validation of differentially-expressed genes in GATA1 knockdown; panel B, validation of differentially-expressed genes in SCL knockdown; panel C, validation of differentially-expressed genes in E2A knockdown. Panel D showed the number of validated and non-validated differentially-expressed genes for each transcription factor under study with the selected fold change cut-off for each knockdown.

The changes in mRNA expression of the chosen differentially-expressed genes were compared between the results obtained in the quantitative PCR and Affymetrix GeneChip (Table 4.4). In general, the changes in mRNA expression were shown to be larger in the results obtained in the Affymetrix GeneChip. The median coefficient of variation of the validated genes (labelled in yellow boxes in Table 4.4) was 9.74% while that of the non-validated genes was 43.37%. This indicates the change in expression of the non-validated genes deviated more than 4 times more from the Affymetrix GeneChip data the validated gene set.

A) GATA1 KD

	% of mRNA remained		
Differentially expressed gene	Affy Gene Chip	qPCR	CV
ASCL2	42.27	40.74	2.60
CITED2	37.03	51.79	23.49
GFI1B	36.94	46.20	15.76
LZTFL1	58.89	83.69	24.60
MYC	42.47	58.41	22.34
NFE2	46.10	71.90	30.93
TBX1	44.80	90.89	48.04
EKLF	13.79	32.08	56.38
LMO2	53.86	103.81	44.81
LDB1	59.38	65.58	7.02
E2A	54.13	73.17	21.16
PPARD*	174.85	200.72	9.74
TNFRSF1A*	310.87	239.91	18.22
DLX4*	149.59	156.94	3.39

B) SCL KD

	% of mRNA remained		
Differentially expressed gene	Affy Gene Chip	qPCR	CV
BCL6	36.88	111.93	71.33
BHLHB2	47.38	98.77	49.73
CITED2	46.59	85.06	41.33
ID2	30.05	99.20	75.66
MYC	63.15	95.43	28.79
ZNF117	54.26	58.38	5.17
ZNF304	57.58	88.42	29.87
E2A	65.21	134.08	48.88
ZNF281*	140.96	68.72	48.72
ATRX*	199.55	78.57	61.51
RUNX1*	142.87	69.39	48.95

C) E2A KD

	% of mRNA remained		
Differentially expressed gene	Affy Gene Chip	qPCR	CV
BHLHB2	46.03	90.83	46.29
CITED2	49.46	91.15	41.93
HHEX	58.86	54.58	5.34
GTF2I	54.57	62.02	9.03
MBNL2	53.32	82.19	30.13
MYC	52.94	92.89	38.74
RUNX2	57.81	85.38	27.23
LMO2	47.14	105.80	54.25
FOXO3A*	159.09	102.03	30.90
HOXA1*	158.44	110.79	25.03

Table 4.4. Comparison of changes in mRNA expression of differentially-expressed genes between Affymetrix GeneChip and quantitative PCR. The % of mRNA remained after siRNA knockdown analysed in Affymetrix GeneChip and qPCR and the coefficient of variation (CV) between the results obtained in the two assays are shown in the tables. Table A: differentially-expressed genes in GATA1 knockdown; Table B: differentially-expressed genes in SCL knockdown; Table C: differentially-expressed genes in E2A knockdown. The validated genes (selected with a cut-off of described above) are highlighted in yellow. Up-regulated genes are marked with an asterisk.

4.4.3.4 Further study and classification of differentially-expressed genes

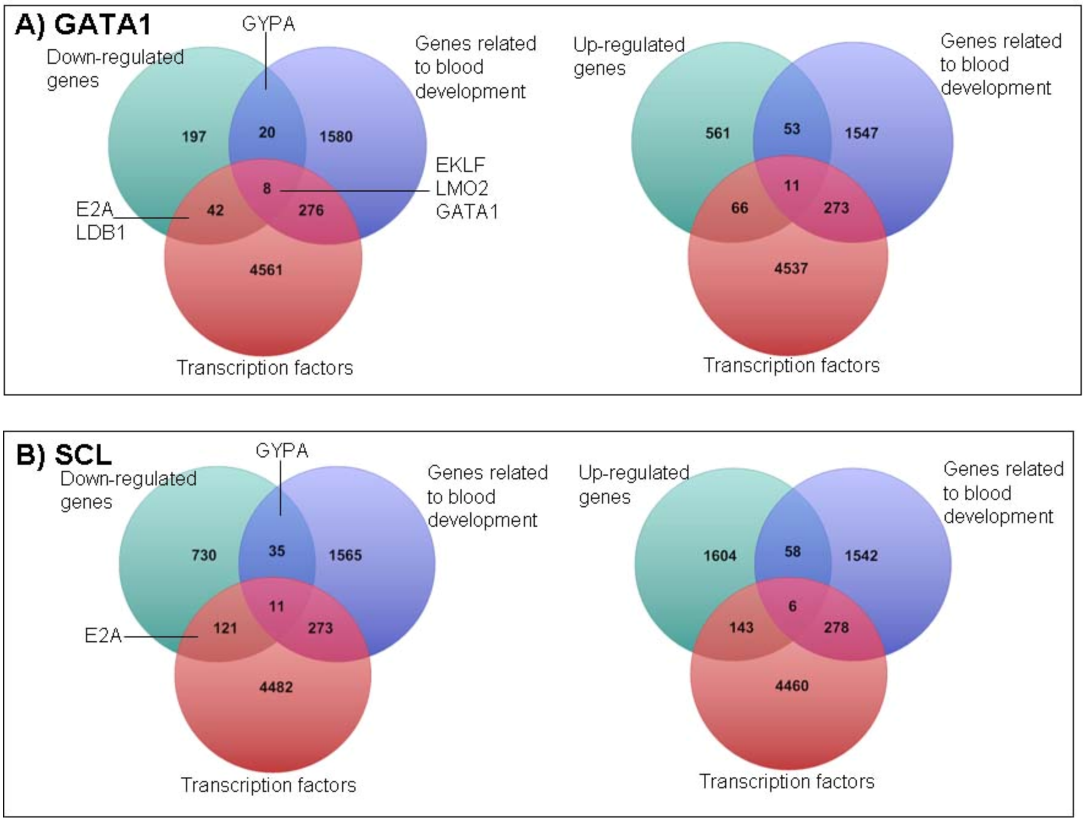
To provide evidence for the reliability of the Affymetrix GeneChip datasets, the up- or down-regulated probes identified from the analysis above were further studied in terms of their functional classifications, comparison with published target genes and auto-regulation of the SCL erythroid complex.

