Chapter 5

ChIP-on-chip analyses of the SCL erythroid complex

5.1 Introduction

The expression analyses of siRNA knockdown described in the previous Chapters allowed us to identify putative target genes regulated by each of five transcription factors found in the SCL erythroid complex. However, the limitations of these types of studies mean that they do not provide direct information regarding the binding of transcription factors to the regulatory regions of target genes. Complementary methods are required to further investigate such protein-DNA interactions at *cis*-regulatory elements of target genes, thus allowing such genes to be considered as *bona fide* direct target genes of the transcription factors. Many methods, both traditional and high-throughput, have been developed and characterised for the study of protein-DNA binding and for the identification of regulatory DNA elements (Chapter 1, section 1.3.3). Traditional low-throughput methods are time-consuming, and in many cases, they are based on DNA-protein binding *in vitro*. The development of ChIP-on-chip as an *in vivo* technique in the last decade has significantly enhanced the scale and spectrum of specificity for identifying transcription factor or other protein-bound DNA elements. At the time this project was first initiated, massively parallel sequencing had not been fully developed - microarrays were still playing the leading role in high-throughput genome-wide ChIP studies. Therefore, ChIP-on-chip analysis was used to identify direct targets for the five transcription factors of the SCL eythroid complex as described in this Chapter.

5.1.1 ChIP-on-chip: principles and issues

In ChIP-on-chip, cells or tissues are extracted and the DNA-protein complexes are cross-linked with formaldehyde. The cross-linked complexes are sonicated to shear the DNA into fragments – the amount of sonication determines the extent of shearing and typically the DNA is sheared to between 200 bp and 1 kb. The DNA-protein complexes are then immunoprecipitated with antibodies specific to a protein bound to the DNA. The immunoprecipitated and (non-immunoprecipitated “input” control sample) DNA-protein complexes are then de-crosslinked, and the ChIP and input DNA are extracted. Because of the amount of ChIP DNA recovered, it is quite often amplified by PCR prior to use in microarray analyses. The ChIP DNA and input DNA are then fluorescently labelled with two different dyes, such as Cy5 and Cy3, and hybridised onto genomic arrays of interest (Figure 5.1).

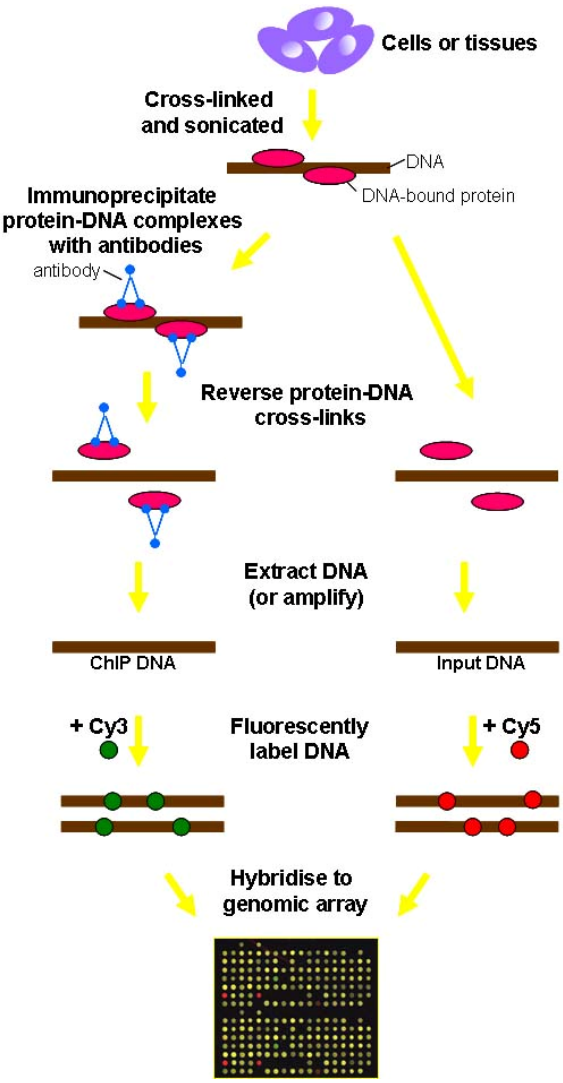


Figure 5.1. The principle of ChIP-on-chip. Flow diagram shows the steps involved in the method. Briefly, DNA-protein complexes are crosslinked, sonicated and immunoprecipitated with specific antibodies. Crosslinks of DNA-protein complexes are reversed and DNA extracted, labelled with fluorescent dyes and hybridised onto genomic arrays.

A number of issues should be considered carefully when performing ChIP-on-chip analyses. These factors ensure that the data obtained is of high quality. Some of them are discussed in details below.

- **Cross-linking**

Cross-linking between protein and DNA is the key factor affecting subsequent steps in a ChIP-on-chip experiment. Formaldehyde is commonly used for cross-linking between protein and DNA, as well as among proteins (Orlando et al., 1997). DNA elements bound by multiprotein complexes, where many of the protein components do not directly bind DNA, can also be studied. The type of protein-DNA interaction being cross-linked depends on the concentration of formaldehyde and the

length of time of cross-linking. As a result, different cell types or different protein-DNA interactions may require optimisation. For large DNA-binding protein complexes, long-range cross-linkers such as dimethyl adipimidate (DMA) can be used in combination with formaldehyde (Zeng et al., 2006). In cases where the protein-DNA interaction is relatively strong such as histone proteins, native ChIP, where no cross-linking is required, can be performed (O'Neill and Turner, 1996; O'Neill et al., 2006).

- **Antibodies**

The quality of antibodies used in ChIP-on-chip experiment is the most critical parameter determining the experimental outcome. Some commercial antibodies are validated and marketed for ChIP applications. However, for most antibodies, validation is performed by the experimenter, and often several antibodies are tested for each ChIP assay. To select antibodies that work well in ChIP, ChIP-qPCR of DNA regions where the protein is known to bind (if known) is useful to perform. Ultimately the best antibodies are those which can pick up specific protein *in vivo* and do not cross-react with other proteins or proteins of the same family. The easiest way to check the specificity of an antibody is by western blotting. To further ensure that the antibody only bind to the protein under study, siRNA knockdown can be used to silence the protein in the cell type and the knockdown of the relevant protein can be quantified by western blotting. Furthermore, the epitope recognised by the antibodies should be carefully selected. DNA-binding motifs or protein-interacting motifs of transcription factors or histone proteins are usually involved directly in DNA or protein binding and are masked during cross-linking. In these cases, it would be difficult for the antibodies to recognise these masked epitopes.

- **Cell numbers**

Traditionally, a large number of cells (usually 10^7 cells) is required for each ChIP assay. This is the main limiting factor for ChIP experiments performed in primary cells or cells/tissue types where the cell number is limiting (such as stem cells), especially in mammalian systems. Many protocols have been developed to circumvent this issue. Carrier ChIP (CChIP) was developed to perform ChIP in combination with qPCR with as few as 100 cells with the addition of *Drosophila* cells as the carrier agent in native non-crosslinked condition. This was successfully applied in mouse for the study of histone modifications (O'Neill et al., 2006). However, one of the drawbacks of this method is that the carrier agents may interfere with the profile of the native protein-DNA interaction. Also, CChIP cannot be used in formaldehyde cross-linked materials due to the low recovery rate. Other methods have also been developed to solve the cell number issue in cross-linked material. MiniChIP was developed for the study of histone modifications in mouse haematopoietic stem cells and progenitor cells with 50,000 cells by qPCR (Attema et al., 2007). The Q^2ChIP protocol has been demonstrated

to detect histone modifications in as few as 100 cells by qPCR (Dahl and Collas, 2007). MicroChIP has been recently developed with 10,000 cells for the study of RNA Pol II and histone H3 modifications in combination with genome-scale microarrays (Acevedo et al., 2007).

- **ChIP DNA yield**

The amount of DNA recovered after the ChIP experiment is usually ten to a few hundred nanograms (based on experience in the Vetrie laboratory). For highly sensitive applications such as qPCR and ChIP-seq, only nanograms of ChIP DNA are required for analyses. However, for hybridisation onto genome-scale microarray, micrograms of DNA are usually needed. Therefore, an amplification of ChIP DNA is often required in most cases to generate enough starting material for hybridisation. Various amplification protocols have been developed and used in ChIP-on-chip studies. These include the ligation-PCR method where a double-stranded linker is ligated to the end of the DNA fragment for PCR amplification (Ren et al., 2000), the random-priming method where random primers are annealed to the DNA for PCR amplification (Iyer et al., 2001), and the T7-based linear amplification where poly dTs are added to the ends of DNA fragments and polyA dT primers are used for PCR (Bernstein et al., 2005). However, all these methods of PCR amplification may introduce biases for certain sequences or fragment lengths which will affect subsequent analyses on microarrays. Unamplified ChIP DNA has also been successfully used in microarray analyses on the ENCODE tiling arrays to study histone modifications (Koch et al., 2007).

- **Array platform and data analysis**

Depending on the type of analysis that is required, different array platforms can be employed for the downstream analysis of ChIP DNA. These include tiling arrays, promoter arrays, CpG island arrays and whole-genome arrays (Chapter 1, section 1.3.3.3 A). A few parameters should be considered when choosing the appropriate array platforms. These include:

(i) genome coverage of array,

(ii) resolution of array elements,

(iii) density and duplicates of array elements, and

(iv) reproducibility of genomic enrichments.

The analysis of ChIP-on-chip datasets obtained from the microarray is critical for identifying significant protein-bound DNA elements. Similar to expression microarray analysis, normalisation is required as the initial step of data analysis for ChIP-on-chip to account for signal-dependent issues, variation between replicates and scanning conditions. In addition, normalisation with arrays

hybridised with samples generated using IgGs as the antibodies should also be considered to eliminate any non-specific binding by the corresponding IgGs (Pawan Dhami, PhD thesis).

5.1.2 Human transcription factor promoter array platform

The array platform used in the ChIP-on-chip studies described in this Chapter was an in-house transcription factor promoter array. This array contains duplicates of array elements of two main components: the SCL tiling path (Pawan Dhami, PhD thesis) and the promoters of the majority of human transcription factors. These will be discussed in more detail in the following sections. The array was generated using a single-stranded technology developed at the Sanger Institute (Dhami et al., 2005). In this system, single-stranded DNA fragments derived from double-stranded PCR products are immobilised on the surface of the array. During the PCR amplification, primers with a 5'-aminolink modification were used to amplify the sequence from genomic DNA resulting in the generation of double-stranded PCR products containing the modification on one strand only. The double-stranded PCR products are spotted onto the array surface and covalent interactions between the aminolink modification and the array surface occurs. The unmodified strand is then removed by chemical or physical denaturation leaving only the modified single-strand attached to the array surface. This single-stranded array system has a high sensitivity as the resultant single-stranded DNAs cannot reanneal making them effective targets for hybridisation with the labelled samples (Figure 5.2). It has been shown that this array system generates a higher signal:noise ratios than double-stranded PCR product arrays (Dhami et al., 2005).

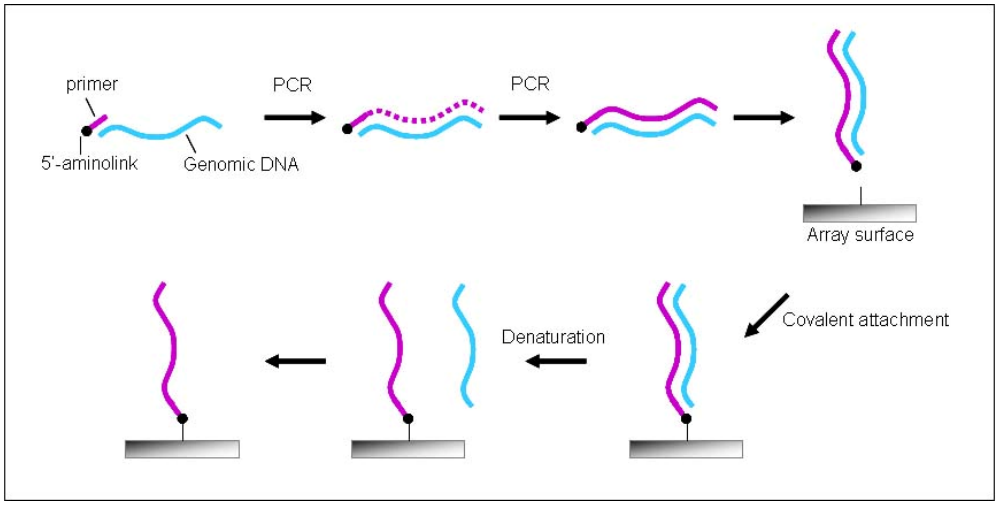


Figure 5.2. The single-stranded array platform. Schematic diagram showed the generation of arrays with the single-stranded array platform. Double-stranded PCR products are generated with a 5'-aminolink primer. 5'aminolink modified strands (purple strands) are attached to the array surface by covalent interaction while the unmodified strands (blue strands) are denatured. Please see text for detailed description. Figure was modified from Dhami et al. 2005 with permission.

- **The SCL tiling path**

GATA1, SCL and LDB1 were shown to bind to the +51 enhancer of SCL (Pawan Dhami, PhD thesis) which is equivalent to the previously described +40 enhancer in mouse (Delabesse et al., 2005) (Chapter 1, section 1.4.2.1 E). As a positive control for selection of ChIP-working antibodies and quality control of the ChIP-on-chip experiment, an SCL tiling path was included on the transcription factor promoter array (see next section). The SCL tiling path was generated by Dr. Pawan Dhami (Pawan Dhami, PhD thesis) which spans approximately 256 kb across the human SCL locus at a resolution of 400 bp. It includes two genes upstream of SCL (KCY and SIL) and three genes downstream (CYP4Z1, CYP4A22 and MAP17) (Figure 5.3 A). Using antibodies against GATA1, SCL and LDB1, significant enrichments were observed by ChIP-on-chip in a novel regulatory region designated as the +51 region (Figure 5.3 B) (Pawan Dhami, PhD thesis). The DNA sequence of this +51 region has hallmarks of the recognition sequence of the SCL erythroid complex originally identified by Wadman et al. (1997) where the E-box and GATA motifs were separated by 9 nucleotides. Therefore, the other members of the SCL erythroid complex may bind to the +51 enhancer.

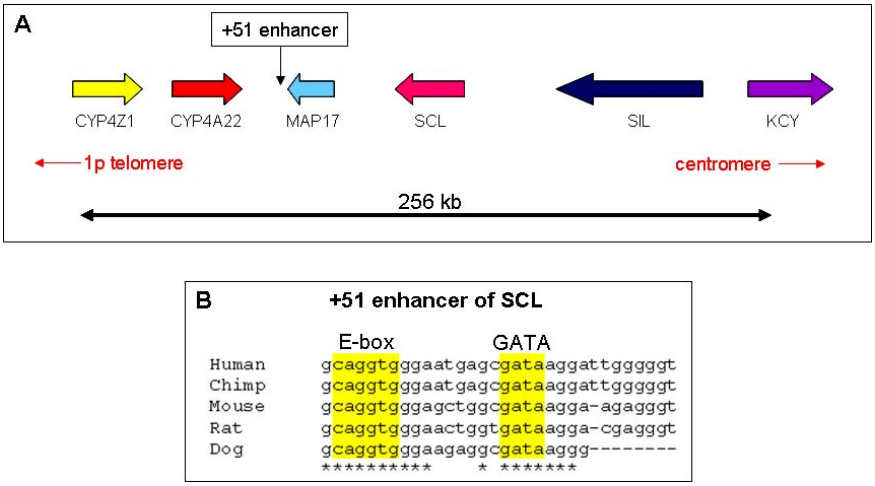


Figure 5.3. The SCL tiling path and the +51 enhancer of SCL. Panel A: schematic diagram showing the genomic region of the SCL locus included on the SCL tiling path array. The black two-way arrow shows the 256 kb region included in the array. The thick coloured arrows represent the genes. The red arrows show the orientation of the locus. The small black arrow shows the position of the +51 enhancer of SCL. Panel B: multiple sequence alignment of the +51 enhancer of SCL. Nucleotides shaded in yellow show the conserved E-box and GATA motifs. Asterisks at the bottom showed the conserved nucleotides across 5 species.

- **The transcription factor promoter array**

As it was not possible to study the entire human genome by ChIP-on-chip when this project was initiated, a sub-set of genomic sequences were studied by ChIP-on-chip. Given that transcription

factors are the key regulators of transcriptional cascades, the focus of the ChIP-on-chip studies for this project was based on the use of an in-house transcription factor promoter array. This array contains approximately 1600 promoters of human transcription factors as well as promoters of a selected handful of haematopoietic genes known to be targets of members of SCL erythroid complex (for example, EPOR, which is a target of GATA1). Gene list for transcription factors was 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). The haematopoietic gene EPOR was included on the array as a positive control for ChIP as GATA1 was shown to bind to the EPOR promoter (Zon et al., 1991). To generate this array, the locations of promoters were first determined using the *in silico* promoter prediction algorithm FirstEF. FirstEF is a software which identifies CpG islands, promoter regions and first exon splice-donor sites in the genome with high accuracy and low false-positive rate (Davuluri et al., 2001). Using FirstEF, predicted promoters of both human and mouse transcription factor genes were shown to be closely clustered within 1 kb around known transcription start sites (Figure 5.4). Therefore; a 1 kb region around the TSS was selected for each transcription factor gene. 1 kb regions for each transcription factor promoter were PCR amplified and included on the array.