(i) Transcription factors and genes involved in haematopoiesis

The putative target gene lists derived from the transcription factor knockdown experiments were examined to (i) identify those genes that have been previously shown to be involved in haematopoiesis, and (ii) those genes that were transcription factors. This would allow us to determine (i) whether targets of the SCL erythroid complex identified in K562 cells were representative of haematopoiesis, and (ii) allow a direct comparison of transcription factor targets found in ChIP-on-chip studies using a transcription factor promoter array (see Chapter 5). Venn diagrams were used to study the number of transcription factors and haematopoietic-specific genes for each activated or repressed gene list (Figure 4.17). Gene lists for transcription factors and haematopoietic-specific genes were defined by Philippe Couttet and David Vetrie (Sanger Institute) using lists of all known human transcription factors downloaded from ENSEMBL (including transcription factors and chromatin modifiers/remodelers) and genes known to be expressed and have specific roles during haematopoiesis (including genes important in both haematopoietic and endothelial lineages since both share a common early precursor, the haemangioblast). In total, 1884 and 4887 probe sets found on the Affymetrix GeneChips were found to represent haematopoietic-specific genes and genes encoding transcription factors respectively. These figures were used to derive the percentages of target genes in each class as shown in Table 4.5. P-values associated with each class were also derived using the chi-squared test, which tests a null hypothesis at the frequency distribution of certain events observed in a sample is consistent with a particular theoretical distribution.

Based on the proportion of probe sets on the Affymetrix arrays which were haematopoietic-specific (1884 out of 54614), the likelihood of detecting differentially-expressed haematopoietic-specific genes by chance was approximately 3.4%. However, 9.5% (P-value <0.0001) of the GATA1 target genes were haematopoietic-specific, suggesting that the GATA1 knockdown experiment was able to enrich for the identification of haematopoietic-specific genes. Similarly, E2A perturbed in the knockdown experiments also showed enrichment for haematopoietic-specific genes (6.8%, P-value 0.0190). This provided confidence that the K562 biological system and the experimental approach were not identifying random events unrelated to blood development. However, no significant enrichments were seen for SCL, LDB1 and LMO2 (3.8% to 4.3%) (see Discussion of this Chapter).

Based on the proportion of probe sets on the Affymetrix arrays which were specific for genes encoding transcription factors (4887 out of 54614), the likelihood of detecting differentially expressed genes encoding transcription factors by chance was approximately 8.9%. However, 14.9% (P-value 0.0360) of target genes identified by the knockdown experiment of E2A were transcription factors (P-value<0.05 is considered to be significant). Between 9-13.3% of the target genes for the other three transcription factors (SCL, GATA1, LMO2 and LDB1) were transcription factors. However, only the down-regulated gene lists of SCL (14.7%, P-value 0.0360) and LMO2 (22.2%, P-value<0.0001) were enriched with transcription factors. This suggests that at least one transcription factor specifically enriched for other transcriptional regulators, suggesting that the SCL erythroid complex may have an important role in regulating transcriptional cascades in K562 cells (see also the discussion for this Chapter).



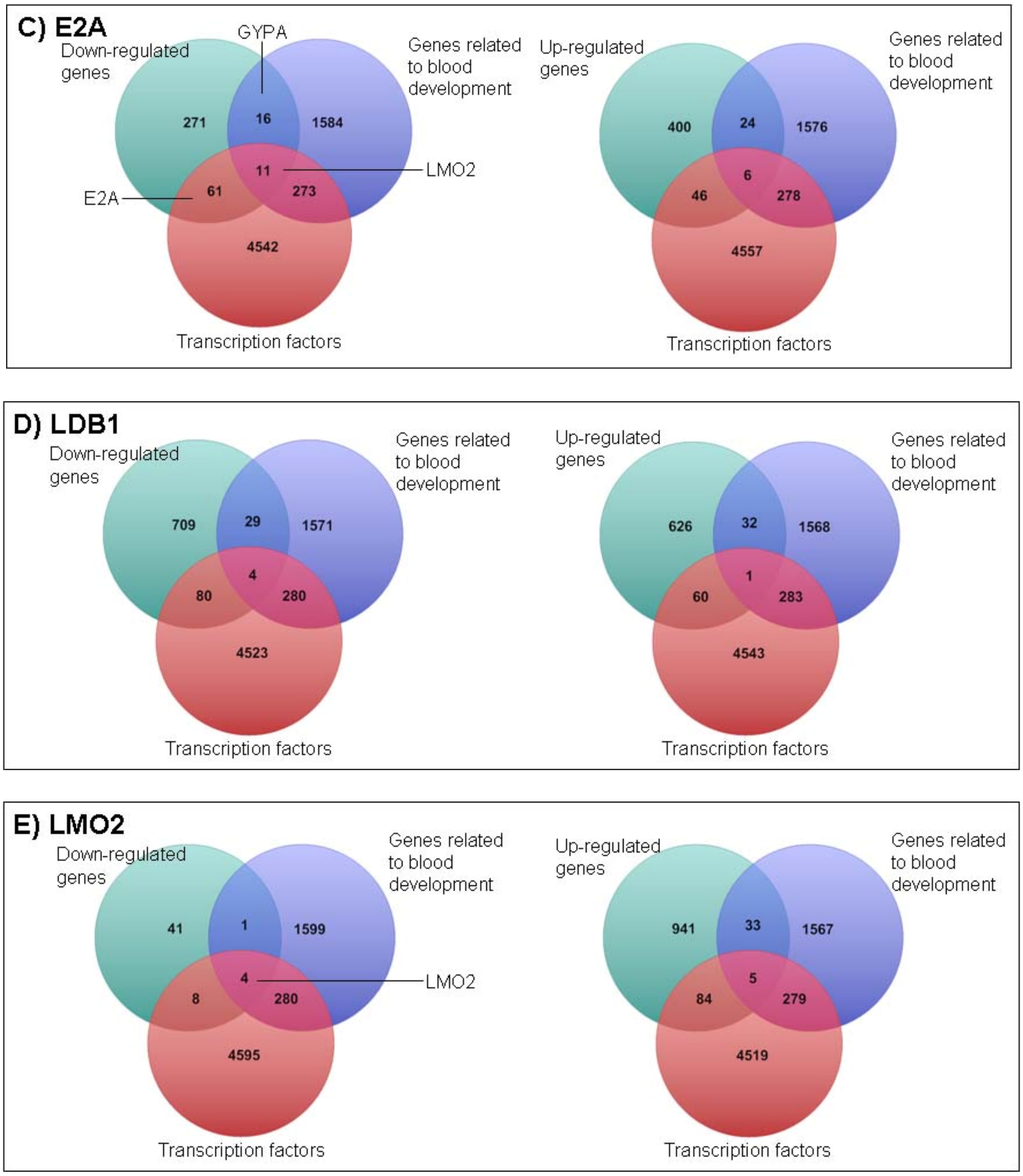


Figure 4.17. Venn diagrams comparison of up- or down-regulated gene lists with haematopoietic-specific gene list and transcription factors. Down-regulated genes are shown in the Venn diagrams on the left while up-regulated genes are shown in the Venn diagrams on the right. Numbers shown in the Venn diagrams are numbers of probe sets representing different or same genes in the human genome. In each Venn diagram, the top left green circle represents the up- or down-regulated genes picked up in the siRNA knockdown study, the top right blue circle represents the haematopoietic genes and the lower red circle represents transcription factors. Some interesting target genes are labelled in the Venn diagram. Panel A: Venn diagrams of GATA1; panel B: Venn diagrams of SCL; panel C: Venn diagrams of E2A; panel D: Venn diagrams of LDB1; panel E: Venn diagram of LMO2.

TF	Category (HG: haematopoietic genes; TF: transcription factor)	Down-regulated genes		Up-regulated genes		All differentially-expressed genes	
		Percentage	P-value	Percentage	P-value	Percentage	P-value
GATA1	HG	10.5%	<0.0001	9.3%	0.0004	9.5%	<0.0001
	TF	18.7%	0.0005	11.1%	0.4846	13.3%	0.1622
SCL	HG	5.1%	0.2410	3.5%	0.5577	4.1%	0.5577
	TF	14.7%	0.0360	8.2%	0.7268	10.4%	0.7268
E2A	HG	7.5%	0.0034	6.3%	0.0786	6.8%	0.0190
	TF	20.1%	0.0001	10.9%	0.4846	14.9%	0.0360
LDB1	HG	4%	0.5577	4.6%	0.2410	4.3%	0.5577
	TF	10.2%	0.7268	8.5%	1.0000	9.4%	1.0000
LMO2	HG	9.3%	0.0004	3.6%	0.5577	3.8%	0.5577
	TF	22.2%	<0.0001	8.4%	0.7268	9%	1.0000

Table 4.5. Percentages and P-values of haematopoietic genes and transcription factors in the differentially-expressed gene lists for each member of the SCL erythroid complex.

(ii) Gene Ontology classification

The activated or repressed genes for each transcription factor were also classified according to the terms found in the Gene Ontology (GO) database. The GO project describes functions of gene products in three different categories: cellular components, biological processes and molecular functions. The differentially-expressed genes of each transcription factor knockdown were studied to identify statistically significant GO terms using GO Term Finder (http://go.princeton.edu/cgi-bin/GOTermFinder) (Boyle et al., 2004). GO terms which are over-represented in the differentially-expressed genes lists compared to the whole human genome with a P-value of <0.01 were identified. The GO terms in the three categories, the associated P-values, the percentage in the differentially-expressed gene lists and in the human genome are included in Appendix 2.

The GO terms in the biological process and molecular function categories which appeared in more than one differentially-expressed gene list are shown Table 4.6. Three GO biological process terms (chromatin modification, regulation of transcription from RNA polymerase II promoter and transcription from RNA polymerase II promoter) are related to the regulation of gene expression. Four molecular function terms (transcription activator activity, transcription regulator activity, transcription cofactor activity and transcription factor binding) are also related to transcription. This illustrates that the five members of the SCL erythroid complex regulate a number of downstream target genes which are related to the regulation of transcription. Nine of the GO biological process terms are related to programmed cell death or apoptosis. This indicates that members of the SCL erythroid complex may also regulate a number of genes related to the apoptotic pathway.