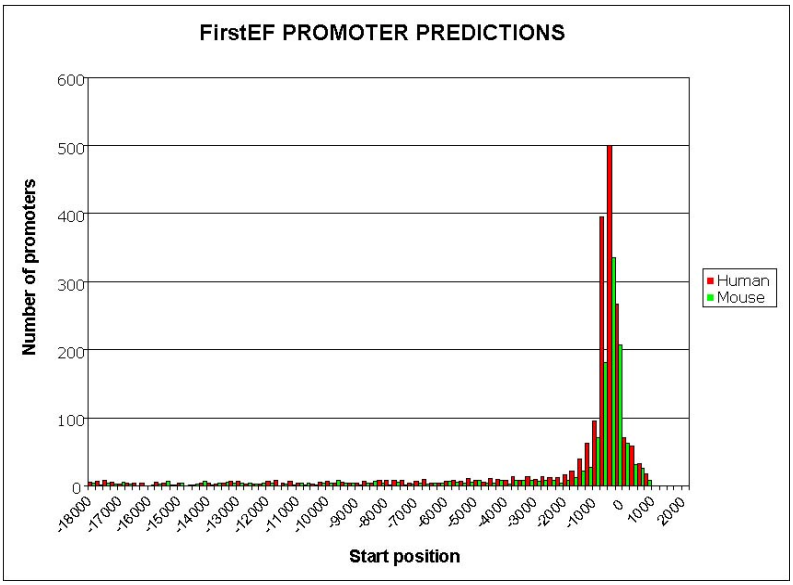


Figure 5.4. FirstEF prediction of human and mouse transcription factor promoters. The histogram shows the number of promoters for transcription factors predicted at various positions relative to the transcription start site. X-axis: start position in bps; y-axis: number of promoters; red bars: human promoters; green bars: mouse promoters. Start position 0 indicates the transcription start site while positive values indicate sequences downstream of the TSS and negative values indicate sequences upstream of the TSS. FirstEF analysis was performed by Dr. Robert Andrews (Wellcome Trust Sanger Institute).

5.1.3 ChIP studies of transcription factors in the SCL erythroid complex in the literature

A number of studies have been performed by previous researchers to identify gene targets bound by members of the SCL erythroid complex. ChIPs in combination with PCR, qPCR or microarray were used to identify and characterise the direct binding of the transcription factors to the promoters or enhancers of their target genes (Table 5.1). In particular, ChIP in combination with a human promoter array was used to identify 71 promoters showing significant binding of SCL in the human T-ALL Jurkat cells (Palomero et al., 2006). ChIP-qPCR of E2A demonstrated the association of E2A with approximately 60% of the SCL target genes in this study. A tiling array across 130 kb of the mouse α-globin locus was used to map GATA1 binding regions at various stages of haematopoietic development in mouse and to study the recruitment of interacting partners using ChIP (Anguita et al., 2004). ChIP-on-chip analysis was used to map GATA1 binding sites in the human β-globin locus in K562 cells identified both known and novel binding regions (Horak et al., 2002). However, a thorough study of the five members of the SCL erythroid complex using ChIP-on-chip in erythroid cells is lacking.

Transcription factor studied	Target gene	Technique used	Organism	Cell type	References
SCL	GYPA promoter	ChIP-PCR	Human	Haematopoietic cell line (TF1)	(Lahlil et al., 2004)
GATA1, SCL and LDB1	P4.2 promoter	ChIP-PCR	Mouse	Erythroid cell line (MEL)	(Xu et al., 2003)
SCL and E2A	c-kit promoter	ChIP-PCR	Human	Haematopoietic cell line (TF1)	(Lecuyer et al., 2002)
SCL, GATA1 and LMO2	β-globin locus control region	ChIP-qPCR	Human	Erythroid progenitor cell line (K562)	(Song et al., 2007)
SCL	71 human genes	ChIP + promoter array	Human	T-ALL cell line (Jurkat)	(Palomero et al., 2006)
GATA1	GFI1B promoter	ChIP-PCR	Human	Erythroid progenitor cell line (K562)	(Huang et al., 2004)
GATA1	HS2 region of the β-globin locus	ChIP-PCR	Mouse	Erythroid cell line (MEL)	(Johnson et al., 2002)
GATA1	MYC promoter	ChIP-PCR	Mouse	GATA1-null erythroblast cell line (G1E-ER4)	(Rylski et al., 2003)
GATA1	FOG-1 enhancer	ChIP-qPCR	Mouse	GATA1-null erythroblast cell line (G1E-ER4)	(Welch et al., 2004)
GATA1	α-globin locus	ChIP + tiling array	Mouse	Erythroid cell line (MEL)	(Anguita et al., 2004)
GATA1	β-globin locus	ChIP + tiling array	Human	Erythroid progenitor cell line (K562)	(Horak et al., 2002)

Table 5.1. ChIP studies of various members of the SCL erythroid complex. The target gene, technique, organism and cell type used in the ChIP studies are listed in the table.

5.2 Aims of this chapter

The aims of work presented in this Chapter were:

1. To test and validate antibodies targeting five members of the SCL erythroid complex for ChIP-on-chip applications.
2. To identify putative promoters bound by each member of the SCL erythroid complex in K562 cells by ChIP in combination with the transcription factor promoter array.
3. To investigate the transcription factor binding sites (TFBS) in the putative promoters and perform comparative genomic sequence analyses of these TFBSs.
4. To validate the putative target genes by ChIP-qPCR in K562 and HEL erythroid cell lines.

5.3 Overall strategy

The overall aim of the work described in this Chapter was to confirm and identify direct target genes regulated by each of five members (GATA1, SCL, E2A, LMO2 and LDB1) of the SCL erythroid complex using ChIP-on-chip. Working ChIP assays for each transcription factor were validated in ChIP-on-chip in K562 cells using the SCL tiling path/transcription factor promoter array. As mentioned in section 5.1.2, GATA1, SCL and LDB1 were shown to bind to the +51 enhancer (Pawan Dhami, PhD thesis) and antibodies against these transcription factors were previously characterised for ChIP assays by Dr. Pawan Dhami. Since the +51 enhancer contains the consensus E-box and GATA1 motifs separated by 9 nucleotides as first described for the SCL erythroid complex by Wadman et al (1997), the other members of the SCL erythroid complex (E2A and LMO2) may also bind to this enhancer. Based on enrichments obtained for the +51 enhancer and promoters at the SCL locus, the best performing antibodies were chosen as the working ChIP assays for these two transcription factors. Three biological replicates of ChIP-on-chip for each of the five transcription factors and their corresponding IgG controls were performed using the SCL tiling path/transcription factor promoter array. Although it has been shown that dye-specific bias is a source of error in 2-colour array experiments and cannot be removed during normalisation (Dobbin et al, 2005), dye-swap experiments were not performed as this will double the cost required for the experiments. The quality of each ChIP-on-chip assay was assessed at various steps during the experiments (Section 5.4.1). Enrichments of each promoter in the ChIP-on-chip study of the transcription factors were normalised with their enrichments in the corresponding IgG controls. Statistical analyses of enriched promoters were carried out for the ChIP-on-chip experiments which passed quality control. Cross-comparison between the putative targets identified by each of the five members of the SCL erythroid complex was performed to identify targets bound by all five

members of the complex. To identify the DNA sequence motifs which were likely to bind the five transcription factors, sequences in the enriched promoters were analysed by TESS and TFSearch together with comparative genomic sequence analyses. Confirmation of promoter binding events was addressed by ChIP in combination with quantitative PCRs. The overall strategy for this ChIP-on-chip study is summarised in Figure 5.5.

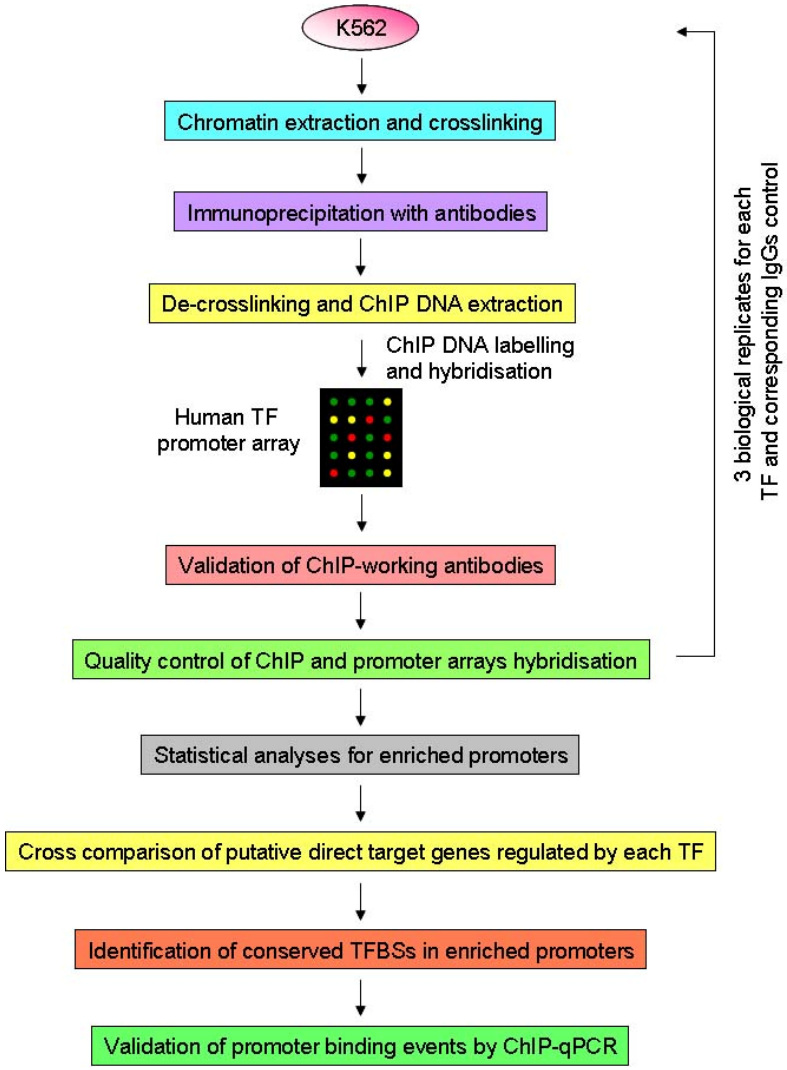


Figure 5.5. Overall strategy for the ChIP-on-chip analyses of the SCL erythroid complex. ChIP assays in K562 were performed as described in Chapter 2. Chromatin from K562 cells was extracted and sonicated while DNA bound by the transcription factor under study was immunoprecipitated by specific antibodies followed by de-crosslinking and extraction. Working ChIP assays for each transcription factor were validated based on enrichments obtained for the +51 enhancer at the SCL locus. Quality control of various steps of the ChIP-on-chip assays was performed. Three biological replicates of ChIP-on-chip for each of the five transcription factors and their corresponding IgG controls were performed. Normalisation of enrichments was done against the IgG controls. Statistical analyses of enriched promoters were carried out for the ChIP-on-chip experiments. Cross-comparison between the enriched promoters identified by

each of the five members of the SCL erythroid complex was performed. Sequences in the enriched promoters were analysed by TESS and TFSearch together with comparative genomic sequence analyses to identify conserved transcription factor binding sites. Confirmation of promoter binding events was addressed by ChIP in combination with quantitative PCRs.

5.4 Results

5.4.1 Quality control of various steps of chromatin-immunoprecipitation

To ensure that experiments done at different times for the various biological replicates were consistent, a variety of steps were analysed throughout the ChIP-on-chip procedure as described below.

5.4.1.1 Culturing of cells

Cell lines (K562 and HEL) were cultured for no more than a week at concentrations of 0.5 million cells per millilitre before chromatin extractions were performed. Fresh media were added one day before extraction. To further reduce the variability across replicates, the same passage of cells was defrosted for biological replicates performed at different times. From the cultured cells, aliquots of cells were taken for flow analysis prior to chromatin extraction. The proportion of actively dividing cells was monitored by flow analysis to determine the DNA content of the cells (Figure 5.6) as a measure of the number of actively dividing cells (actively dividing cells in S or G2/M phases of the cell cycle have higher DNA contents due to DNA replication). Only cell populations with similar growth characteristics were used for subsequent analyses. For example, for all of the experiments performed for K562, approximately 60-70% of cells were actively dividing in all three bioreplicates.

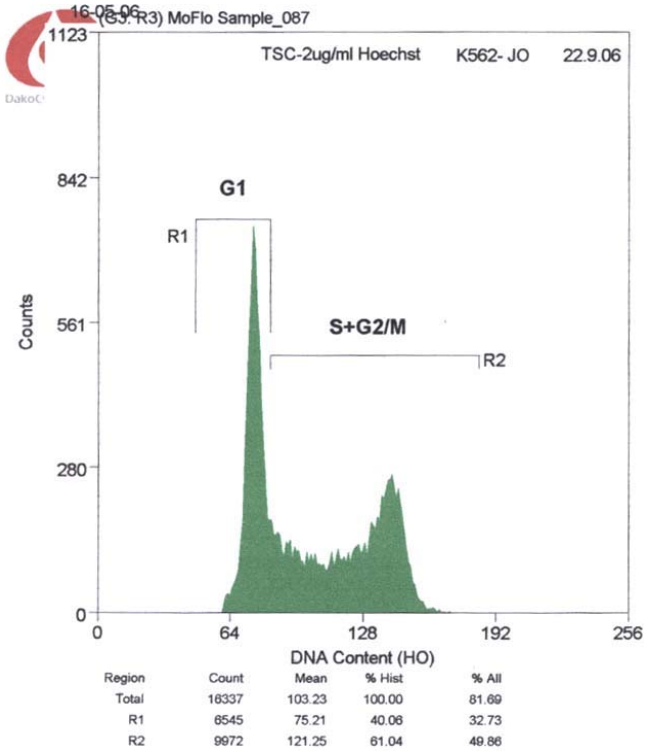


Figure 5.6. Flow analysis of growth pattern of cell lines. K562 or HEL cells used in the ChIP-on-chip experiment described in this Chapter were subjected to flow-analysis by staining with the DNA binding dye Hoechst 33342 to determine the DNA content of each cell. The percentages of cells in the G1 (labelled as R1) and S and G2 or M (labelled as R2) phases of the cell cycle were determined by measuring the fluorescence intensity (shown at the bottom of the image). Actively dividing cells in S or G2/M phases have higher DNA contents due to DNA replication and this could be used as a measure of the proportion of cells in the population which were actively dividing (this study was performed by Bee Ling Ng, Wellcome Trust Sanger Institute).

5.4.1.2 Preparation of cross-linked chromatin

The initial step of ChIP-on-chip is the cross-linking of protein and DNA in the chromatin. The cross-linking condition used here was 1% formaldehyde for 10 minutes (this was based on titration experiments performed by Dr. Pawan Dhami in the laboratory). The resultant protein-DNA complexes were sonicated to shear the DNA into fragments with a size distribution in the range of of 600-3000 bp (average size around 1000 bp). To ensure that the cross-linking and sonication consistently resulted in DNA fragments of the correct size distribution; a small aliquot of the cross-linked and sonicated material was analysed by agarose gel electrophoresis (Figure 5.7). A smear was observed with an average size distribution of approximately 1000-1500 bp. Purified DNA from this crude chromatin extract was subsequently shown to give a size distribution with an average DNA fragment size of approximately 1000 bp (See Figure 5.8).

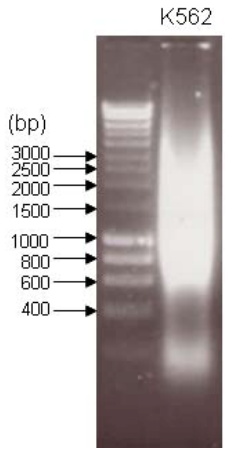


Figure 5.7. Agarose gel electrophoresis of cross-linked and sonicated chromatin. Chromatin extracted from K562 after cross-linking and sonication was analysed by electrophoresis of a 1% agarose gel made with 1 X TBE and visualised by ethidium bromide staining. A 1 kb ladder was loaded in the left lane and 5 µl of the K562 chromatin cross-linked in 1% formaldehyde for 10 minutes is shown in the lane to the right of the size markers.

5.4.1.3 Extraction of ChIP DNA

Similarly, agarose gel electrophoresis was used to examine the size distributions and recoveries of input and ChIP DNAs (Figure 5.8). Input DNA is the material extracted after de-crosslinking of the chromatin which did not undergo any immunoprecipitation. ChIP DNA, in contrast, is the DNA extracted after immunoprecipitation with specific antibodies. On agarose gels, input DNA normally showed a visible DNA smear of similar size distribution to the crude chromatin (Figure 5.7). ChIP DNAs, in contrast, were difficult to visualise on agarose gels because of the amount of material recovered from ChIP assays. Thus, relative amounts of DNA recovered in ChIP samples were monitored by comparing the intensity of the yeast tRNA which was co-precipitated in the ChIP samples.

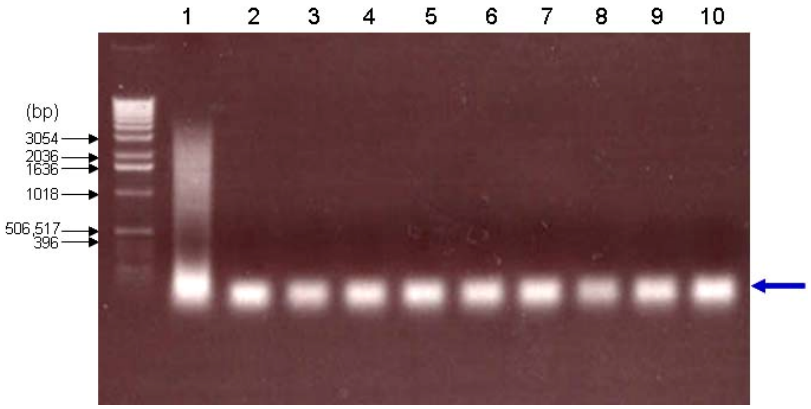


Figure 5.8. Agarose gel electrophoresis of input and ChIP DNA. 5 µl of input and ChIP DNAs using antibodies for transcription factor and IgGs were extracted and electrophoresed on a 1% agarose gel in 1 X TBE and visualised by ethidium bromide staining. A 1 kb ladder was loaded in the lane on the left of the image. Lane 1: input DNA; lane 2: ChIP DNA of LMO2 G16 antibody; lane 3: ChIP DNA of LMO2 N16 antibody; lane 4: ChIP DNA of LMO2 Abcam antibody; lane 5: ChIP DNA of LDB1 N18 antibody; lane 6: ChIP DNA of SCL serum; lane 7: ChIP DNA of E12 H208 antibody; lane 8: ChIP DNA of goat IgG; lane 9: ChIP DNA of mouse IgG; lane 10: ChIP DNA of rabbit IgG. The

input DNA in lane 1 shows a smear of the appropriate size distribution while only yeast tRNA was observed in the ChIP DNA samples at the bottom of the gel (shown by the blue arrow).