	GATA1		SCL		E2A		LDB1		LMO2	
Biological process	Down-regulated	Up-regulated	Down-regulated	Up-regulated	Down-regulated	Up-regulated	Down-regulated	Up-regulated	Down-regulated	Up-regulated
regulation of cell proliferation		🟦	🟦		🟦					
regulation of developmental process		🟦	🟦	🟦						
regulation of macromolecule metabolic process			🟦		🟦					
regulation of programmed cell death		🟦	🟦							
regulation of signal transduction		🟦	🟦							
anatomical structure development		🟦		🟦				🟦		🟦
apoptosis		🟦	🟦							
biological regulation		🟦	🟦							
biopolymer metabolic process			🟦				🟦			
cell cycle			🟦		🟦		🟦			
cell cycle process			🟦				🟦			
cell death		🟦	🟦							
cell motion		🟦		🟦						
cell proliferation		🟦	🟦		🟦					
cellular component organization and biogenesis			🟦		🟦		🟦			
cellular developmental process			🟦	🟦						🟦
cellular metabolic process			🟦		🟦					
chromatin modification	🟦		🟦							
death		🟦	🟦							
developmental process		🟦	🟦	🟦				🟦		🟦
endomembrane system			🟦		🟦		🟦			
establishment of protein localization			🟦		🟦		🟦			
gene expression			🟦		🟦		🟦			
intracellular signaling cascade		🟦	🟦					🟦		
macromolecule metabolic process			🟦				🟦			
localization of cell		🟦		🟦						
macromolecule localization			🟦		🟦		🟦			
mRNA processing			🟦				🟦			
multicellular organismal development		🟦		🟦				🟦		🟦
multicellular organismal process		🟦		🟦				🟦		
negative regulation of apoptosis		🟦	🟦							
negative regulation of biological process		🟦	🟦	🟦		🟦		🟦		🟦
negative regulation of cell proliferation		🟦	🟦							
negative regulation of cellular process		🟦	🟦	🟦		🟦		🟦		🟦
negative regulation of developmental process		🟦	🟦							
negative regulation of programmed cell death		🟦	🟦							
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process			🟦		🟦		🟦			
organ development		🟦	🟦							
positive regulation of biological process		🟦	🟦	🟦	🟦					
positive regulation of cellular process		🟦	🟦	🟦	🟦					
positive regulation of developmental process		🟦		🟦						
primary metabolic process			🟦				🟦			
programmed cell death		🟦	🟦							
protein kinase cascade		🟦	🟦					🟦		
protein localization			🟦		🟦		🟦			
protein transport			🟦		🟦		🟦			
regulation of apoptosis		🟦	🟦							
regulation of cell cycle			🟦		🟦		🟦			
regulation of cell proliferation		🟦	🟦							
regulation of cellular metabolic process			🟦		🟦					
regulation of developmental process		🟦	🟦	🟦						
regulation of macromolecule metabolic process			🟦		🟦					
regulation of metabolic process			🟦		🟦					
regulation of programmed cell death		🟦	🟦							
regulation of transcription from RNA polymerase II promoter	🟦				🟦					
RNA metabolic process	🟦		🟦		🟦		🟦			
RNA splicing			🟦				🟦			
system development		🟦		🟦				🟦		🟦
transcription from RNA polymerase II promoter	🟦		🟦		🟦					
Molecular function										
enzyme binding		🟦	🟦		🟦					
kinase binding		🟦	🟦							
protein binding	🟦	🟦	🟦		🟦	🟦	🟦	🟦		
RNA binding	🟦		🟦		🟦		🟦			
transcription activator activity			🟦		🟦					
transcription cofactor activity					🟦		🟦			
transcription factor binding	🟦				🟦					
transcription regulator activity	🟦				🟦					

Table 4.6. Gene Ontology classification of differentially-expressed genes for each of the five members of the SCL erythroid complex. The GO terms associated with biological processes (top) and molecular functions (bottom) and significantly enriched in more than one of the differentially-expressed gene lists are shown. The blue boxes indicate the GO terms which are statistically significant in the up- or down-regulated gene lists in the knockdown study of five members of the SCL erythroid complex (P value < 0.1).

(iii) Identification of published target genes

The differentially-expressed genes were compared with the published target genes of the transcription factors (see Chapter 1, section 1.4.2). GATA1, SCL and E2A were all found to regulate one of the three known target genes of the SCL erythroid complex - GYPA (c-kit and α-globin being the other two). GATA1 was shown to regulate 6 out of the 11 published target genes and these included GYPA, EKLF, NFE2, EPOR, MYC and GFI-1B. This indicates that the siRNA-induced knockdown in combination with expression profiling with the GeneChip identified at least some published targets for these transcription factors.

(iv) Auto-regulation of the SCL erythroid complex

Based on the Affymetrix data, the transcription factors of the SCL erythroid complex were also found to regulate other members of the complex itself. For example, GATA1 activated expression of E2A, LDB1 and LMO2. SCL activated expression of E2A while E2A activated expression of LMO2. This indicates that members of the SCL erythroid complex are involved in auto-regulatory loops to regulate the transcription of other proteins involved in the complex.

4.4.3.5 Co-regulation of transcription factors in the SCL erythroid complex

Whilst each of the transcription factors studied here may function alone or in combination with other transcription factors in regulating gene expression, the aim of this project was to identify targets of the SCL erythroid complex. Therefore comparing the differentially-expressed gene lists for each transcription factor and determining which genes were found in more than one list would provide insights into which genes are targets of the SCL erythroid complex. Gene lists were analysed in several ways, by varying the number of members of the SEC in the comparisons, and by including data at the expression outcome (activated or repressed) of the targets (since it was likely that *bona fide* targets of the SCL erythroid complex would be affected in the same way during knockdown of any one of the five transcription factors).

(i) Identification and classification of co-regulated genes

Initially, the putative target genes of GATA1, SCL and E2A were compared as these three transcription factors are bound to DNA directly in the SCL erythroid complex. 102 probe sets representing 92 genes were found to be co-regulated by GATA1, SCL and E2A (Figure 4.18 A).

These 102 probe sets were further classified and studied (see below). To further assess the roles played by the bridging proteins LDB1 and LMO2 in the SCL erythroid complex, the putative target gene lists of GATA1, SCL and E2A were also compared against those of LDB1 and LMO2. Unlike the co-regulation among GATA1, SCL and E2A, only a very small portion of genes (up to 7 probes) were found to be co-regulated by the 3 transcription factors: LDB1, LMO2 and either GATA1, SCL or E2A (Figure 4.18 B-D). However, no genes were found to be co-regulated by all five members of the complex.

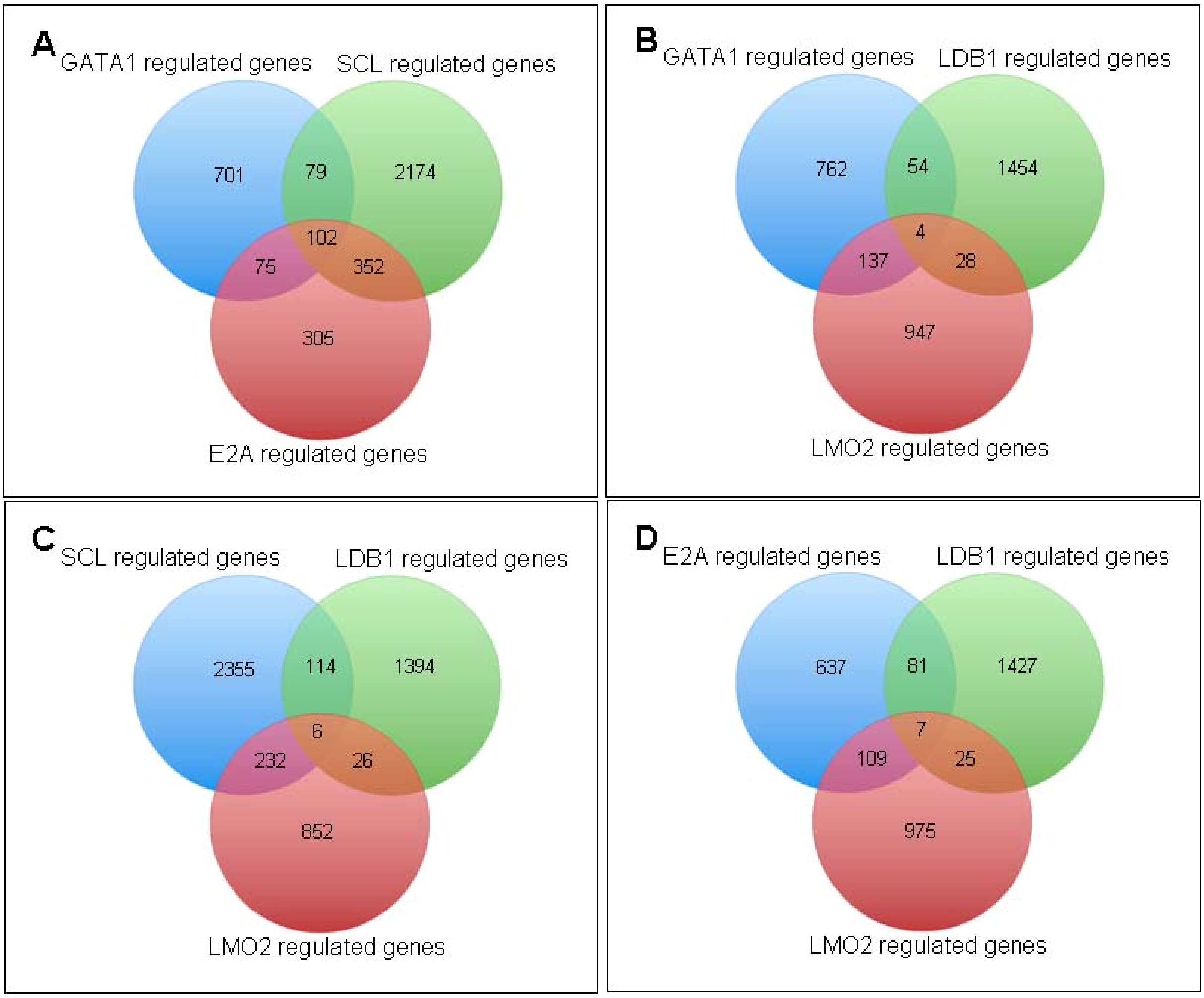


Figure 4.18. Co-regulation of target genes by members of the SCL erythroid complex. Numbers shown in the pie charts are numbers of probe sets representing genes in the human genome. Panel A: co-regulation of GATA1, SCL and E2A; panel B: co-regulation of GATA1, LDB1 and LMO2; panel C: co-regulation of SCL, LDB1 and LMO2; panel D: co-regulation of E2A, LDB1 and LMO2.

Within the group of 92 genes found to be co-regulated by GATA1, SCL and E2A, 19 were transcription factors. Therefore, not surprisingly, these 92 genes were enriched in GO terms related to transcription from RNA polymerase II promoter (Table 4.7). In addition, these 92 genes were also enriched in the protein binding GO term which indicates that these genes may be involved in protein-protein interaction required for the regulation of transcription. These GO classifications again reinforce the idea that the SCL erythroid complex may play a critical role in transcriptional regulation by regulating other transcription factors and associated factors which are involved in the regulation of transcription and signal transduction activities.

Biological process	**P-value**	**% in gene list**	**% in genome**
regulation of transcription from RNA polymerase II promoter	0.00347	8.51	1.28
RNA metabolic process	0.00467	26.60	12.33
transcription from RNA polymerase II promoter	0.00602	9.57	1.82
Molecular function			
Protein binding	3.6E-06	48.94	26.66
RNA binding	6.3E-05	13.83	2.87
Cellular component			
intracellular part	8.6E-05	58.51	39.25
intracellular organelle	8.8E-05	52.13	32.57
Organelle	8.9E-05	52.13	32.58
intracellular membrane-bounded organelle	0.00013	45.74	26.73
membrane-bounded organelle	0.00013	45.74	26.74
Nucleus	0.00369	32.98	18.30
Intracellular	0.00773	58.51	44.61

Table 4.7. Gene Ontology classification of GATA1, SCL and E2A co-regulated genes. GO terms associated with biological process, molecular function and cellular component significantly enriched in the GATA1, SCL and E2A co-regulated genes are shown. The P-values associated with each GO term, percentage of genes belonging to the GO term in the gene list and in the human genome are also shown.

(ii) Identification of known co-regulated target genes

Within these 92 genes co-regulated by GATA1, SCL and E2A, glycophorin A (GYPA) - which is a published known target of the SCL erythroid complex - was identified. This confirmed that the Affymetrix expression data could detect at least one co-regulated target out of the 3 published co-regulated genes (GYPA, c-kit and α-globin) found in the SCL erythroid complex.

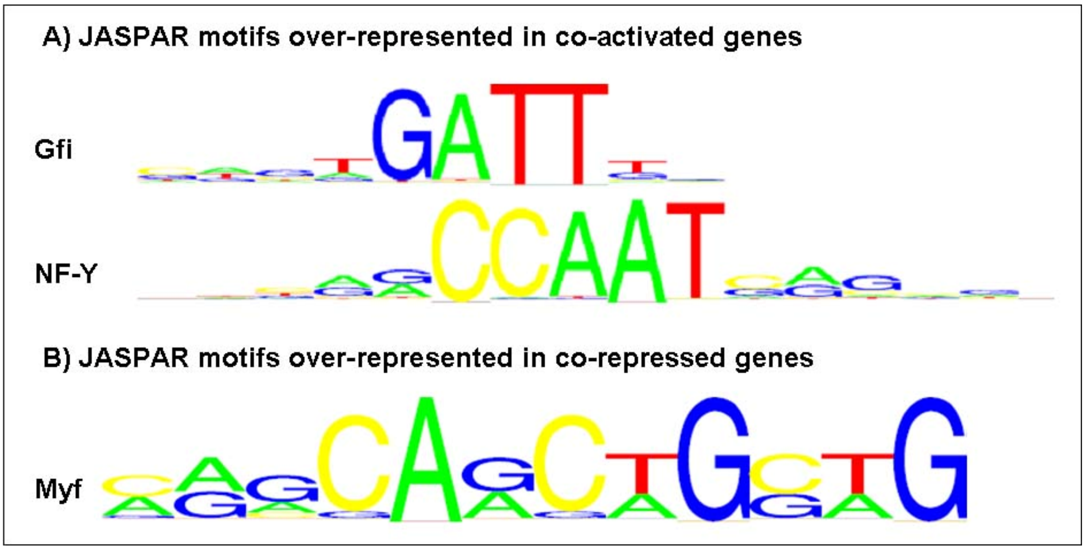
(iii) Auto-regulation of the SCL erythroid complex