5.4.1.4 Labelling of input and ChIP DNA

Input and ChIP DNAs were labelled with cyanine dyes (Cy3 and Cy5) for array hybridisations. The DNA labelling process was performed by random priming with Klenow fragments lacking the 3' to 5' and 5' to 3' exonuclease activity (Lieu et al., 2005). Due to the intrinsic strand displacement activity of Klenow, the labelled fragments were exponentially amplified (Walker, 1993) (Figure 5.9).

The labelled input and ChIP DNAs were analysed by agarose electrophoresis to evaluate the labelling and amplification efficiency (Figure 5.10). In both input DNA and ChIP DNA, smears were observed across a broad size range, with the majority of the labelled fragments being less than 200 bp in size. Compared with the original unlabelled ChIPs DNA (Figure 5.8) where $1/10^{th}$ the original material was loaded onto agarose gels, more obvious smears were observed after labelling (1/30 of the labelled material) indicating large quantities of DNA were amplified during labelling.

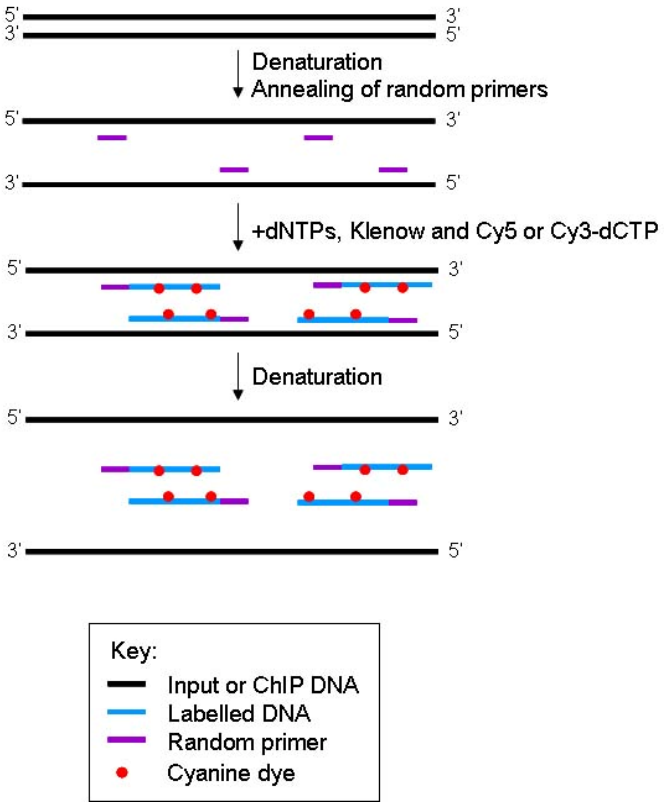


Figure 5.9. Random priming and cyanine labelling of input and ChIP DNA. Input and ChIP DNAs were labelled with a random priming method involving the use of the Klenow fragment with a strand displacement activity. The DNA being labelled was first denatured and primers were annealed. The resulted DNA were amplified with Klenow enzyme and labelled with cyanine dyes.

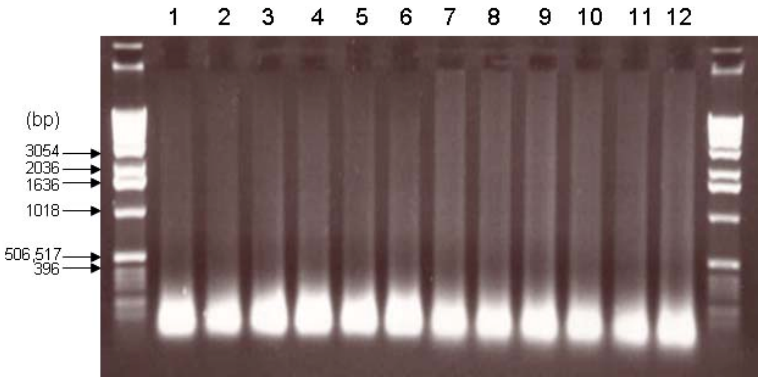


Figure 5.10. Agarose gel electrophoresis of labelled input and ChIP DNA. 5 µl of input and ChIP DNA samples from K562 were labelled by the random priming method and electrophoresed on a 1% agarose gel made with 1 X TBE and visualised by ethidum bromide staining. 1 kb ladder is loaded in the first and last lane. Lanes 1-6: Input samples; lane 7: labelled ChIP DNA of rabbit IgG; lane 8: labelled ChIP DNA of goat IgG; lane 9: labelled ChIP DNA of LMO2 N16 antibody; lane 10: labelled ChIP DNA of LDB1 N18 antibody; lane 11: labelled ChIP DNA of E47 N649 antibody; lane 12: labelled ChIP DNA of E12 H208 antibody.

5.4.1.5 Hybridisation and analyses of the transcription factor promoter array

After hybridisation and scanning, the resultant array images were quality-controlled. Initially they were assessed by eye to identify any visible problems with the array hybridisation which may affect the quantification of spots (high background and various hybridisation artifacts). Array which showed such problems were discarded and the hybridisations were repeated. Given that the array elements were spotted in duplicate, the coefficients of variation (CVs) for the duplicated elements were calculated for each spot to determine reproducibility of datapoints within a single hybridisation. Typically, the median CV (median of all CVs obtained from duplicate array elements) for a given hybridisation was approximately 5% and arrays which deviated substantially from this median value were not included in further analyses (a CV of 6% was used as a cut-off). An example of good quality hybridisation is shown in Figure 5.11.

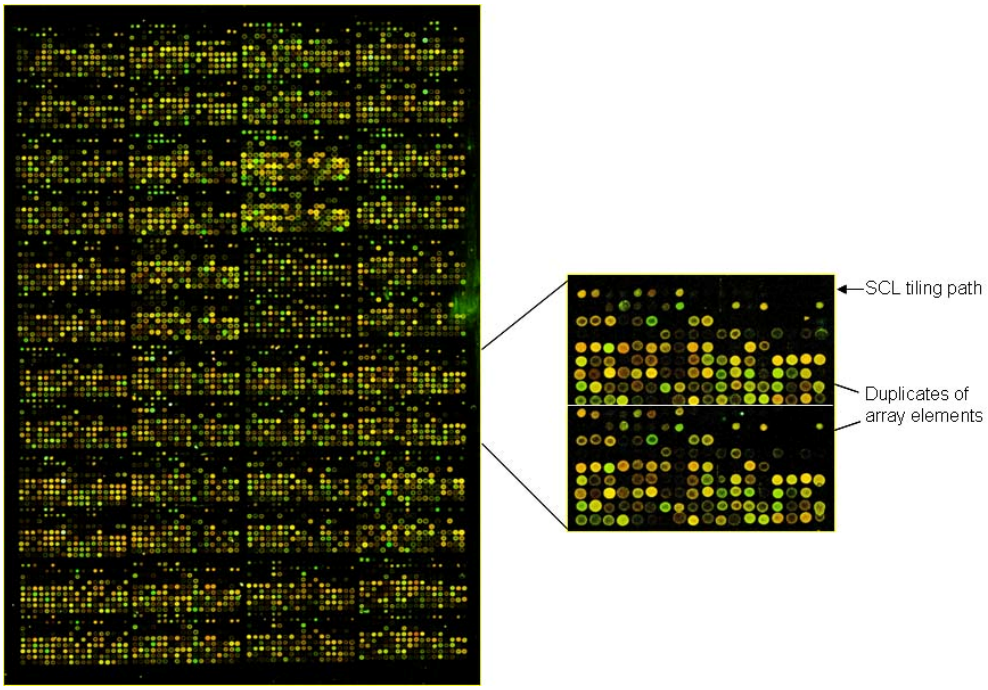


Figure 5.11. A composite image of the human transcription factor promoter array. The promoter array was hybridised with (i) ChIP DNA derived from ChIP with the GATA1 M20 antibody in K562 cells and (ii) the input DNA of K562 cells. The array contains 24 sub-arrays and 4132 spots where each spot represents an array element for either a human transcription factor promoter or a tile of the SCL tiling path. The zoomed-in image on the right illustrates one of the sub-arrays containing the spots for the SCL tiling path and the duplicates of each array element. Green spots represent array elements enriched in the ChIP sample. Red spots represent array elements under-represented in the ChIP sample. Yellow spots represented array elements equally represented in the ChIP and input samples. White spots showed array elements with saturated pixel values in the image for the ChIP sample.

5.4.2 Evaluation of working antibodies by positive control elements of the array

Three criteria were used for the selection of high quality antibodies for use in ChIP-on-chip assays for the five transcription factors of the SCL erythroid complex:

(i) they must show significant enrichments at the +51 enhancer of the SCL locus. The promoter array contained the SCL tiling path (section 5.1.2) which acted as the positive control region for testing the antibodies against the 5 members of the SCL erythroid complex. Each member was expected to bind the +51 enhancer of SCL which contains the consensus E-box/GATA motif and had been shown to bind GATA1, SCL and LDB1 (Pawan Dhami, PhD thesis). This +51 is the equivalent to the +40 enhancer of SCL in mouse (Ogilvy et al., 2007).

(ii) the background in the negative regions must be low. As previously demonstrated by the ChIP-on-chip data of GATA1, SCL and LDB1 on the SCL tiling array, many regions on the locus show

enrichments at or near a value of 1 (baseline). These regions are regarded as the negative regions for assessment of non-specific binding.

(iii) they must be specific as detected by western blotting as described in Chapter 3.

Appendix 3B summarised the characteristics of, and the results obtained for, a variety of antibodies tested in the ChIP-on-chip experiment. Six of these antibodies were used for further ChIP-on-chip experiments and the results of these across the SCL tile path are described below.

(i) GATA1: 15- to 30-fold enrichments for the +51 SCL enhancer were observed for the GATA1 M20 ChIP assay (Figure 5.12 A), replicating results obtained previously with this assay (Pawan Dhami, PhD thesis). Significant enrichments of up to 5-fold were also observed for SCL promoter 1a and +3 and -9/-10 enhancers (Pawan Dhami, PhD thesis). This antibody was also shown to be highly specific on western blotting (Chapter 3, section 3.4.2).

(ii) SCL: Both antibodies tested for SCL showed substantial enrichments at the +51 enhancer where 8- to 12-fold enrichments were observed for the TAL1 Active Motif antibody (Figure 5.12 B) and 20-fold enrichments were shown for the SCL serum. Although the SCL serum showed higher enrichment in the +51 enhancer, the quantity of this antibody was limiting (as it was a gift from a collaborator) and it did not yield the appropriate bands for the SCL protein in western analysis (Chapter 3, section 3.4.2). In contrast, the TAL1 Active Motif antibody was shown to be specific for SCL in western analysis and was therefore used in subsequent ChIP-on-chip analyses.

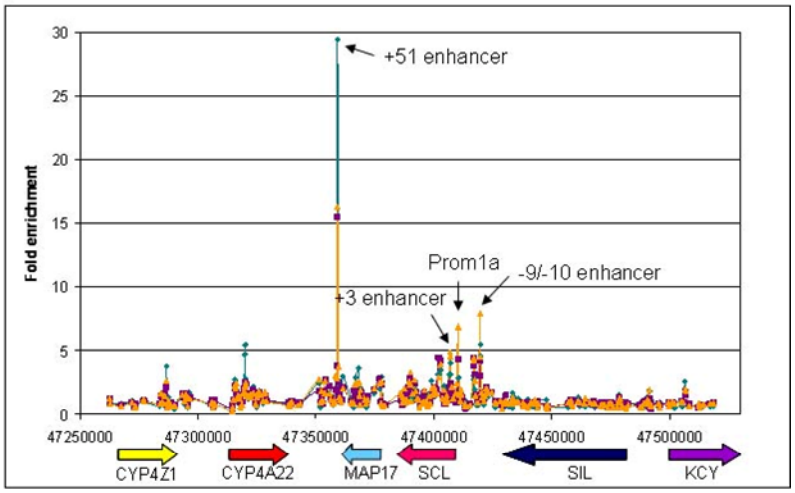
(iii) E2A: Two antibodies were tested for E2A which recognised both E12 and E47 isoforms. The TCF3 antibody from Abcam showed no substantial enrichments in any of the SCL enhancers or promoters while the E2A antibody from BD Biosciences showed an approximately 12-fold enrichment in the +51 region. However, the E2A antibody from BD Bioscience could not identify specific bands for E2A in western analysis (Chapter 3, section 3.4.2). Specific antibodies for the E12 and E47 isoforms were also characterised. The E12 H208 and E47 N649 antibodies both showed up to 60-fold enrichments in the +51 enhancer and enrichments of approximately 8-fold in the +3 and -9/-10 enhancers and the promoter 1a (Figure 5.12 C and D). These two antibodies were also shown to be specific for E12 and E47 in western analysis (Chapter 3, section 3.4.2) and were used for further ChIP-on-chip analyses. However, no information is known about the cross-reactivity of these two antibodies with the other isoform.

(iv) LDB1: Up to 45-fold enrichments for the +51 enhancer were observed for the one antibody tested for LDB1 (Figure 5.12 E). In addition, substantial enrichments of more than 10-fold were also observed for promoter 1a and -9/-10 enhancers and 5-fold enrichments were shown for the +3

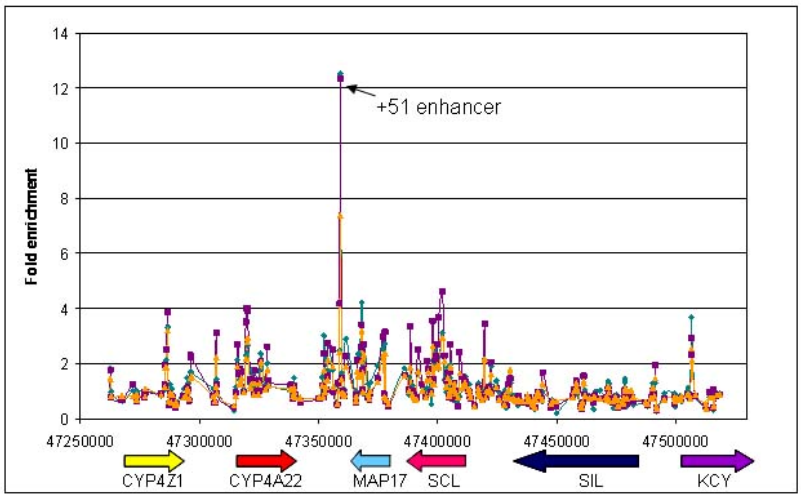
enhancer. This antibody was also shown to be highly specific on western blotting (Chapter 3, section 3.4.2). Therefore, this antibody was used for further ChIP-on-chip analyses.

(v) LMO2: Both antibodies tested for LMO2 did not show high enrichments across the SCL locus and generated a lot non-specific noise (Figure 5.12 F). However, the N16 antibody was slightly better than G16 in terms of the enrichments at the +51 enhancer i.e. up to 10-fold for LMO2 N16 versus 7-fold for LMO2 G16. LMO2 N16 was therefore used in subsequent ChIP-on-chip analyses despite there being no western data to support its specificity.