The auto-regulation of the SCL erythroid complex (first mentioned in section 4.4.3.4 part iv) also appeared to extend to co-regulation by more than one member of the complex. Two members of the SEC were identified as putative target genes of GATA1, SCL and E2A. E2A was shown to be activated by both GATA1 and SCL while LMO2 was shown to be activated by GATA1 and E2A.

4.4.4 Motif discovery of co-regulated putative target genes

It is known that the SCL erythroid complex binds to a composite E-box/GATA motif (Wadman et al., 1997) which directly binds to the SCL/E2A heterodimer and GATA1. As a means of confirming whether the putative target genes identified by the knockdown experiments were *bona fide*, motif analysis was performed to determine whether this motif, or any similar ones, was found in common for the putative target genes of the transcription factor knockdown experiments. The 92 genes co-regulated by SCL, GATA1 and E2A (described in section 4.4.3.5) were studied by two methods to identify transcription factor binding motifs in a one kilobase region covering their known promoter regions identified using FirstEF (Davuluri et al., 2001). Promoter regions were chosen for this analysis, although regulation involving enhancers may play a crucial role as well – however the location of any enhancers was not known. In the first method, the vertebrate motif database JASPAR CORE was used to identify known transcription factor binding motifs that were over-represented within this 1 kb region around the transcription start sites of the genes. The JASPAR CORE database is an open-access database containing curated, non-redundant transcription factor binding site profiles for multicellular eukaryotes which were derived from experimentally verified DNA sequences bound by transcription factors (Sandelin et al., 2004). Two transcription factor binding motifs, Gfi and NF-Y, were found to be over-represented in the co-activated gene list while one motif Myf was over-represented in the co-repressed gene list (Figure 4.19 A and B). The Gfi motif is recognised by the zinc finger protein Gfi family containing the C2H2 motif (Zweidler-Mckay et al., 1996). The NF-Y motif is recognised by the nuclear transcription factor Y family and has a characteristic CCAAT motif (Becker et al., 1991). The Myf motif is a bHLH motif recognised by the myogenic factor family (Wasserman and Fickett, 1998). The composite E-box/GATA motif was not identified by this analysis.

In the second method, the NestedMICA programme was used to perform unbiased motif discovery (Down and Hubbard, 2005) (Chapter 1, section 1.3.4.2). This method allowed us to identify possible novel DNA motifs in the promoter regions of the 92 genes co-regulated by GATA1, SCL and E2A. In this case, three DNA motifs were identified reproducibly in the promoters of the activated genes; no recurrent motifs were found in the repressed genes (Figure 4.19 C). Once again, the composite E-box/GATA motif was not identified by this analysis.


A) JASPAR motifs over-represented in co-activated genes
Gfi
NF-Y
B) JASPAR motifs over-represented in co-repressed genes
Myf

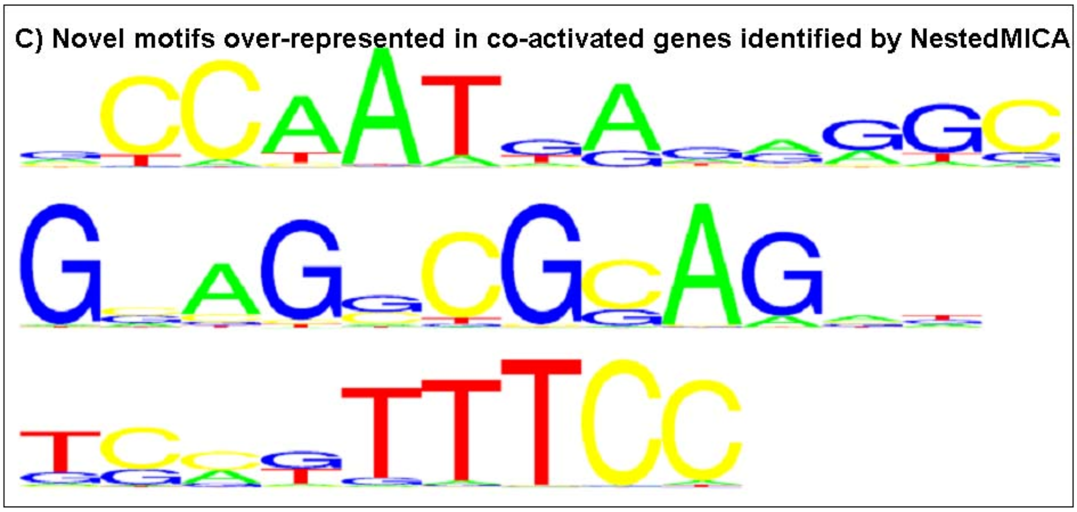

C) Novel motifs over-represented in co-activated genes identified by NestedMICA

Figure 4.19. Motif discovery of co-regulated genes by GATA1, SCL and E2A. Panel A: annotated motifs identified for the co-activated genes in the JASPAR database; panel B: annotated motifs identified for the co-repressed genes in the JASPAR database; panel C: novel motifs identified for the co-activated genes using NestedMICA. DNA logos are presented for each motif and the height of each letter indicates relative occurrence of nucleotides in the identified binding sites.

4.5 Discussion

The work presented in this Chapter demonstrated the use of a commercial expression array platform, the Affymetrix GeneChip, to study the effects on gene expression in the K562 cell line when members of the SCL erythroid complex were knocked down using siRNAs. This array analysis permitted the identification of putative target genes regulated by each transcription factor in the complex, as these targets would be very likely to change in their levels of expression during the knockdown conditions. The data presented in this Chapter will now be discussed as follows:

4.5.1 Affymetrix GeneChips as a platform of expression profiling

- **Low validation rate by qPCR**

The results obtained from the expression profiling of downstream effects of siRNA knockdown using Affymetrix GeneChip demonstrated that it is not a particular good method of identifying target genes of the transcription factors under study. This conclusion was based on the low validation rate of differentially-expressed genes by qPCR. This suggests that the differentially-expressed genes identified on the Affymetrix GeneChip included a large number of false positive targets. This may be because the TF knockdowns were not sufficient to elicit profound and reproducible changes in target gene expression profiles (because a proportion of the knocked down TF was still present). Thus, quantitative measurements on the Affymetrix platform may not detect such subtleties in expression changes or may have detected changes which were not reproducible between bioreplicates. This was supported by the fact that GATA1, which had the highest knockdown efficiency, showed the highest rate of qPCR validation.

- **Identification of published targets**

Despite the caveats mentioned above, known targets of members of the complex were identified by Affymetrix analysis. Taking GATA1 as an example, the majority of its published target genes was detected on the Affymetrix expression array. These included EPOR, GYPA, GFI1B, NFE2, MYC and EKLF. However, other published targets, α- and β-globin genes, Epo, FOG-1 and GATA2, were not detected. One of the key downstream targets of the SCL erythroid complex, glycophorin A (GYPA), was shown to be co-activated by GATA1, SCL and E2A in the expression study. However, GYPA was not detected as an activated gene by LDB1 and LMO2, although it is known to be a target gene of the whole SCL erythroid complex (Lahlil et al., 2004). Furthermore, the two other published targets of the SCL erythroid complex - c-kit and α-globin - were not detected by any members of the complex.

A number of reasons may explain why some of the published and novel targets of the complex may not have been identified in the GeneChip analysis:

(i) The knockdown of the transcription factors under study was not 100%. The remaining level of the transcription factors may be sufficient to drive the expression of their target genes. Thus, the change in expression of these target genes during the knockdown may not be significant or reproducible for detection on the expression array.

(ii) For target genes which are regulated by the whole SCL erythroid complex, some members of the complex may be dispensable for the regulation. This may be particularly relevant to LDB1 and LMO2, which do not bind DNA directly, but are bridging proteins. The roles of these proteins may be to stabilise the complex and not participate in direct regulation *per se*.

(iii) It is also possible that other transcription factors, apart from the five members of the complex in question, can compensate for the knockdown effects, thus allowing regulation of target genes even in the absence of a member of the complex.

(iv) The three DNA binding partners (GATA1, SCL and E2A), or combinations thereof, are able to interact within other regulatory complexes which do not include LMO2 or LDB1. This would add an additional layer of complexity onto the analysis and make gene list comparisons more complex.

(v) It may be difficult to determine co-regulation by multiple members of the SCL eyrthroid complex because they could also be acting on target genes independent of the SCL erythroid complex. Thus major effects may be elicited by some knockdowns, but not others.

(vi) Only one time point was studied on the expression array for each knockdown assay. The effects on gene expression may be transient, occur earlier than that was monitored or may take substantially more time after the silencing of the transcription factor. Therefore; not every target gene can be detected at the time point selected.

(vii) The stringency of fold change use in the statistical analyses may also be an issue. The expression changes of some target genes may be very subtle and not satisfy the criteria for selection as differentially-expressed genes.

4.5.2 The SCL erythroid complex regulates transcription factors

The differentially-expressed genes identified in the gene expression profiling for each transcription factor were over-represented for transcription factors. Between 9% and 14.9% of target genes identified by the five transcription factor knockdown experiments were transcription factors (section 4.4.3.4). This suggests that the SCL erythroid complex may play a crucial role regulating haematopoietic transcriptional networks in K562 cells. This makes sense, given the role of SCL as a master regulator of haematopoiesis.

4.5.3 Identification of haematopoietic-related genes regulated by members of the SCL erythroid complex

Enrichments of haematopoietic-related genes were observed in the differentially-expressed gene lists for members of the SCL erythroid complex during knockdown (section 4.4.3.4). The percentage of haematopoietic-specific genes of the 5 transcription factors ranged from 3.8% to 9.5%. This data confirms that knockdown of members of the SCL erythroid complex does induce changes to genes which have known roles in haematopoietic development. The percentage of haematopoietic-specific genes in the GATA1 study was the highest among the 5 transcription factors (9.5%). This is because it is an important regulator of erythroid development and the knockdown efficiency with siRNA was the highest for the five TFs studied. The percentages for SCL and LDB1 were the lower (4.1% and 4.3% respectively). For SCL, only one siRNA was used in the expression profiling and thus many of the differentially-expressed genes identified may be off-targets, thus resulting in a lower haematopoietic-specific effect. LDB1 is a ubiquitously-expressed gene. Therefore, its target gene list may reflect other cellular events than those associated purely with haematopoiesis. Yet, E2A is a ubiquitously-expressed gene and should also have identified a high degree of non-haematopoietic-related target genes. However, E2A identified 1.5 times as many haematopoietic targets as LDB1 (6.8%). This may be due to the fact that it is a known interacting partner of SCL, and such dimerisation is a requirement for DNA-binding (Hsu et al., 1994). LMO2 is expressed in haematopoietic progenitors and is required for erythropoiesis and theoretically, a large percentage of its putative target genes should also have been haematopoietic-related. However, LMO2 showed the lowest levels of enrichment for haematopoietic-related genes (3.8%). One possible reason is that the siRNA-induced knockdown could not be monitored at the protein level for LMO2 (due to the lack of an antibody which worked well in western analysis). Therefore, there was no way of knowing whether the time point chosen to identify relevant targets was appropriate.

4.5.4 Auto-regulation of the SCL erythroid complex

Previous studies have demonstrated that GATA1 is a regulator of SCL expression. Indeed, the SCL +51 enhancer has also been shown to be bound by at least three members of the SCL erythroid complex SCL, GATA1 and LDB1 (Chapter 1, section 1.4.2.1). Thus, SCL may indeed be regulated by the whole SEC complex or by no fewer than three of its members. The expression data obtained from this Chapter further characterised the auto-regulatory role of this complex. E2A was found to be activated by both GATA1 and SCL while LMO2 was activated by E2A and GATA1. LDB1 was also an activated target gene by GATA1. This data suggests that there are tightly controlled regulatory complexities which may govern the activity of the SEC.