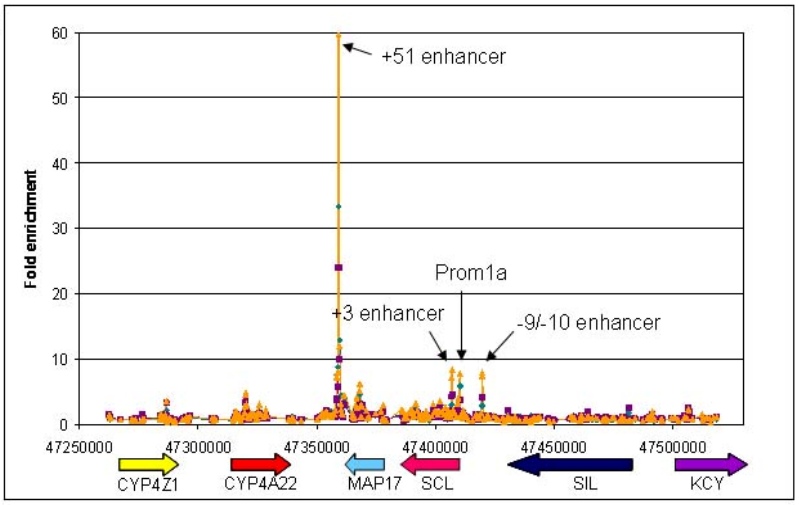
A) GATA1 M20



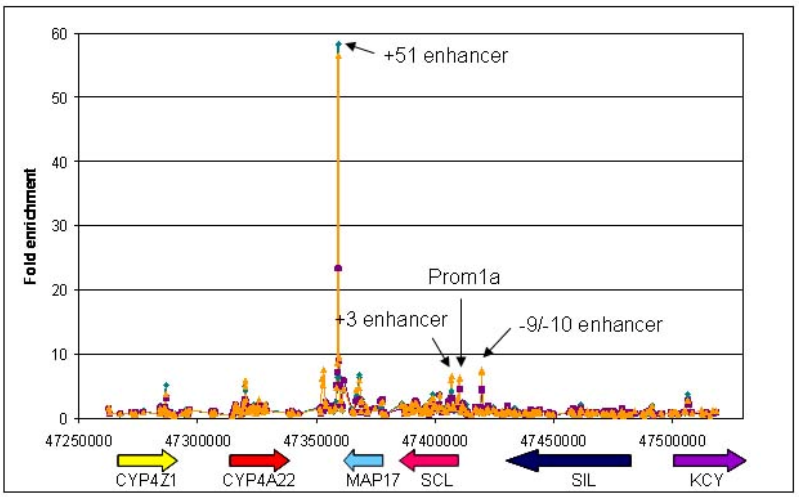
B) TAL1 Active Motif



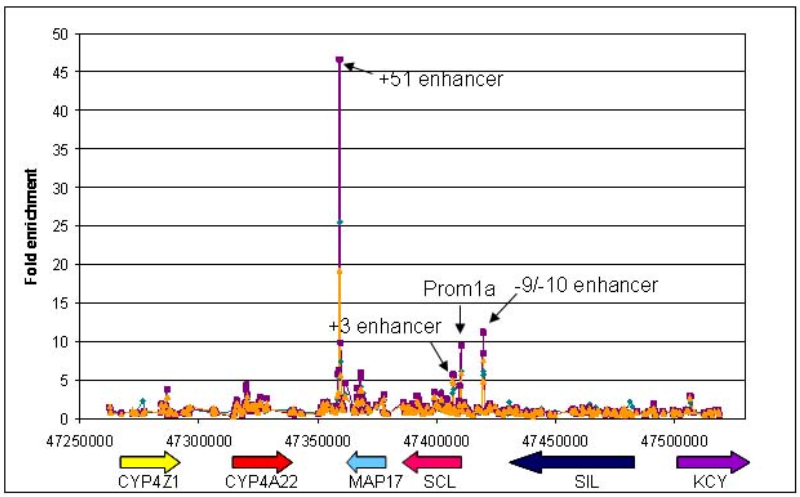
C) E12 H208



D) E47 N649



E) LDB1 N18



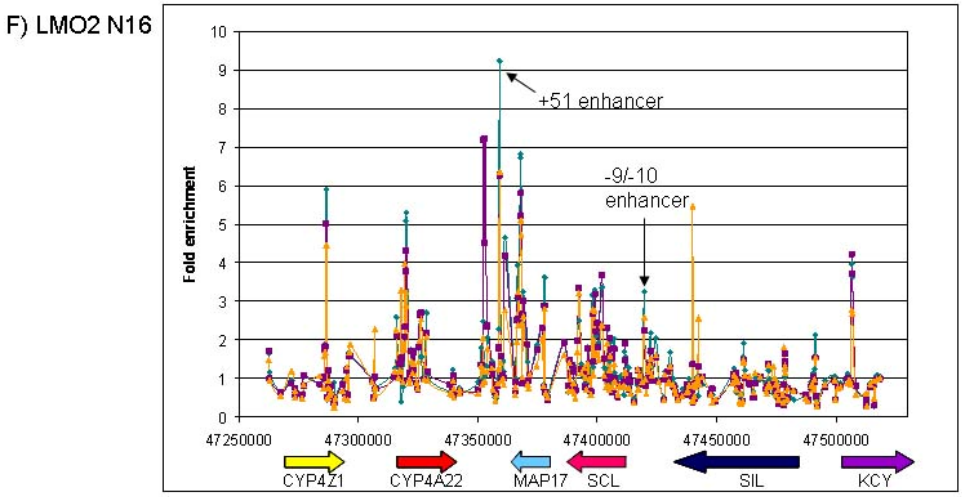


Figure 5.12. ChIP-on-chip profiles of selected working antibodies for the SCL erythroid complex across the SCL locus in K562 cells. The ChIP-on-chip profiles across the SCL locus of antibodies selected for subsequent analyses were shown. Panel A: SCL locus profile of GATA1 M20 antibody; panel B: SCL locus profile of TAL1 Active Motif antibody; panel C: SCL locus profile of E12 H208 antibody; panel D: SCL locus profile of E47 N649 antibody; panel E: SCL locus profile of LDB1 N18 antibody; panel F: SCL locus profile of LMO2 N16 antibody. The x-axis represented the genomic coordinates across the SCL tiling path and the y-axis represented the fold enrichments. The thick coloured arrows showed the position of the genes included on the SCL tiling path. Light blue curve: biological replicate 1; violet curve: biological replicate 2 and orange curve: biological replicate 3. SCL enhancers or promoters which showed enrichments were labelled by black arrows on the graph.

5.4.3 Data analyses of enriched promoters

Having validated the performance of antibodies for each of the five transcription factors in ChIP-on-chip, three bioreplicates for each of the chosen assays were performed on the SCL tiling path/transcription factor promoter array. Two technical replicates were also performed for each biological replicate. Similarly, the host IgG control ChIP-on-chip experiments were performed for each transcription factor assay across three bioreplicates and two technical replicates. The quality of each ChIP and hybridisation was monitored as described in section 5.4.1 and the array hybridisations that passed the quality control criteria were subject to statistical analyses for the selection of enriched promoters which are likely to be bound by the transcription factors under study.

5.4.3.1 Overall strategy of statistical analyses

Figure 5.13 outlines the procedures used for statistical analyses of enriched promoters in the ChIP-on-chip experiment. Signal intensities of the array elements for all the scanned array hybridisations were first quantitated in Scanarray Express. Ratios of Cy3 (ChIP sample) against Cy5 (Input sample) were also generated in Scanarray Express. The ratios for the duplicated array elements in a

given hybridisation were averaged. Ratios for the two technical replicates were averaged to provide a mean ratio for each bioreplicate. The ratio data was transformed by normalisation, at various stages, in three ways:

(i) signal intensities for both channels in each hybridisation were scaled by total intensity in Scanarray Express,

(ii) each ratio measurement for every array element in a given hybridisation was normalised to the median ratio of all measurements.

(iii) the ratios for all array elements in each experiment (either in each bioreplicate or as the mean of bioreplicates) were normalised against the ratios obtained for the host IgG controls. This normalisation procedure would help account for non-specific enrichments from the host IgGs and effectively remove them from the datasets.

Two methods were used to carry out the statistical analyses of the enriched promoters (Figure 5.13). In method A, each biological replicate was treated separately with respect to the generation of mean enrichments and normalisation between the transcription factor ChIP-on-chip assays and the host IgG ChIP-on-chip assays. Enriched promoter array elements which were two standard deviations above the mean were chosen as the putative target promoters. Two standard deviations were used as a cut-off as it represented a 95.45% confidence level – in other words, the promoters identified were statistically significant in terms of enrichment levels away from background. The promoter lists from each of the three biological replicates were compared in a Venn diagram and promoters found to be significantly enriched in all three bioreplicates were chosen as the putative target promoters of the transcription factor under study. In method B, the average ratio of each promoter was obtained from the 3 biological replicates for the transcription factor ChIP-on-chip assay and normalised with the corresponding average ratio of each promoter from the 3 bioreplicates for the host IgG ChIP-on-chip experiments. Promoters which were enriched 2 standard deviations above the mean were chosen as the putative target genes.

Comparatively speaking, method A was a more stringent approach for selecting promoters which are likely to be bound by the transcription factor. Only promoters which were statistically significant in all three biological replicates were chosen as putative target genes. This requires that the transcription factor-promoter binding is strong and significant to show enrichment in each ChIP-on-chip experiment. Method B, however, is less stringent but it was possible to detect binding events which showed a degree of variability in enrichment across the three bioreplicates.

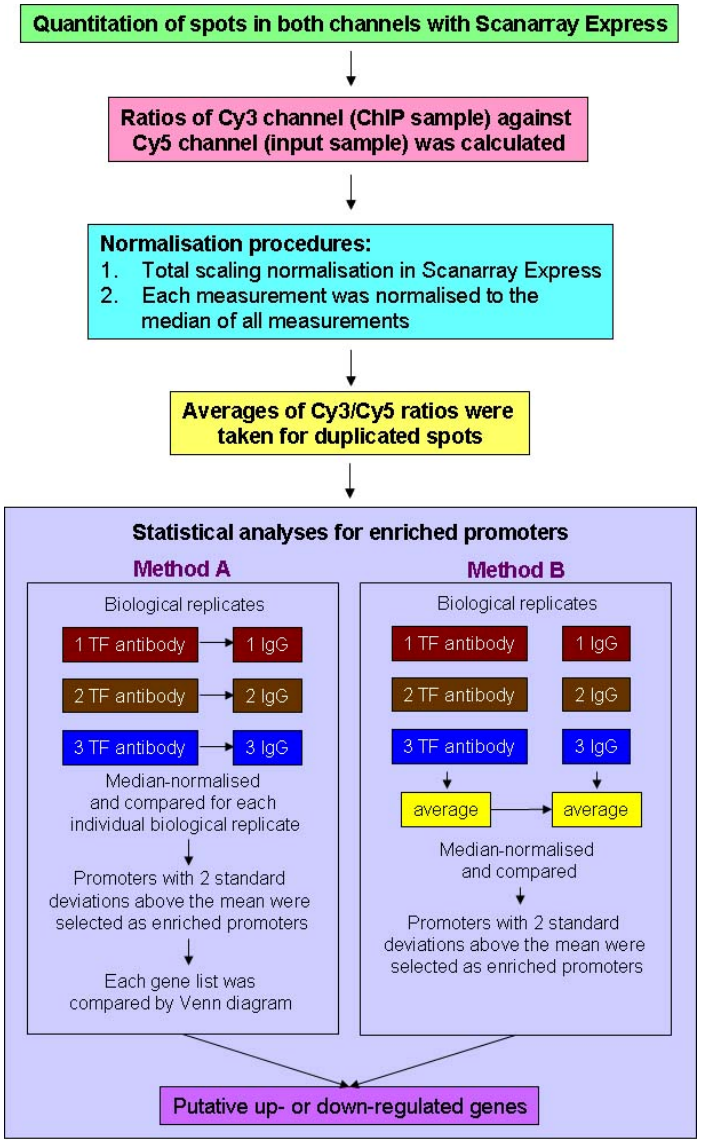


Figure 5.13. Flow diagram of statistical analyses of enriched promoters in ChIP-on-chip. Signal intensities of the array elements were first quantitated in Scanarray Express. Ratios of Cy3 (ChIP sample) against Cy5 (Input sample) were also generated and ratios for the duplicated elements were averaged. Ratios for the two technical replicates were also averaged to provide a mean ratio for each bioreplicate. The ratio data was transformed by normalisation at various levels as described in the text. Statistically significant enriched promoters were identified for each of methods A or B.

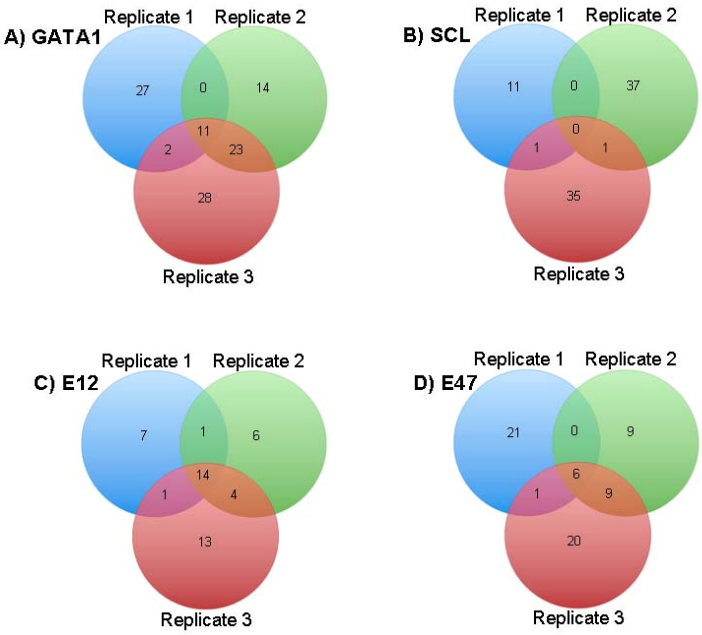
5.4.3.2 Data analyses for the selection of putative target genes

Using the two strategies outlined above, a number of promoters were selected as putative regulatory target genes for each of the transcription factors under study.

Figure 5.14 shows the results for each transcription factor ChIP-on-chip analysis using method A. In ChIP-on-chip analysis for GATA1, E12, E47 and LDB1, between 6 and 14 promoters were

identified in all the three biological replicates. In particular, the known direct target genes (EPOR and EKLF), were found to be enriched in the ChIP-on-chip study of GATA1. Overall, the percentages of promoters being significantly enriched in all bioreplicates for each of these four transcription factor ChIP-on-chip assays was approximately 1% of the total number of promoters on the array. However, no promoters were found to be consistently enriched in all three of the biological replicates for SCL and LMO2. Both of these ChIP-on-chip assays were consistently the worst performing (in terms of enrichments) of all of the assays used.

A larger set of significantly enriched promoters were identified using method B. The number of promoters identified by the transcription factor ChIP-on-chip assays in method B ranged from 15 to 41. Unlike the results obtained for method A, a number of promoters were found to be significantly enriched for SCL and LMO2 using this method. Promoter targets identified by methods A and B were also compared in Venn diagrams. All the promoters identified by method A for a given transcription factor were also identified by method B for the same transcription factor (Figure 5.15). In total, using the two different statistical methods of analyses described here, over 100 promoters of putative target genes were found to be enriched in ChIP-on-chip analysis.



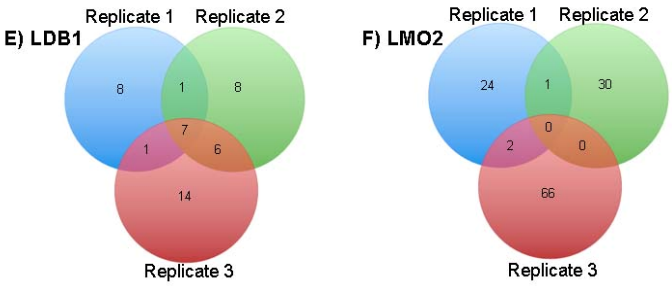


Figure 5.14. Venn diagram comparison of putative target promoters identified in all three bioreplicates for each transcription factor ChIP-on-chip assay using statistical method A. Numbers shown in the Venn diagrams were numbers of promoters identified in each biological replicate of the ChIP-on-chip studies. Panel A: Venn diagram of GATA1 ChIP-on-chip study; panel B: Venn diagram of SCL ChIP-on-chip study; panel C: Venn diagram of E12 ChIP-on-chip study; panel D: Venn diagram of E47 ChIP-on-chip study; panel E: Venn diagram of LDB1 ChIP-on-chip study; panel F: Venn diagram of LMO2 ChIP-on-chip study.

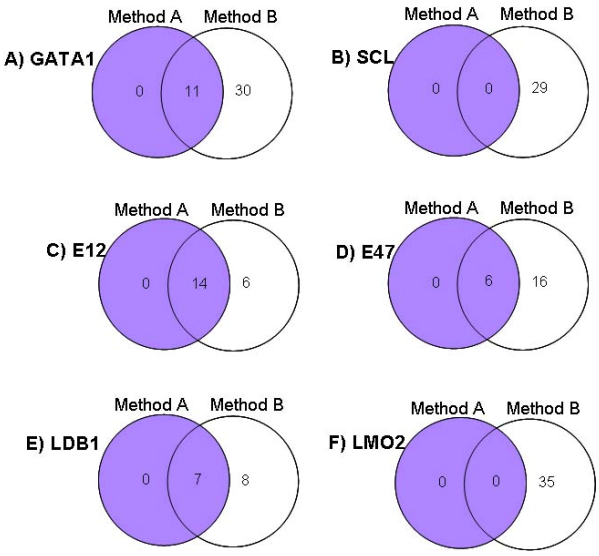


Figure 5.15. Venn diagram comparison of putative target promoters identified in ChIP-on-chip studies for each of the five transcription factors of SCL erythroid complex using statistical methods A and B. Numbers shown in the Venn diagrams were numbers of promoters identified in each biological replicate of the ChIP-on-chip studies. Panel A: Venn diagram of GATA1 ChIP-on-chip study; panel B: Venn diagram of SCL ChIP-on-chip study; panel C: Venn diagram of E12 ChIP-on-chip study; panel D: Venn diagram of E47 ChIP-on-chip study; panel E: Venn diagram of LDB1 ChIP-on-chip study; panel F: Venn diagram of LMO2 ChIP-on-chip study.

5.4.3.3 Classification and literature review of putative target genes

The transcription factor promoters which were identified as being significantly enriched in the ChIP-on-chip analysis described above were considered to represent regulatory interactions of putative direct target genes of members of the SCL erythroid complex. The putative target genes for

GATA1, E12, E47 or LDB1 selected by method A are summarised in Table 5.2. Some of the targets were enriched by more than one transcription factor using the method A criteria (for example, EPOR was identified with GATA1, E12, E47 and LDB1 and the SCL +51 enhancer was identified with all six transcription factor assays). The promoters identified in method B for the 6 transcription factor ChIP-on-chip assays were also cross-compared with the enriched promoters obtained by method A. Indeed, some of these target genes identified by method A for one transcription factor were also identified by method B for another transcription factor. This is also shown in Table 5.2. Taken all together, this data provided further evidence that at least some of the target gene promoters may be bound by the whole erythroid complex or variations thereof.

The vast majority of the targets identified were transcription factors, with the exception of EPOR (the known target of GATA1 in the erythroid lineage). To further understand the nature of the putative targets of members of the SCL erythroid complex, information was obtained from public databases including iHOP (http://www.ihop-net.org/UniPub/iHOP/), OMIM (http://www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM) and Gene Expression Atlas (http://expression.gnf.org/cgi-bin/index.cgi) and from performing literature searches. This information is summarised in Table 5.2 and Figure 5.16. The ChIP-on-chip studies were able to identify additional putative targets with known function in the haematopoietic compartment. Based on the method A analysis, ten of the target genes were known to be expressed in the lymphoid lineage while five others (including SCL) were found to be expressed in early blood progenitors found in the bone marrow. Six of the genes (including SCL) were shown to be involved in haematopoietic development. Furthermore, eight of the target gene encoded proteins involved in chromatin remodelling/chromatin modifications. The putative promoters identified for the 6 transcription factor ChIP-on-chip assays using statistical methods A and method B were classified by function as shown in Figure 5.16.