This also highlights a further challenge to analysing the targets of the SEC - the knockdown of one member of the complex may change the expression of another member of the complex. This idea will be explored further in Chapter 6 of this thesis.

4.5.5 Motif discovery at target genes

The motif discovery analyses for the 92 genes co-regulated by GATA1, SCL and E2A offered little insights into their regulation by the transcription factors of the SCL erythroid complex. This was because the expected E-box/GATA composite motif was not identified in the promoters of this set of target genes. However, there are plausible explanations for this. Firstly, a 1 kb region around the TSS was used in these motif discovery analyses. Gene regulation and binding by transcription factors may occur outside this 1 kb window at regulatory elements such as proximal or distal enhancers, or silencers or even at distal promoters. Secondly, the siRNA-induced knockdown in combination with Affymetrix expression analysis identifies both direct and indirect target genes regulated by the transcription factor. Thus, the gene list which was used in the motif discovery contains both types of target genes - and it is highly unlikely that indirect targets would require an E-box GATA consensus motif. This would make consensus motifs difficult to derive from such a mixed set of targets. Despite this, a few additional motifs were identified in these targets. These motifs could possibly represent (i) new TFBS for other transcription factors which are required by direct targets which also bind the SEC, or (ii) are sites which bind factors required by both direct and indirect targets of the SEC to mediate transcriptional control at various levels of a transcriptional cascade. The delineation of direct targets of the SCL erythroid complex, from indirect ones, will be the basis of the next chapter in this thesis.

4.6 Conclusions

The work presented in this Chapter identified genes involved in haematopoiesis and transcription factors as the downstream targets of members of the SCL erythroid complex. However, the validation data demonstrated that the Affymetrix GeneChip platform generated a high false positive rate. Therefore; results from this Chapter should be carefully interpreted when further analyses are being performed and compared in the following Chapters.

Becker, D.M., Fikes, J.D., and Guarente, L. (1991). A cDNA encoding a human CCAAT-binding protein cloned by functional complementation in yeast. Proceedings of the National Academy of Sciences of the United States of America *88*, 1968-1972.

Boyle, E.I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J.M., and Sherlock, G. (2004). GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics (Oxford, England) *20*, 3710-3715.

Davuluri, R.V., Grosse, I., and Zhang, M.Q. (2001). Computational identification of promoters and first exons in the human genome. Nat Genet *29*, 412-417.

Down, T.A., and Hubbard, T.J. (2005). NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence. Nucleic Acids Res *33*, 1445-1453.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States of America *95*, 14863-14868.

Hsu, H.L., Huang, L., Tsan, J.T., Funk, W., Wright, W.E., Hu, J.S., Kingston, R.E., and Baer, R. (1994). Preferred sequences for DNA recognition by the TAL1 helix-loop-helix proteins. Mol Cell Biol *14*, 1256-1265.

Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B., and Speed, T.P. (2003a). Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res *31*, e15.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. (2003b). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics *4*, 249-264.

Lahlil, R., Lecuyer, E., Herblot, S., and Hoang, T. (2004). SCL assembles a multifactorial complex that determines glycophorin A expression. Mol Cell Biol *24*, 1439-1452.

Landry, J.R., Kinston, S., Knezevic, K., de Bruijn, M.F., Wilson, N., Nottingham, W.T., Peitz, M., Edenhofer, F., Pimanda, J.E., Ottersbach, K., *et al.* (2008). Runx genes are direct targets of Scl/Tal1 in the yolk sac and fetal liver. Blood *111*, 3005-3014.

Lin, Y.W., and Aplan, P.D. (2007). Gene expression profiling of precursor T-cell lymphoblastic leukemia/lymphoma identifies oncogenic pathways that are potential therapeutic targets. Leukemia *21*, 1276-1284.

Muntean, A.G., and Crispino, J.D. (2005). Differential requirements for the activation domain and FOG-interaction surface of GATA-1 in megakaryocyte gene expression and development. Blood *106*, 1223-1231.

Palomero, T., Odom, D.T., O'Neil, J., Ferrando, A.A., Margolin, A., Neuberg, D.S., Winter, S.S., Larson, R.S., Li, W., Liu, X.S., *et al.* (2006). Transcriptional regulatory networks downstream of TAL1/SCL in T-cell acute lymphoblastic leukemia. Blood *108*, 986-992.

Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W., and Lenhard, B. (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res *32*, D91-94.

Siddiqui, A.S., Delaney, A.D., Schnerch, A., Griffith, O.L., Jones, S.J., and Marra, M.A. (2006). Sequence biases in large scale gene expression profiling data. Nucleic Acids Res *34*, e83.

Singh-Gasson, S., Green, R.D., Yue, Y., Nelson, C., Blattner, F., Sussman, M.R., and Cerrina, F. (1999). Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. Nat Biotechnol *17*, 974-978.

Tusher, V.G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences of the United States of America *98*, 5116-5121.

Wadman, I.A., Osada, H., Grutz, G.G., Agulnick, A.D., Westphal, H., Forster, A., and Rabbitts, T.H. (1997). The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, E47, GATA-1 and Ldb1/NLI proteins. Embo J *16*, 3145-3157.

Wasserman, W.W., and Fickett, J.W. (1998). Identification of regulatory regions which confer muscle-specific gene expression. J Mol Biol *278*, 167-181.

Welch, J.J., Watts, J.A., Vakoc, C.R., Yao, Y., Wang, H., Hardison, R.C., Blobel, G.A., Chodosh, L.A., and Weiss, M.J. (2004). Global regulation of erythroid gene expression by transcription factor GATA-1. Blood *104*, 3136-3147.

Wu, Z., and Irizarry, R.A. (2005). Stochastic models inspired by hybridization theory for short oligonucleotide arrays. J Comput Biol *12*, 882-893.

Zweidler-Mckay, P.A., Grimes, H.L., Flubacher, M.M., and Tsichlis, P.N. (1996). Gfi-1 encodes a nuclear zinc finger protein that binds DNA and functions as a transcriptional repressor. Molecular and cellular biology *16*, 4024-4034.