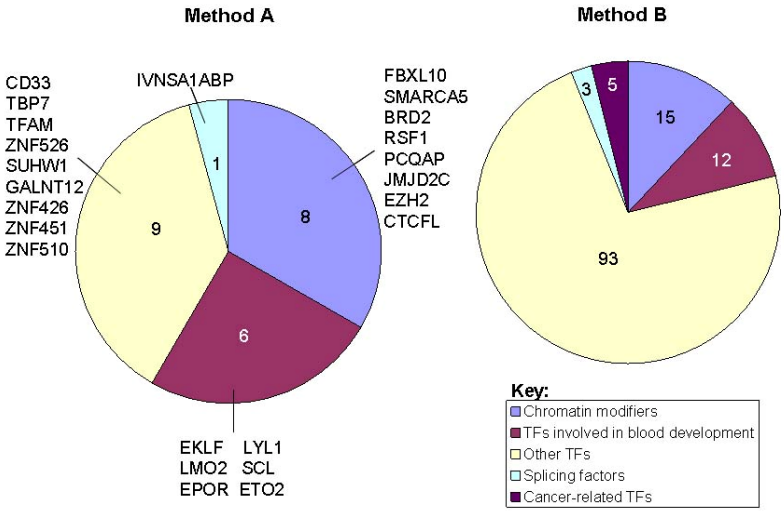


Figure 5.16. Classification of putative target genes of members of the SCL erythroid complex based on ChIP-on-chip studies. Pie charts show the classification of transcription factors identified by one or more transcription factors in the SCL erythroid complex using statistical method A (left panel) or method B (right panel). Numbers indicated in the pie charts show number of target genes in each category. The gene symbols shown in the method A pie chart are further summarised in Table 5.2. Each functional category is depicted by the colour code shown in the key.

Name of putative target gene	TF regulating target (method A)	TF regulating target (method B)	Expression pattern	Functions of putative target gene
FBXL10*	GATA1		Low expression in thymus and $CD4^+$ T-cells	JmjC domain-containing histone demethylation protein. Involved in chromatin modification and recruitment to chromatin. Cooperate with MBD1 to regulate transcription at methylated CpG sequences.
EKLF*	GATA1, E12, E47, LDB1	SCL	Bone marrow	Regulator of erythropoiesis. Transcriptional activator of β-globin expression.
ZNF526	GATA1			
TFAM	GATA1		Low expression in $CD4^+$ T-cells	Regulator of mitochondrial DNA replication.
TBP7	GATA1			Subunit of 26S protease required for ubiquitination.
SMARCA5*	GATA1		Low expression in $CD4^+$ T-cells	Associates with RSF1 and is required for chromatin assembly. Component of a chromatin-remodelling complex.
IVNSA1ABP	GATA1	LDB1	Low expression in $CD4^+$ T-cells	Involved in mRNA nuclear export and pre-mRNA splicing.
SUHW1	GATA1		Low expression in all cell types	
LMO2*	GATA1	E12, LDB1	Bone marrow	Regulator of erythroipoietic and endothelial development. Member of the SCL erythroid complex.
GALNT12	GATA1		$CD4^+$ T-cells and lung	Plays an important role in the initial step of mucin-type oligosaccharide biosynthesis in digestive organs.
EPOR*	GATA1, E12, E47, LDB1		Bone marrow	Required for differentiation and maturation of erythrocytes and programmed cell death
BRD2*	E12	GATA1, E47, LDB1	Low expression in thymus	Associated with E2F and involved in H4 acetylation
CTCFL*	E12	GATA1	Low expression in all cell types	Paralogue of the insulator CTCF which shares the same DNA-binding domain as CTCF and expressed in a mutual exclusive manner as CTCF. Its expression is activated in a wide-range of cancers. Possibly involved in epigenetic reprogramming of CTCF-binding sites.
ZNF 426	E12		$CD4^+$ T-cells	
RSF1*	E12		Low expression in $CD4^+$ T-cells	Associates with SMARCA5 and is required for chromatin assembly.
LYL1*	E12, E47, LDB1		Bone marrow	Dimerises with E2A. Chromosomal translocation leads to T-ALL

				Regulator of erythroid differentiation Highly similar in expression and function with SCL.
ZNF451	E12, E47, LDB1	GATA1	Low expression in $CD4^+$ T-cells	
PCQAP*	E12	E47	Low expression in $CD4^+$ T-cells	Mediates chromatin-directed transcriptional activation through protein complex formation.
JMJD2C*	E12	LDB1	$CD4^+$ T-cells	Contains histone demethylase activity. PHD finger domain protein. Overexpression leads to progression of cancer.
ETO2*	E12, E47, LDB1	SCL	Thymus	Breast-tumor suppressor gene. Repressor of early erythroid gene expression. Fusion partner of RUNX1 in leukemia-related translocation. Member of SCL erythroid complex.
ZNF510	E47, LDB1	GATA1	Low expression in all cell types	
EZH2*	LDB1	E47, GATA1, SCL	Thymus	Histone lysine methyltransferase. Associated with transcriptional repression. Methylate histone H1 and H3.
CD33	LDB1			Antigen expressed in myeloid lineage
SCL*	GATA1, E12, E47, SCL, LDB1, LMO2		High expression in HSCs and erythroid progenitors	Regulator of haematopoeitic development. Member of the SCL erythroid complex.

Table 5.2. Putative target promoters of members of the SCL erythroid complex. This table shows the expression pattern and function of the putative target genes identified for one or more of the 5 members of the SCL erythroid complex using method A (second column) or method B (third column). The genes marked with an asterisk were chosen for further characterisation (section 5.4.4).

5.4.4 Characterisation of a subset of putative target genes

5.4.4.1 Criteria for selection of subset of genes for further studies

In order to make additional characterisation of putative targets possible in the context of this project, the following criteria were used to select a subset of genes for further analyses.

- **Significantly enriched in all biological replicates**

Since method A was the more stringent approach for selecting statistically significant putative target genes, this gene list was used for choosing a subset of genes for further analyses. However, this gene list did not include any putative targets for SCL and LMO2. Therefore, putative target genes of SCL and LMO2 selected by method B were also included for further analysis.

- **Putative targets of more than one member of the SCL erythroid complex**

Since the main objective of this project was to identify direct transcriptional targets of the entire SCL erythroid complex in haematopoietic development, target genes which were identified by more than one of the transcription factor ChIP-on-chip assays were prioritised for further study.

- **Haematopoietic function**

Given that the SCL erythroid complex has been shown to regulate genes in the erythroid lineage, genes with known involvement in erythropoiesis or expression in the erythroid lineage were prioritised for further analysis.

- **Chromatin function**

Surprisingly, a number of target genes were identified whose functions were related to chromatin structure and function. As there is currently tremendous scientific interest in these proteins, the functions of which have widespread effects on the regulation of all genes in transcriptional programmes, these target genes were also selected for follow-up studies.

In summary, fourteen target genes were chosen which satisfied the first or second criteria and at least one of the functional criteria. These targets are highlighted with an asterisk in Table 5.2. Additional studies were then performed to validate the promoter binding events and further characterise the putative target genes. These included:

1. **Transcription factor motif identification in promoter regions:** Given that each promoter array element on the transcription factor array was approximately 1 kb in size, the potential binding site of the transcription factors to DNA sequence motifs in the promoter region were likely to be found within this one kilobase segment (however, the sites of binding could also be close to, but not within, this one kilobase of sequence). To identify the possible binding site of

the transcription factors in the promoter region, the DNA sequences of the promoters of each of the fourteen targets were screened for consensus transcription factor binding sites and the conservation of these sites were then compared across species. The presence of the relevant transcription factor binding motifs in regions of sequence conservation would provide additional evidence that the transcription factors had *bone fide* binding sites within these promoters.

2. **ChIP-qPCR validation of transcription factor binding events:** transcription factor-promoter interactions were then validated by ChIP-qPCR using the putative transcription factor binding motifs as locations around which the qPCR assays were designed. This validation was performed in K562 cells, where the interaction was initially identified and also in a second erythroid cell line, HEL. The validation of the transcription factor binding events in a second cultured cell line of a similar developmental state as K562 would support the biological relevance of these binding sites in regulating these target genes *in vivo*.

3. **Effect of knockdown of transcription factors on target gene expression:** The expression changes of these target genes were also investigated in time-course experiments of siRNA knockdowns of members of the SCL erythroid complex (to be discussed in Chapter 6). This would provide evidence that perturbations of the SCL erythroid complex affect the expression of the target genes.

All these studies together would provide further evidence that these genes are direct targets of the transcription factors which are found in the SCL erythroid complex. Figure 5.17 summarised the studies performed in characterising the putative direct target genes of the SCL erythroid complex.

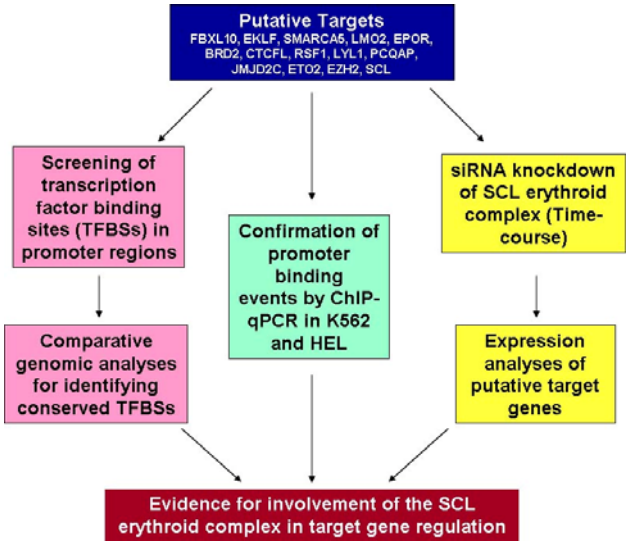


Figure 5.17. Follow-up characterisation of selected putative targets from the ChIP-on-chip experiments. The 14 putative target genes selected in the ChIP-on-chip study were further characterised in three analyses: (i) screening of conserved transcription factor binding sites; (ii) confirmation of promoter binding by the transcription factors by ChIP-qPCR in K562 and HEL; (iii) expression analyses of the 15 putative target genes in siRNA knockdown of the TFs. The information generated in these analyses will provide evidence for the involvement of the SCL erythroid complex in target gene regulation.

5.4.4.2 Transcription factor binding sites (TFBS) studies and comparative genomic analyses of enriched promoters

For screening of transcription factor binding sites in the promoter regions of the target genes, a 4 kb window (1 kb downstream of the TSSs and 3 kb upstream of the TSSs) was used to identify all possible transcription factor binding sites (TFBSs) using TESS and TFSEARCH. Transcription factor binding sites are conserved sequences with a certain degree of degeneracy which TFs recognise and bind. TESS and TFSEARCH are two web-based motif search algorithms which use the TRANSFAC TFBS database to identify TFBSs in genomic sequences (Chapter 1, section 1.3.4.2). Whilst the transcription factor binding events were likely to be present in the region encompassed by the approximately 1 kb contained within the promoter array elements, the windows for TFBS search were expanded to 4 kb to ensure that all possible TFBSs mapping near the TSS were identified. This would ensure that any motifs (and possible locations for transcription factor binding events) which were located close to, but not within, the sequences represented on the array elements, could be identified. In particular, E-box motifs of the E2A/SCL type and GATA motifs were identified within these promoter sequences. Given that the SCL erythroid complex binds to a composite E-box/GATA site separated by 9-12 bases in regulatory elements of its target genes, the location of clusters of E-box and GATA consensus sequences was of particular interest.

Following the mapping of relevant TFBS, the conservation of these binding sites across species was investigated. This allowed us to identify evolutionarily and functionally important DNA-binding motifs (Chapter 1, section 1.3.4.3). Multiple sequence alignments of the 4 kb of sequences around the TSSs were downloaded from the UCSC genome browser (http://genome.ucsc.edu/). These were derived from a variety of species including human, mouse, rat, rabbit, dog and chicken. Any DNA binding sites for E2A/SCL and GATA1 were carefully scrutinised for sequence conservation.

Relevant TFBSs and multi-species sequence alignments at the promoters of the fifteen target genes are shown in Appendix 4. Detailed descriptions of the possible TFBSs are given as follows:

A. BRD2: E-box and GATA motifs separated by 12 bases were identified. They are highly, though not completely, conserved across species.

B. CTCFL: Only one conserved GATA site was identified.

C. EKLF: Two possible TFBSs were identified. In the first one, three conserved E-box motifs separated by 6-12 bases were identified. In the second one, two conserved GATA sites separated by 47 bases were identified. However, no E-box and GATA motifs in close proximity were found in the promoter region studied.

D. EPOR: Three E-box motifs with high conservation across species separated by 6-13 bases were identified.

E. ETO2: E-box and GATA motifs separated by 9 bases were identified. Both motifs are fully conserved across species.

F. EZH2: Three possible TFBSs were identified. In the first two, E-box and GATA motifs separated by 10 to 23 bases were identified. They are highly, though not completely, conserved across species. In the third one, an E-box motif was identified with high conservation across species.

G. FBXL10: E-box and GATA motifs separated by 55 bases were identified. They are highly, though not completely, conserved across species.

H. JMJD2C: Three possible TFBSs were identified. In the first two, E-box and GATA motifs separated by 19 to 56 bases were identified. They are highly, though not completely, conserved across species. In the third one, an E-box motif was identified with full conservation across species.

I. LMO2: Two possible TFBSs were identified. In the first one, two conserved E-box motifs separated by 29 bases were identified. In the second one, one conserved GATA site was identified. However, no E-box and GATA motifs in close proximity were found in the promoter region studied.

J. LYL1: Two possible TFBSs were identified. In the first one, two fully-conserved GATA sites separated by 25 bases were identified. In the second one, one conserved E-box was identified. However, no E-box and GATA motifs in close proximity were found in the promoter region studied.

K. SCL: The +51 enhancer of SCL was selected for qPCR validation. This +51 enhancer contains the consensus E-box/GATA motifs separated by 9 bases. This was included in the validation as a positive control and reference for the qPCR assays.

L. SMARCA5: Three possible TFBSs were identified. In the first one, E-box and GATA motifs separated by 9 bases were identified. They are highly, though not completely, conserved across

species. In the third one, two GATA sites separated by 5 to 61 bases were identified. However, the conservation of these GATA sites was not high.

M. PCQAP: Four possible TFBSs were identified. In the first two, a single E-box motif was identified. The conservation for the first one was not high whereas there was no alignment with other species for the second one. In the other two, E-box and GATA motifs separated by 32 to 61 bases were identified. The conservation for the first one was not high whereas there was no alignment with other species for the second one.

N. RSF1: Only one conserved E-box motif was identified in the promoter region. No E-box and GATA motifs in close proximity were found in the promoter region studied.

5.4.4.3 ChIP-qPCR validation of promoter binding events

Based on the locations of conserved E-box and GATA motifs, qPCR assays were designed and validation of the transcription factor-promoter binding events was performed using ChIP-qPCR. In ChIP-qPCR, the input and ChIP DNAs were subjected to SYBR Green real-time quantitative PCR analyses. TFBS regions amplified in both input and ChIP DNA were quantified and compared. To normalise the fold enrichments above background, ChIP-qPCR was also performed for eleven negative control regions in the SCL locus which do not give enrichments above background for members the SCL erythroid complex (Appendix 1E). The average enrichment for these eleven regions was determined for every ChIP-qPCR assay and this value was used to scale the ChIP-qPCR enrichments of the promoter binding events so that the enrichment for negative control regions was a baseline of 1.

To identify statistical significant enrichments for the transcription factor binding sites tested for the selected putative target, cut-offs for significant enrichment were chosen. These cut-offs were different for different ChIP assays as the efficiency of antibodies differed. The enrichments of the eleven negative regions on the SCL locus were used as the baseline for determining significant fold enrichments. The standard deviations and average of these eleven regions in the two biological replicates for each ChIP assay of transcription factor were calculated. A fold enrichment cut-off was identified as the two standard deviations above the mean of enrichment i.e. a 99.45% confidence level is reached.

Three levels of validations were performed:

(i) Confirmation of ChIP-on-chip data was performed so that the identified TFBSs were tested for enrichment in ChIP-qPCR with ChIP DNAs from the assay which showed enrichments in ChIP-on-chip in K562 cells (analysed by method A).

(ii) Promoter binding events were tested for all transcription factors in the SCL erythroid complex in K562 cells to detect binding events which were missed using ChIP-on-chip.

(iii) Confirmation of binding events in a second, but somewhat similar, cell line was done to test the biological relevance of *in vivo* promoter-binding events in K562 cells.

A. Confirmation of ChIP-on-chip data

Since more than one region was found to be conserved with E-box and GATA motifs in many of the target promoters, all of them were first tested for enrichments in ChIP-qPCR. Initially, these regions were tested for enrichment in ChIP-qPCR with ChIP DNAs from the assay which showed enrichments in ChIP-on-chip in K562 cells (analysed by method A) (Table 5.3). In 10 out of the 14 chosen promoters, at least one qPCR region per promoter was shown to have a significant enrichment above the cut-off for at least one member of the SCL erythroid complex. However, enrichments for all tested regions of FBXL10, PCQAP, EZH2 and RSF1 were less than the cut-offs and unfortunately no other regions in the promoters showed conserved E-box and GATA motifs. These four genes were excluded in subsequent ChIP-qPCR analyses.

Region tested	GATA1 ChIP (>4.4)	E12 ChIP (>3.1)	E47 ChIP (>3.8)	LDB1 ChIP (>2.7)
BRD2		8.28		
CTCFL		6.76		
EKLF (1)	0	3	1.1	0.9
EKLF (2)	8.02	7.94	2.28	5.27
EPOR	20.57	4.24	4.14	3.14
ETO2		31.35	10.83	29.84
FBXL10	2.79			
JMJD2C (1)		31.35		
JMJD2C (2)				
JMJD2C (3)		3.55		
LMO2 (1)	4.53			
LMO2 (2)	1.2			
LYL1 (1)		37.99	14.91	62.64
LYL1 (2)		2.15	0.63	1.1
SCL	13.88	59.98	17.04	44.07
SMARCA5 (1)	0.71			
SMARCA5 (2)	1.53			
SMARCA5 (3)	19.77			
PCQAP (1)		0.78		
PCQAP (2)		0.64		
PCQAP (3)		0.97		
PCQAP (4)		1.55		
EZH2 (1)				1.3
EZH2 (2)				1.1
EZH2 (3)				1
RSF1		1.6		

Table 5.3. Fold enrichments putative target promoters tested in ChIP-qPCR. The fold enrichments of all the regions selected for the 14 putative target promoters in the confirmation by ChIP-qPCR are shown. The cut-offs of fold

enrichments used for each ChIP assay are shown in the first row in brackets. The regions which show a fold enrichment above the cut-offs are highlighted in yellow. These regions were also highlighted with an asterisk in Appendix 4 (where more than one region tested) and were chosen for subsequent ChIP-qPCR analyses.

B. Study of promoter binding for 5 members of the SCL erythroid complex in K562

From the results of the ChIP-on-chip studies, not all the selected putative target gene promoters were found to be bound by all 5 of the transcription factors in the SCL erythroid complex. However, all selected putative target promoters were tested by ChIP-qPCR for all the 5 transcription factors in K562 cells. This would allow for binding events which were missed using ChIP-on-chip to be detected by the more sensitive PCR-based assay. Two independent biological replicates were performed for each experiment.

For the ChIP-qPCR in K562, nine out of ten of putative target promoters showed significant enrichments above background passing the cut-offs in ChIP for at least one of the five transcription factors (Figure 5.18 and Table 5.4). In some cases where enrichments were only observed in ChIP-on-chip experiments for one transcription factor, they were shown to be enriched for some of the other transcription factors by ChIP-qPCR. For example, LYL1 was originally identified in the E12, E47 and LDB1 ChIP-on-chip but not in GATA1 and SCL - but it was shown to be enriched in the GATA1 and SCL ChIP-qPCR.

None of the putative target promoters were shown to be enriched in the ChIP of all five members of the SCL erythroid complex. However, four promoters or enhancers for SCL (CTCFL, LYL1, SCL and SMARCA5) were found to be enriched in the ChIP of four members of the complex including both E2A isoforms, GATA1, LDB1 and SCL. Promoters of EPOR and ETO2 were enriched in the ChIP of three members including GATA1, both E2A isoforms and LDB1 while promoters of BRD2 and EKLF were also enriched in the ChIP of three members including GATA1, the E12 isoform of E2A and LDB1. Promoter of LMO2 was only enriched in the GATA1 ChIP assay. Although the JMJD2C promoter showed significant enrichments above cut-off in the E12 ChIP assay in the initial screening (Table 5.3), no significant enrichments were shown in the ChIP of all five members of the SCL erythroid complex in the current study (Table 5.4).

Among the five members of the SCL erythroid, GATA1 bound to the largest number of promoters or enhancers (for SCL) (9 of them) while E12 and LDB1 bound to 8 of them. LMO2 was shown to bind to none of the promoters or enhancers tested (possibly due to the poor quality of the ChIP assay for LMO2).

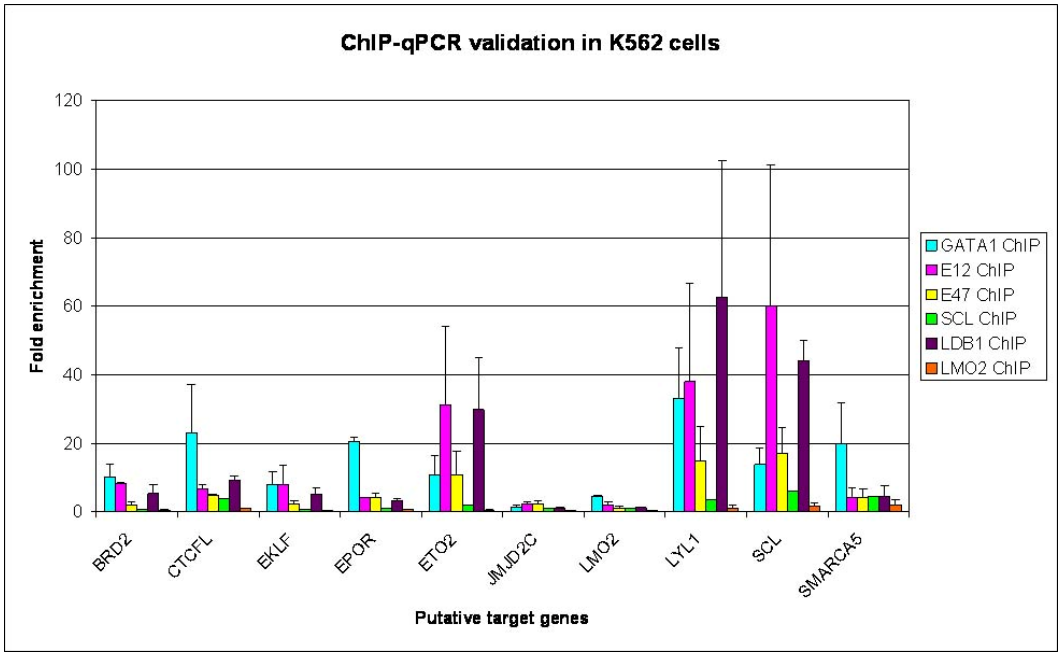


Figure 5.18. ChIP-qPCR analyses of selected putative target genes in K562. Histogram shows the fold enrichments of selected regions for putative target genes in ChIP-qPCR. Y-axis: fold enrichments above background. X-axis: putative target genes. The ChIP experiments represented by the colour bars are shown in the key on the right. Error bars show standard errors of two biological replicates.

Putative target	Fold enrichment					
	GATA1 (>4.4)	E12 (>3.1)	E47 (>3.8)	SCL (>3.1)	LDB1 (>2.7)	LMO2 (>3.1)
BRD2	10.13	8.28	1.97	0.69	5.37	0.44
CTCFL	23.03	6.76	4.99	3.97	9.15	1.02
EKLF	8.02	7.94	2.28	0.74	5.27	0.46
EPOR	20.57	4.24	4.14	1.21	3.14	0.66
ETO2	10.9	31.35	10.83	1.85	29.84	0.37
JMJD2C	1.35	2.33	2.42	1.2	1.17	0.5
LMO2	4.53	1.98	1.2	1.14	1.24	0.43
LYL1	32.98	37.99	14.91	3.58	62.64	1.1
SCL	13.88	59.98	17.04	6.22	44.07	1.67
SMARCA5	19.77	4.3	4.32	4.46	4.52	2.1
Total Validated	9	8	6	4	8	0

validated above threshold

Table 5.4. Fold enrichments of selected putative target promoters in ChIP-qPCR in K562 cells. The fold enrichments of the regions selected for the 10 putative target promoters in ChIP studies of five members of the SCL erythroid complex by ChIP-qPCR are shown. The cut-offs of fold enrichments used for each ChIP assay are shown in

the first row in brackets. The promoters which show a fold enrichment above the cut-offs are highlighted in green boxes. The total number of validated target genes for each member is shown in the bottom of the table.

A comparison of the interactions between promoters or enhancers (in the case of the +51 region of SCL) and transcription factors observed in ChIP-on-chip (analysed with methods A and B) and ChIP-qPCR was performed (Tables 5.5). Twenty-five binding events (56.8%) were observed in both assays (shown in green boxes in Table 5.5). Nine binding events (20.5%) were only observed in ChIP-on-chip (shown in blue boxes in Table 5.5), and six of these were identified by the less stringent method B analysis (which may be less reliable at identifying real binding events). Ten binding events (22.7%) were only observed in ChIP-qPCR (shown in pink boxes in Table 5.5). Overall, this analysis shows that the majority of ChIP-on-chip interactions were confirmed and that both approaches are complimentary at detecting interactions missed by the other method.

	Comparison between ChIP-on-chip and ChIP-qPCR in K562					
Putative target genes	GATA1	E12	E47	SCL	LDB1	LMO2
BRD2			*			
CTCFL						
EKLF				*		
EPOR						
ETO2				*		
JMJD2C					*	
LMO2		*			*	
LYL1						
SCL						
SMARCA5						

Validated in ChIP-on-chip only
Validated in both ChIP-on-chip and ChIP-qPCR
Validated in ChIP-qPCR only
* Interaction picked up by method B

Table 5.5. Comparison between ChIP-on-chip and ChIP-qPCR in K562 cells. The green boxes indicate promoter/enhancer binding events which were observed in both ChIP-on-chip and ChIP-qPCR. The pink and blue boxes indicate promoter/enhancer binding events which were observed in only in ChIP-qPCR or in ChIP-on-chip respectively

C. Study of promoter binding for 5 members of the SCL erythroid complex in HEL

Given that K562 is a cell line originally derived from a patient with chronic myeloid leukaemia (CML), the information derived from the ChIP-on-chip experiment performed in this study may not reflect the *bona fide* binding events found in normal erythroid cells. Confirming the promoter binding events in a second, but somewhat similar, cell line would provide further confidence of the true *in vivo* promoter-binding events (although cell culture may affect these binding events in both cell lines). Therefore, to further characterise the transcription factor binding at specific E-box or

GATA motifs in the selected promoter/enhancer regions, ChIP material from another cell line (HEL) was used. K562 and HEL are both erythroid progenitor cell lines which can be spontaneously differentiated into erythroid cells. However, developmentally, HEL cells represent a more mature erythroid cell population than K562. This was confirmed by flow-analysis of the erythrocytic surface marker glycophorin A which showed that a larger proportion of HEL cells expressed GPA than was found in K562 (Figure 5.19). Furthermore, HEL cells do not contain the BCL-ABL translocation (which is known to affect gene expression) (Martin and Papayannopoulou, 1982), suggesting that gene expression patterns in this cell line may reflect normal erythroid development more so than K562.

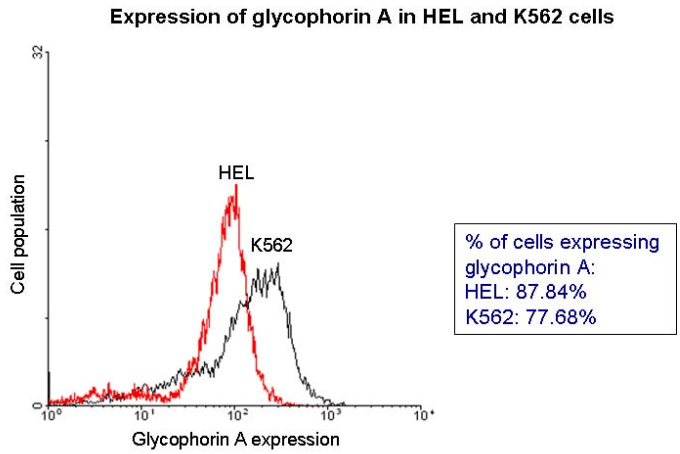


Figure 5.19. Flow analysis of glycophorin A expression in HEL and K562 cell lines. X-axis: Glycophorin A expression; y-axis: number of cells in population. The red curve shows the pattern for HEL cells while the black curve shows the pattern of K562 cells. % of cells in each population expressing GPA was calculated by WINMDI software and is shown in the box on the right.

For the ChIP-qPCR in HEL, eight out of ten of putative target promoters showed significant enrichments above baseline cut-offs in ChIP for at least one of the five transcription factors (Figure 5.20 and Table 5.6). Only the SCL +51 enhancer was shown to be enriched in the ChIP of all five members of the SCL erythroid complex. Two promoters (LYL1 and BRD2) were found to be enriched in the ChIP of four members of the complex including both E2A isoforms, GATA1, LDB1 and SCL. Promoter of ETO2 was enriched in the ChIP of three members including GATA1, both E2A isoforms and LDB1. Promoter of EKLF was enriched in the GATA1 and E12 ChIP assays. Promoters of LMO2 and EKLF were only enriched in the GATA1 ChIP assay. Two promoters were not enriched in the ChIP assays for all members of the SCL erythroid complex (CTCFL and JMJD2C). Again, the validation rates were in agreement with the quality of the ChIP assay – with validation for SCL and LMO2 showing the lowest levels.

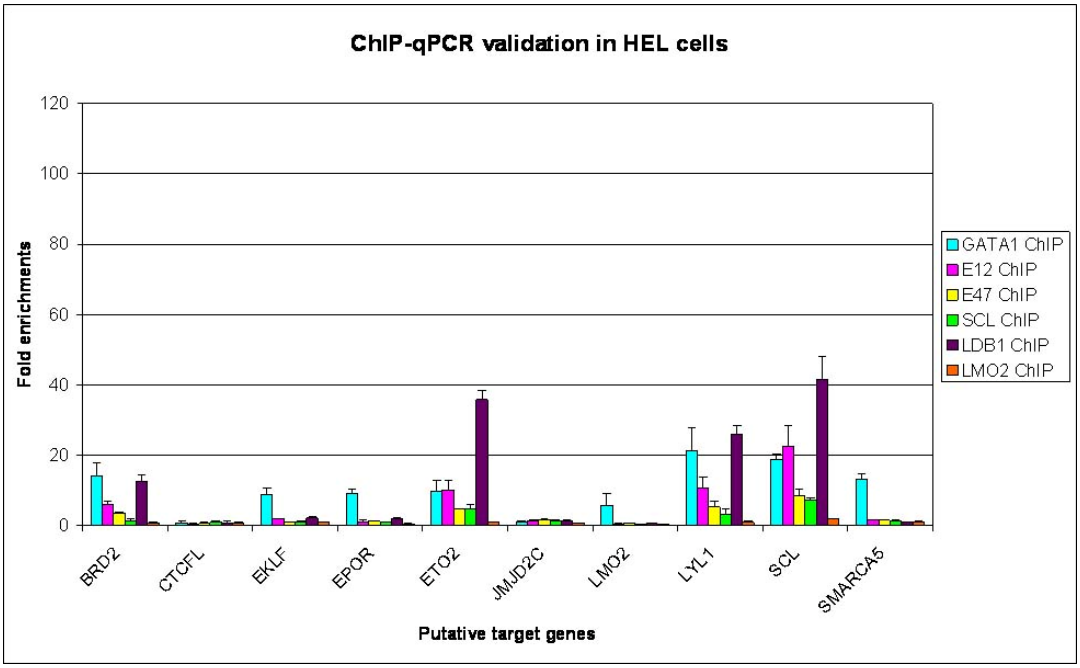


Figure 5.20. ChIP-qPCR analyses of selected putative target genes in HEL. Histogram shows the fold enrichments of selected regions for putative target genes in ChIP-qPCR. Y-axis: fold enrichments above background. X-axis: putative target genes. The ChIP experiments represented by the colour bars were shown in the key on the right. Error bars showed standard errors of two biological replicates.

Putative target	Fold enrichment					
	GATA1 (>3.6)	E12 (>1.5)	E47 (>2.1)	SCL (>1.4)	LDB1 (>2.2)	LMO2 (>1.3)
BRD2	14.02	5.93	3.47	1.41	12.67	0.75
CTCFL	0.74	0.41	0.55	0.95	0.77	0.58
EKLF	8.91	2.02	1.04	1.17	2.12	1
EPOR	9.05	1.05	1.26	1.02	2.1	0.47
ETO2	9.75	10.06	4.63	4.88	35.83	0.91
JMJD2C	1.15	1.35	1.61	1.3	1.33	0.69
LMO2	5.7	0.52	0.53	0.37	0.55	0.43
LYL1	21.28	10.63	5.24	3.06	25.94	0.86
SCL	18.81	22.5	8.37	7.33	41.67	1.81
SMARCA5	13.3	1.54	1.5	1.35	0.88	0.9
Total Validated	8	6	4	3	4	1

validated above threshold

Table 5.6. Fold enrichments selected putative target promoters in ChIP-qPCR in HEL cells. The fold enrichments of the regions selected for the 10 putative target promoters in ChIP studies of five members of the SCL erythroid complex by ChIP-qPCR are shown. The significance cut-offs of fold enrichments used for each ChIP assay are shown in the first row in brackets. The promoters which show a fold enrichment above the cut-offs are highlighted in green boxes. The total number of validated target genes for each member is shown in the bottom of the table.

D. Comparison between the validated targets of K562 and HEL

A comparison of the binding of transcription factors in the SCL erythroid complex to the promoters of their putative target genes was made between the data obtained in K562 and HEL cells. This was done to study the biological relevance of promoter binding in K562 cells. Table 5.7 compared the interactions found in each cell line. Twenty-two binding events (60%) were shown to be the same in both K562 and HEL cells (shown in green boxes in Table 5.7). Twelve binding events (32%) were only observed in K562 (shown in pink boxes in Table 5.7) while three binding events (8%) were only observed in HEL (shown in blue boxes in Table 5.7). Particularly, the CTCFL promoter was only enriched in all the ChIP assays in K562 cells but none in HEL cells. As a large proportion of promoter binding events were found in both cell lines, there is a high level of confidence that the data obtained in K562 is biologically relevant.

Deleted: the

Putative target genes	Validation in the ChIP assays in K562 and HEL cells					
	GATA1	E12	E47	SCL	LDB1	LMO2
BRD2						
CTCFL						
EKLF						
EPOR						
ETO2						
JMJD2C						
LMO2						
LYL1						
SCL						
SMARCA5						

HEL
K562 & HEL
K562

Table 5.7. Comparison between the ChIP-qPCR assays in K562 and HEL cells. The green boxes indicate promoter/enhancer binding events which were observed in both K562 and HEL. The pink and blue boxes indicate promoter/enhancer binding events were observed in only in K562 and HEL respectively.

5.4.4.4 Comparison of ChIP-on-chip, ChIP-qPCR and motif analyses

The results obtained in ChIP-on-chip, ChIP-qPCR and the *in silico* motif analyses were used to deduce whether a particular promoter was regulated by one or more members of the SCL erythroid complex or the whole complex. A summary of the combined data for the 24 putative target genes (described first in Table 5.2) are summarised in Table 5.8. The criteria used to make these deductions were as follows:

1. Target of any one of the five transcription factors:

- There must be evidence of significant enrichments in at least one of the ChIP analyses (ChIP-on-chip, ChIP-qPCR in K562 and ChIP-qPCR in HEL) for a gene to be considered as a direct target of any one transcription factor.

2. Target of the whole SCL erythroid complex (all five members):

- Significant enrichments must be observed in at least three ChIP assays (GATA1, E12 or E47, and LDB1) in either ChIP-on-chip, ChIP-qPCR in K562 or ChIP-qPCR in HEL. Also, both GATA and E-box motifs with the spacing of 9-12 bp must be identified in the promoters. Due to the poor quality of the SCL and LMO2 ChIP assays, a target was not required to demonstrate enrichments for these two TFs.

OR

- Significant enrichments must be observed in at least four ChIP assays (GATA1, E12 or E47, LDB1, and either SCL or LMO2) in either ChIP-on-chip, ChIP-qPCR in K562 or ChIP-qPCR in HEL. No motif data was required (this would allow *trans* interactions between enhancers (containing a motif) and promoters (not containing a motif) to be included as targets.

Based on this analysis, all 24 genes were considered as targets of at least one transcription factor, while 8 genes were considered to be direct targets of the whole SCL erythroid complex. These eight genes were BRD2, CTCFL, EKLF, ETO2, LYL1, SCL, SMARCA5 and EZH2. Four genes (EPOR, LMO2, ZNF451 and ZNF510) were found to be direct targets of GATA1, E2A (E12 or E47 or both) and LDB1. These genes may be direct targets of a novel complex containing GATA1, E2A and LDB1 but they may also be possible targets of the whole SCL erythroid complex due to the poor quality of the SCL and LMO2 antibodies. Only conserved E-box motifs were found in the EPOR and LMO2 promoters while no motif analyses was performed for ZNF451 and ZNF510.

Putative target gene	ChIP-on-chip (methods A & B)						ChIP-qPCR (K562)						ChIP-qPCR (HEL)						Motif analysis				Interpretation
	GATA1	E12	E47	SCL	LDB1	LMO2	GATA1	E12	E47	SCL	LDB1	LMO2	GATA1	E12	E47	SCL	LDB1	LMO2	E-box motif	GATA motif	Motif sequence conservation	Spacing between E-box and GATA	
BRD2																					High	12 bases	Target of SEC
CTCFL																					High		Target of SEC
EKLF																					High		Target of SEC
EPOR																					High		Target of GATA1, E2A and LDB1
ETO2																					Complete	9 bases	Target of SEC
JMJD2C																					High	19 bases	Target of E12 and LDB1
LMO2																					High		Target of GATA1, E12 and LDB1
LYL1																					High		Target of SEC
SCL																					Complete	9 bases	Target of SEC
SMARCA5																					High		Target of SEC
FBXL10																					High	55 bases	Target of GATA1
PCQAP																					Low	32 or 61 bases	Target of E2A
EZH2																					High	10 or 23 bases	Target of SEC
RSF1																					High		Target of E12
TFAM																							Target of GATA1
TBP7																							Target of GATA1
IVNSA1ABP																							Target of GATA1 and LDB1
SUHW1																							Target of GATA1
GALNT12																							Target of GATA1
CD33																							Target of LDB1
ZNF 426																							Target of E12
ZNF 526																							Target of GATA1
ZNF510																							Target of GATA1, E47 and LDB1
ZNF451																							Target of GATA1, E2A and LDB1

Validated targets
Non-validated targets
Not tested
Motif identified
Motif not identified

Table 5.8. Comparision of ChIP-on-chip, ChIP-qPCR and motif analyses. Putative targets identified in ChIP-on-chip, ChIP-qPCR in K562 and HEL are shown as green boxes for validated targets and red boxes for non-validated targets. E-box or GATA motifs identified in the motif analysis and contained within the sequence of the ChIP-qPCR assay are shown as blue boxes. Black boxes indicate no data available. The interpretation in the last column shows whether the putative target is confirmed as a direct target of one or more transcription factor or of the whole SCL erythroid complex (SEC) using the criteria detailed in section 5.4.4.4.

5.5 Discussion

The results of this Chapter describe the use of the ChIP-on-chip method to study the binding events of transcription factors of the SCL erythroid complex to promoter regions of target genes. Five transcription factors (GATA1, SCL, LMO2, LDB1, and two isoforms of E2A – E47 and E12) in the SCL erythroid complex were studied by ChIP-on-chip in K562 cells using an in-house transcription factor promoter array. A number of transcription factors related to haematopoietic development and chromatin remodelling were identified as putative targets of some or all members of the SCL erythroid complex in these ChIP-on-chip studies. These targets were confirmed in subsequent ChIP-qPCR, and by *in silico* transcription factor binding site and comparative sequence analysis.

5.5.1 Validation of promoter-binding events

Three levels of validation were performed in section 5.4.4.3 including (i) confirmation of ChIP-on-chip data, (ii) testing of promoter binding events for all transcription factors in the SCL erythroid complex in targets identified in K562 cells, and (iii) confirmation of biological relevance of K562 cells by identifying TF-binding events of K562 targets in HEL cells. This validation was all performed using ChIP-qPCR. The findings of each of these validation studies are discussed below.

(i) Confirmation of ChIP-on-chip data

ChIP-qPCR was first performed to confirm binding events obtained in ChIP-on-chip. Promoter binding events of members 10 out of 14 putative target genes (71%) were validated in ChIP-qPCR assay in K562 cells. This is approximately twice the validation rate obtained for the Affymetrix GeneChip platform for studies carried out in Chapter 4. Those promoters not validated may be false-positive targets identified on the promoter array. However, given the stringent statistical criteria by which these four were selected for validation (method A criteria), this is unlikely. Alternatively, the ChIP-qPCR assays may have been designed around TFBSs where the transcription factors do not bind. Thus, the ChIP-qPCR assays used here may result in false negatives. Therefore, these four non-validated genes may still be targets of the corresponding transcription factors. Other conserved TFBS inside or outside the promoter regions should be tested by ChIP-qPCR to confirm this.

(ii) Promoter binding events for five in the SCL erythroid complex in K562 cells

The first level of validation provided a general confidence that the promoter was bound by at least one member of the SCL erythroid complex. In the second level of confirmation, ChIP-qPCR was performed to detect binding events for all five members of the complex on each target promoter in K562 cells which may have been missed by ChIP-on-chip. Of 44 TF-DNA interactions detected by either ChIP-PCR or ChIP-on-chip, 56.8% were found by both methods, and 22.7% and 20.5% were found only in ChIP-qPCR or in ChIP-on-chip assays respectively. This would suggest that false negatives as well as false positives could be present in either assay. In fact, a large proportion of the "ChIP-on-chip only" interactions (6 out of 10) were only picked up using the less stringent method B of analyses. This may suggest that method B generates a high level of false positive target promoters.

(iii) Validation of K562 promoter-binding events in HEL cells

K562 is a cancerous cell line containing a BCR-ABL translocation which may induce changes in the expression pattern of genes. In fact, one piece of evidence showing that K562 may be abnormal is that thirteen out of the 24 putative target genes selected from the ChIP-on-chip study are normally expressed in the T-cell lineage (Table 5.2), despite K562 being an erythroid cell line. This indicates that there may be abnormal regulation of genes in K562. The HEL cell line lies at a similar stage of haematopoietic development as K562 cells and is BCR-ABL negative. Therefore it is a good cell line to validate K562 targets by ChIP-qPCR confirmation. However, since HEL is also a cell line, many issues associated with gene regulation in cell culture are not resolved.

A large proportion of promoter binding events (60%) for the 5 TFs of the SCL erythroid complex were observed in both K562 and HEL cells. This indicates that the majority of the data obtained in K562 is likely to be biologically relevant. Furthermore, these common binding events may mean that similar transcriptional programmes are found in both cell lines. Yet, 32% of the binding events were only observed in K562 but not HEL. These transcription/promoter interactions may be induced by the BCR-ABL translocation. However, they may also represent interactions which are found earlier in erythroid differentiation, given that K562 cells may be slightly more immature cells in the erythroid lineage than HEL cells (as determined by GPA expression). Following from that, the 8% of the binding events which were only observed in HEL cells may only be found later in erythroid development. Furthermore, the possibility that variations of the composition of the SCL erythroid complex, or different modes of regulation of these targets are present in K562 and HEL cells. Thus, while all of the targets found in K562 may also be targets in HEL, they may be regulated in different ways or by different TFs.

5.5.2 Validation of known target genes

Only three published target genes have previously been identified for the SCL erythroid complex, namely GYPA, c-kit and α-globin. However, since they are not transcription factors and were not included on the promoter array, they could not be validated in the ChIP-on-chip study in this Chapter. However, other direct target genes of members of the SCL erythroid complex were validated in the ChIP-on-chip assays. GATA1 has previously been shown to bind to the promoters of EPOR and EKLF (Anderson et al., 1998; Zon et al., 1991) and these bindings were confirmed in the data shown in this Chapter. The SCL +51 enhancer was previously reported to be bound by GATA1, SCL and LDB1 (Pawan Dhami, PhD thesis) and these binding events were also confirmed here.

A ChIP-on-chip study was performed for a TF complex containing SCL, HEB and E2A in the leukaemic Jurkat T-cell line (Palomero et al., 2006). The putative target genes obtained in this Chapter for SCL and E2A were compared with those obtained in the Jurkat study. No target genes were found in common in both studies. One possible reason to explain this is that the two studies were performed in cell lines derived from entirely different haematopoietic lineages (erythroid versus lymphoid) and the regulatory pattern may be very different, especially since SCL is expressed in Jurkat because of its involvement in T-acute lymphocytic leukaemia. Also, the transcription factors in these two cell lines may form different multiprotein complexes and thus may bind and regulate different target genes.

5.5.3 Novel targets of the SCL erythroid complex

Based on the criteria used in this study to define TF binding events at promoters, eight genes were likely to be direct targets of the whole SCL erythroid complex. These included BRD2, CTCFL, EKLF, ETO2, LYL1, SCL, SMARCA5 and EZH2 (section 5.4.4.4). However, apart from SCL where there is evidence that LMO2 binds to the +51 enhancer, there is no experimental data showing that the other 7 promoters are bound by LMO2. However, TFBS motif analysis confirmed the presence of conserved E-box and GATA motifs in the SCL, ETO2, BRD2, SMARCA5 and EZH2 promoters, providing additional confidence that they are direct targets of the whole SCL erythroid complex. However, there is still a possibility that LMO2 is not present in the SCL eythroid complex binding to the promoters of these seven genes. Thus, these eight genes may not be regulated by the whole SCL erythroid complex. The possibility that other LMO family members are part of the complex cannot be excluded. This line of reasoning could also explain why it was not possible to confirm binding events for all five members of the complex for these eight target genes. GATA2 has been shown to play the same role as GATA1 in the SCL erythroid complex in binding to the c-kit promoter (Lecuyer et al., 2002). In addition, other transcription factors may also form

novel complexes containing other transcription factors at the promoters of these targets. In fact, transcription factors such as SP1 and ETO2 have been identified as part of the SCL erythroid complex in certain contexts (Goardon et al., 2006; Lecuyer et al., 2002). Binding events for these TFs were not performed in the present study.

It is also important to note only 14 of the 24 genes listed in Table 5.2 were studied by ChIP-qPCR in this Chapter. The remaining 10 may also be direct target of the whole SCL erythroid complex. Thus, this gene list serves as an additional source of targets for analysis in the future.

5.5.4 The sequences of the putative binding sites of the SCL erythroid complex

For the eight targets described in section 5.5.3, conserved E-box and GATA motifs with spacing ranging from 9 to 12 bp were found in the promoter or enhancers of five of them - BRD2, ETO2, SCL, SMARCA5 and EZH2 (Figure 5.21). However, ChIP-qPCR assays showed enrichment around these composite sites for only three of them – SCL, ETO2 and BRD2. According to Wadman et al. (1997), the SCL erythroid complex binds to an E-box motif with consensus sequence of CAGGTG, followed 9 bp downstream by a GATA site. However, this canonical sequence with exactly the same sequence and spacing was only observed in the SCL +51 enhancer (Figure 5.21). Collectively, for the three sites which showed enrichment in ChIP-qPCR assays (SCL, BRD2 and ETO2) variations in (i) sequence, (ii) spacing and (iii) orientation of the sites were observed. This suggests that there is flexibility in terms of the requirements for TF binding to allow the components of the complex to reside on the same face of the DNA molecule.

SCL	n **caggtg**	nnnnnnnn	c**gata**a
BRD2	n **catctg**	nnnnnnnnnnn	t**tatc**t
ETO2	n **catctg**	nnnnnnnn	t**gata**a
SMARCA5[1]	n **cagctg**	nnnnnnnn	t**tatc**c
EZH2[2]	n **catctg**	nnnnnnnnn	g**tatc**c

Figure 5.21. Alignment of composite E-box/GATA motifs found in promoter sequences of five targets of whole SCL eythroid complex. The sequences of the E-box (green) and GATA (red) are highlighted for each target. n = any nucleotide; [1]= no significant ChIP-qPCR enrichment around this site; [2] = no significant ChIP-qPCR enrichment around this site.

In four of the eight target genes (LYL1, SMARCA5, CTCFL, ELKF), GATA sites were found in the regions assayed by ChIP-qPCR. However, the ChIP-qPCR data also suggests that either E2A or SCL binds to these regions in the absence of an obvious E-box motif. A possible reason to explain ChIP enrichment of TFs in these regions is that there is a looping of DNA sequences which brings GATA and E-box motifs on different regulatory elements into close proximity, allowing for the whole complex to bind (also see section 5.5.7). This mechanism could also be invoked to explain targets which have only an E-box motif in their promoters, although none of these eight targets fall

into that category. It is also possible that the SCL erythroid complex may bind to DNA sequences in addition to the consensus sequence suggested by Wadman et al (1997). To further characterise any of these TFBSs, additional assays would need to be performed in order to provide empirical evidence that these are indeed the actual sites of TF binding of members of the SCL erythroid complex. Gel shift assays are *in vitro* analyses that can be employed to confirm the binding of transcription factors to these DNA sequences. Moreover, mutation analysis can be used to investigate the requirement of these binding sites for driving expression in reporter assays.

5.5.5 Biological roles of novel targets of the SCL erythroid complex

The identification of novel targets of the SCL erythroid complex sheds new light on the role that this complex has in controlling transcriptional programmes in erythroid development. Of the eight genes thought to be novel direct targets of the whole SCL erythroid complex, four of them have known roles in haematopoietic development (EKLF, ETO2, LYL1 and SCL). SCL is a member of the complex itself and has been shown to be indispensable for haematopoietic development (see Chapter 1). ETO2 has previously been shown to be an interacting partner of the SCL where such interaction is related to down-regulation of early erythroid gene expression (Schuh et al., 2005). It was later demonstrated to be a novel member of the SCL erythroid complex (Goardon et al., 2006). Knockdown experiments of ETO2 demonstrated its involvement in governing erythroid and megakaryocytic differentiation (Goardon et al, 2006; Hamlett et al, 2008). EKLF is a transcription factor required for terminal erythropoiesis which regulates the expression of β-globin gene (Nuez et al., 1995). In EKLF knockout mice, definitive fetal liver erythropoiesis is disrupted, leading to lethality by embryonic day 15 (Nuez et al., 1995). LYL1 has overlapping expression patterns with SCL in mouse and is expressed in the erythroid and myeloid lineages and in ascular tissues (Visvader et al., 1991) (Chapter 1, section 1.4.2.1 C). LYL1 knockout mice were shown to be viable and have normal blood counts except for a reduced number of B-cells while Lyl1$^{-/-}$ haematopoietic stem cells showed severe defects in repopulation activities (Capron et al., 2006). Therefore, the SCL erythroid complex may play important roles in controlling specific aspects of erythroid development.

Deleted: ing

Deleted: leading to

Formatted: Superscript

A somewhat more surprising set of targets suggests that the SCL erythroid complex plays more generalised roles in controlling wide programmes of gene expression. Four direct target genes of the whole SCL erythroid complex are involved in regulating chromatin (BRD2, CTCFL, SMARCA5 and EZH2). Such regulators are known to have roles in regulating expression of a wide range of genes in many cell types. BRD2 dimerises with E2F and binds to acetylated histone H4 tails (Nakamura et al., 2007). It has also been shown to bind to the entire length of transcribed genes allowing RNA polymerase II to transcribe through nucleosomes (LeRoy et al., 2008). CTCFL

(BORIS) is a paralogue of the insulator CTCF which shares the same DNA-binding domain as CTCF and is expressed in a mutual exclusive manner to CTCF (Loukinov et al., 2002). These insulator proteins are involved in regulating chromatin domains, and three-dimensional chromatin looping structures, thus ensuring appropriate expression of genes. SMARCA5 associates with RSF1 and is required for chromatin assembly (Loyola et al., 2003). It is also a component of some chromatin-remodelling complexes (Bochar et al., 2000; Poot et al., 2000) (it should also be noted that RSF1 was also considered a target for the complex based on the genes listed in Table 5.2). Expression of SMARCA5 was also shown to be dysregulated in acute myeloid leukaemia (AML) and knockout studies also indicated that SMARCA5 is required for proliferation of haematopoietic progenitors (Stopka et al., 2000; Stopka and Skoultchi, 2003). EZH2 is a histone lysine methyltransferase which methylates histone proteins (Cao et al., 2002). Thus, through regulating chromatin factors, the SCL erythroid complex exerts transcriptional control over a large number of genes through epigenetic reprogramming or chromatin structure. This further emphasises its role as a key regulator of blood development.

5.5.6 Autoregulation of members of the SCL erythroid complex

In Chapter 4, evidence was provided from Affymetrix GeneChip analysis that members of the SCL erythroid complex were involved in regulation of the genes for other members of the complex. The data described in this Chapter further provides evidence of this regulation and that the whole SCL erythroid complex directly regulates expression of the genes of its own members. GATA1 was shown to bind to the promoter of LMO2 (section 5.4.4.3) – this confirms the findings of the Affymetrix GeneChip analysis in the GATA1 siRNA knockdown (although this was not confirmed by the qPCR validation of Affyemtrix expression changes). Furthermore, based on ChIP-on-chip and ChIP-qPCR, SCL and ETO2 were shown to be direct targets of the whole SCL erythroid complex [ETO2 can be a member of this SCL erythroid complex (Goardon et al., 2006)]. Regulation by individual members of the complex and regulation by the complex as a whole provides two levels of regulation - ensuring that the expression level of various members of the complex are tightly regulated in erythroid development. This further highlights the complex regulatory network that controls expression of the SEC.

5.5.7 Limitations of the ChIP-on-chip studies

The ChIP-on-chip assays in this Chapter have been demonstrated to identify DNA elements bound by proteins of interest. Over the last 11 years, since the discovery of the SCL erythroid complex in 1997 by Wadman et al, only three direct target genes (GYPA, c-kit, and α-globin) had been identified. In the study described in this Chapter, 8 additional direct target genes of the SCL erythroid complex were identified. However, there are likely to be many more targets of this

complex which have not been identified here. The limitations of using a transcription factor promoter array in ChIP-on-chip studies are discussed below.

- **Off-promoter binding**

The ChIP-on-chip study in this Chapter focused on an in-house array containing 1 kb array elements of promoters of transcription factors. The promoter sequences were identified in the genome using the FirstEF algorithm. However, if promoters were not accurately identified by FirstEF, the actual promoters would not be represented on the array. This may mean that promoter binding events for some target genes are missed in ChIP-on-chip. Furthermore, transcription factors may bind to other regulatory elements such as enhancers, silencers or distal promoters to regulate transcription. These binding events cannot be detected on the promoter array used in this study. Therefore, the current study only allowed the identification of a subset of genes regulated by the transcription factors in the SCL erythroid complex. One possible solution to this limitation is to increase the coverage of the genome represented on the array. Indeed promoter arrays having coverage of 10 kb around promoter regions are commercially available. Ultimately, the best solution would be to use whole genome tiling arrays which would remove any representation bias and ensure all possible binding events to be detected.

- **Resolution of the array**

In ChIP-on-chip studies of TF binding, the resolution of array elements plays a crucial role for localising the binding sites of transcription factors. The promoter array used in this Chapter has a resolution of 1 kb. Thus, the ChIP-on-chip analyses could only detect binding to the 1 kb fragment but could not identify the precise location of TF binding sites. Higher resolution arrays (using oligonucleotides as array elements) which have a greater coverage around promoters (>5 -10 kb around promoters) would resolve this issue. However, given this limitation, one can use TFBS and comparative sequence analysis to help refine the search for the site of TF binding, and then use ChIP-qPCR to identify and confirm specific interactions at precise locations. This was used extensively in this Chapter.

- **Efficiency of ChIP assays**

The antibodies used in the studies of this Chapter were evaluated by both western blotting and by binding to the +51 enhancer element on the SCL tiling array. Although they were shown to perform well in ChIP-on-chip and pick up high enrichments at the +51 region, some of them performed less well than the others. Particularly, the antibodies for SCL and LMO2 showed the lowest enrichments at the +51 region among all the antibodies and the specificity of LMO2 antibody could not be evaluated on western blotting. The inconsistency of the results obtained for SCL and LMO2 on the

promoter array was evident in the datasets generated by both statistical methods A and B. Furthermore, enrichments were also lower in the ChIP-qPCR for SCL and LMO2 assays. Evaluation of additional antibodies for SCL and LMO2 in ChIP would help resolve these issues and provide more reliable information on the binding profiles of these two TFs.

To circumvent issues with specific antibodies, another possible solution would be to express a tagged protein of the transcription factor under study in the cell line of interest. Some researchers have tried to express an epitope-tagged protein for ChIP studies (Greenbaum and Zhuang, 2002; Lee et al., 2002). Others have tried to co-express the target protein fused to a short biotin acceptor domain together with the biotinylating enzyme BirA from *Escherichia coli* (Viens et al., 2004). The resulting protein-DNA complexes could then be purified by streptavidin affinity. However, whether expression of a tagged protein can completely reflect the native binding patterns of the TF is always an issue for these types of studies.

- **Indirect protein-DNA interaction**

As shown from the data obtained in the ChIP-qPCR studies, enrichments were observed in regions where no TFBSs were found. DNA elements not directly bound by the transcription factor under study could be identified in the ChIP study via indirect protein-DNA interaction. During cross-linking, proteins are cross-linked with any DNA sequence in close proximity. It is possible that transcription factors bound to a primary DNA sequence, which interacts with a secondary DNA sequence by chromatin looping, are cross-linked all together with both DNA elements. As a result, enrichments could be observed in both DNA sequences even though there is no direct binding of the transcription factor with the secondary DNA sequence. Such chromatin looping events have been previously described in the literature. Long range interactions between distal *cis*-elements and the promoter of the α- and β-globin genes were reported to co-ordinate the expression of the genes (Song et al., 2007; Vernimmen et al., 2007). A study of the topoisomerase IIα gene confirmed that the recognition sites of Sp1 and Sp3 transcription factors in the distal and proximal promoters interact with each other via DNA looping (Williams et al., 2007). In fact, from the data obtained in the ChIP-qPCR assays in this Chapter, binding events were confirmed for both E2A and SCL on promoters where only GATA motifs were found (section 5.5.4). To test for such long-range interactions, chromosome conformation capture (3C) could be used (Dekker et al., 2002).

5.6 Conclusions

The work described in this Chapter demonstrated the use of ChIP-on-chip as a robust technique to identify promoters bound by the SCL erythroid complex in erythroid cells. Both published and novel direct target genes were identified. The data obtained in this Chapter provides useful

information for the generation of a transcription network governing aspects of erythroid development which will be described in Chapter 6 of this thesis.

Acevedo, L.G., Iniguez, A.L., Holster, H.L., Zhang, X., Green, R., and Farnham, P.J. (2007). Genome-scale ChIP-chip analysis using 10,000 human cells. BioTechniques *43*, 791-797.
Anderson, K.P., Crable, S.C., and Lingrel, J.B. (1998). Multiple proteins binding to a GATA-E box-GATA motif regulate the erythroid Kruppel-like factor (EKLF) gene. J Biol Chem *273*, 14347-14354.
Anguita, E., Hughes, J., Heyworth, C., Blobel, G.A., Wood, W.G., and Higgs, D.R. (2004). Globin gene activation during haemopoiesis is driven by protein complexes nucleated by GATA-1 and GATA-2. Embo J *23*, 2841-2852.
Attema, J.L., Papathanasiou, P., Forsberg, E.C., Xu, J., Smale, S.T., and Weissman, I.L. (2007). Epigenetic characterization of hematopoietic stem cell differentiation using miniChIP and bisulfite sequencing analysis. Proc Natl Acad Sci U S A *104*, 12371-12376.
Bernstein, B.E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D.K., Huebert, D.J., McMahon, S., Karlsson, E.K., Kulbokas, E.J., 3rd, Gingeras, T.R., *et al.* (2005). Genomic maps and comparative analysis of histone modifications in human and mouse. Cell *120*, 169-181.
Bochar, D.A., Savard, J., Wang, W., Lafleur, D.W., Moore, P., Cote, J., and Shiekhattar, R. (2000). A family of chromatin remodeling factors related to Williams syndrome transcription factor. Proceedings of the National Academy of Sciences of the United States of America *97*, 1038-1043.
Cao, R., Wang, L., Wang, H., Xia, L., Erdjument-Bromage, H., Tempst, P., Jones, R.S., and Zhang, Y. (2002). Role of histone H3 lysine 27 methylation in Polycomb-group silencing. Science (New York, NY *298*, 1039-1043.
Dahl, J.A., and Collas, P. (2007). Q2ChIP, a quick and quantitative chromatin immunoprecipitation assay, unravels epigenetic dynamics of developmentally regulated genes in human carcinoma cells. Stem cells (Dayton, Ohio) *25*, 1037-1046.
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.
Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science *295*, 1306-1311.
Delabesse, E., Ogilvy, S., Chapman, M.A., Piltz, S.G., Gottgens, B., and Green, A.R. (2005). Transcriptional regulation of the SCL locus: identification of an enhancer that targets the primitive erythroid lineage in vivo. Mol Cell Biol *25*, 5215-5225.
Dhami, P., Coffey, A.J., Abbs, S., Vermeesch, J.R., Dumanski, J.P., Woodward, K.J., Andrews, R.M., Langford, C., and Vetrie, D. (2005). Exon array CGH: detection of copy-number changes at the resolution of individual exons in the human genome. Am J Hum Genet *76*, 750-762.
Goardon, N., Lambert, J.A., Rodriguez, P., Nissaire, P., Herblot, S., Thibault, P., Dumenil, D., Strouboulis, J., Romeo, P.H., and Hoang, T. (2006). ETO2 coordinates cellular proliferation and differentiation during erythropoiesis. Embo J *25*, 357-366.
Greenbaum, S., and Zhuang, Y. (2002). Identification of E2A target genes in B lymphocyte development by using a gene tagging-based chromatin immunoprecipitation system. Proc Natl Acad Sci U S A *99*, 15030-15035.
Horak, C.E., Mahajan, M.C., Luscombe, N.M., Gerstein, M., Weissman, S.M., and Snyder, M. (2002). GATA-1 binding sites mapped in the beta-globin locus by using mammalian chIp-chip analysis. Proc Natl Acad Sci U S A *99*, 2924-2929.
Huang, D.Y., Kuo, Y.Y., Lai, J.S., Suzuki, Y., Sugano, S., and Chang, Z.F. (2004). GATA-1 and NF-Y cooperate to mediate erythroid-specific transcription of Gfi-1B gene. Nucleic Acids Res *32*, 3935-3946.
Iyer, V.R., Horak, C.E., Scafe, C.S., Botstein, D., Snyder, M., and Brown, P.O. (2001). Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature *409*, 533-538.
Johnson, K.D., Grass, J.A., Boyer, M.E., Kiekhaefer, C.M., Blobel, G.A., Weiss, M.J., and Bresnick, E.H. (2002). Cooperative activities of hematopoietic regulators recruit RNA polymerase II to a tissue-specific chromatin domain. Proc Natl Acad Sci U S A *99*, 11760-11765.

Koch, C.M., Andrews, R.M., Flicek, P., Dillon, S.C., Karaoz, U., Clelland, G.K., Wilcox, S., Beare, D.M., Fowler, J.C., Couttet, P., *et al.* (2007). The landscape of histone modifications across 1% of the human genome in five human cell lines. Genome Res *17*, 691-707.
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.
Lecuyer, E., Herblot, S., Saint-Denis, M., Martin, R., Begley, C.G., Porcher, C., Orkin, S.H., and Hoang, T. (2002). The SCL complex regulates c-kit expression in hematopoietic cells through functional interaction with Sp1. Blood *100*, 2430-2440.
Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., *et al.* (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. Science *298*, 799-804.
LeRoy, G., Rickards, B., and Flint, S.J. (2008). The double bromodomain proteins Brd2 and Brd3 couple histone acetylation to transcription. Molecular cell *30*, 51-60.
Lieu, P.T., Jozsi, P., Gilles, P., and Peterson, T. (2005). Development of a DNA-labeling system for array-based comparative genomic hybridization. J Biomol Tech *16*, 104-111.
Loukinov, D.I., Pugacheva, E., Vatolin, S., Pack, S.D., Moon, H., Chernukhin, I., Mannan, P., Larsson, E., Kanduri, C., Vostrov, A.A., *et al.* (2002). BORIS, a novel male germ-line-specific protein associated with epigenetic reprogramming events, shares the same 11-zinc-finger domain with CTCF, the insulator protein involved in reading imprinting marks in the soma. Proceedings of the National Academy of Sciences of the United States of America *99*, 6806-6811.
Loyola, A., Huang, J.Y., LeRoy, G., Hu, S., Wang, Y.H., Donnelly, R.J., Lane, W.S., Lee, S.C., and Reinberg, D. (2003). Functional analysis of the subunits of the chromatin assembly factor RSF. Molecular and cellular biology *23*, 6759-6768.
Martin, P., and Papayannopoulou, T. (1982). HEL cells: a new human erythroleukemia cell line with spontaneous and induced globin expression. Science *216*, 1233-1235.
Nakamura, Y., Umehara, T., Nakano, K., Jang, M.K., Shirouzu, M., Morita, S., Uda-Tochio, H., Hamana, H., Terada, T., Adachi, N., *et al.* (2007). Crystal structure of the human BRD2 bromodomain: insights into dimerization and recognition of acetylated histone H4. The Journal of biological chemistry *282*, 4193-4201.
Nuez, B., Michalovich, D., Bygrave, A., Ploemacher, R., and Grosveld, F. (1995). Defective haematopoiesis in fetal liver resulting from inactivation of the EKLF gene. Nature *375*, 316-318.
O'Neill, L.P., and Turner, B.M. (1996). Immunoprecipitation of chromatin. Methods in enzymology *274*, 189-197.
O'Neill, L.P., VerMilyea, M.D., and Turner, B.M. (2006). Epigenetic characterization of the early embryo with a chromatin immunoprecipitation protocol applicable to small cell populations. Nat Genet *38*, 835-841.
Ogilvy, S., Ferreira, R., Piltz, S.G., Bowen, J.M., Gottgens, B., and Green, A.R. (2007). The SCL +40 enhancer targets the midbrain together with primitive and definitive hematopoiesis and is regulated by SCL and GATA proteins. Mol Cell Biol *27*, 7206-7219.
Orlando, V., Strutt, H., and Paro, R. (1997). Analysis of chromatin structure by in vivo formaldehyde cross-linking. Methods *11*, 205-214.
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.
Poot, R.A., Dellaire, G., Hulsmann, B.B., Grimaldi, M.A., Corona, D.F., Becker, P.B., Bickmore, W.A., and Varga-Weisz, P.D. (2000). HuCHRAC, a human ISWI chromatin remodelling complex contains hACF1 and two novel histone-fold proteins. The EMBO journal *19*, 3377-3387.
Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., *et al.* (2000). Genome-wide location and function of DNA binding proteins. Science *290*, 2306-2309.

Rylski, M., Welch, J.J., Chen, Y.Y., Letting, D.L., Diehl, J.A., Chodosh, L.A., Blobel, G.A., and Weiss, M.J. (2003). GATA-1-mediated proliferation arrest during erythroid maturation. Mol Cell Biol *23*, 5031-5042.
Schuh, A.H., Tipping, A.J., Clark, A.J., Hamlett, I., Guyot, B., Iborra, F.J., Rodriguez, P., Strouboulis, J., Enver, T., Vyas, P., *et al.* (2005). ETO-2 associates with SCL in erythroid cells and megakaryocytes and provides repressor functions in erythropoiesis. Mol Cell Biol *25*, 10235-10250.
Song, S.H., Hou, C., and Dean, A. (2007). A Positive Role for NLI/Ldb1 in Long-Range beta-Globin Locus Control Region Function. Molecular cell *28*, 810-822.
Sun, L., Huang, L., Nguyen, P., Bisht, K.S., Bar-Sela, G., Ho, A.S., Bradbury, C.M., Yu, W., Cui, H., Lee, S., *et al.* (2008). DNA methyltransferase 1 and 3B activate BAG-1 expression via recruitment of CTCFL/BORIS and modulation of promoter histone methylation. Cancer research *68*, 2726-2735.
Valverde-Garduno, V., Guyot, B., Anguita, E., Hamlett, I., Porcher, C., and Vyas, P. (2004). Differences in the chromatin structure and cis-element organization of the human and mouse GATA1 loci: implications for cis-element identification. Blood *104*, 3106-3116.
Vernimmen, D., De Gobbi, M., Sloane-Stanley, J.A., Wood, W.G., and Higgs, D.R. (2007). Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression. The EMBO journal *26*, 2041-2051.
Viens, A., Mechold, U., Lehrmann, H., Harel-Bellan, A., and Ogryzko, V. (2004). Use of protein biotinylation in vivo for chromatin immunoprecipitation. Analytical biochemistry *325*, 68-76.
Visvader, J., Begley, C.G., and Adams, J.M. (1991). Differential expression of the LYL, SCL and E2A helix-loop-helix genes within the hemopoietic system. Oncogene *6*, 187-194.
Walker, G.T. (1993). Empirical aspects of strand displacement amplification. PCR methods and applications *3*, 1-6.
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.
Williams, A.O., Isaacs, R.J., and Stowell, K.M. (2007). Down-regulation of human topoisomerase IIalpha expression correlates with relative amounts of specificity factors Sp1 and Sp3 bound at proximal and distal promoter regions. BMC molecular biology *8*, 36.
Xu, Z., Huang, S., Chang, L.S., Agulnick, A.D., and Brandt, S.J. (2003). Identification of a TAL1 target gene reveals a positive role for the LIM domain-binding protein Ldb1 in erythroid gene expression and differentiation. Mol Cell Biol *23*, 7585-7599.
Zeng, P.Y., Vakoc, C.R., Chen, Z.C., Blobel, G.A., and Berger, S.L. (2006). In vivo dual cross-linking for identification of indirect DNA-associated proteins by chromatin immunoprecipitation. BioTechniques *41*, 694, 696, 698.
Zon, L.I., Youssoufian, H., Mather, C., Lodish, H.F., and Orkin, S.H. (1991). Activation of the erythropoietin receptor promoter by transcription factor GATA-1. Proc Natl Acad Sci U S A *88*, 10638-10641.