Chapter 3

Construction and Validation of the SCL Genomic Tiling Path Microarrays and Characterization of SCL Expressing and Non-expressing Cell Lines

3.1. Introduction

In recent years, genomic microarrays have emerged as a preferred platform for genomic analysis including annotating the transcriptome, elucidating DNA-protein interactions, and for comparative genomic hybridisations to detect genomic copy number changes among many other applications. The development and use of chromatin-immunoprecipitation coupled with genomic microarrays (ChIP-chip) has greatly facilitated the elucidation and annotation of functional elements and site-specific DNA-protein interactions in the genomes of various organisms (discussed in chapter 1, section 1.6.4) including mammalian genomes.

Genomic tiling arrays (representing non-repetitive DNA across a genomic region) were first used in the mammalian system to map the binding distribution of the haematopoietic transcription factor GATA-1 across ~75 kb of the human β-globin locus in K562 cells (Horak et al. 2002). Since then, tiling arrays have been used extensively to map DNA-protein binding sites across regions of interest, whole chromosomes or in some cases across whole genomes (Martienssen et al. 2005; Martone et al.2003; Euskirchen et al. 2004; Cawley et al. 2004; Pokholok et al. 2005; Schübeler et al. 2004; Kim et al. 2005). Although complete tiling arrays of the human genome studies would prove very useful in elucidating genome-wide DNA-protein interactions, a major limitation is the huge costs involved to carry out such studies. To this end, it is more realistic to use tiling path arrays for a discrete genomic region which would be an ideal resource from which to identify novel regulatory elements and interactions in their genomic context. Thus, a tiling array across the SCL locus, would greatly aid in the understanding of the complex regulation of SCL expression.

The SCL locus is well-characterized as a number of regulatory regions directing SCL expression in distinct compartments during embryonic development are already known (see chapter 1, section 1.9.5). Several studies have identified regulatory sequences located distant from SCL, for example, the erythroid enhancer located 40 kb downstream of SCL promoter 1a in mouse (Delabesse et al. 2005). The identification of this enhancer, suggests that in order to fully understand the regulation of SCL and identify all

of its regulatory elements, it is necessary to examine sequences which are located distant from SCL – both upstream and downstream from the gene itself.

Long range sequence comparisons between the human and mouse loci revealed that the genomic regions surrounding the SCL gene on either side are structurally similar in the two species (Gottgens et al. 2000). The upstream region of SCL contains the SIL gene which is an immediate-early gene and ubiquitously expressed in proliferating cells (Izraeli et al. 1997). The downstream region contains the MAP17 gene which is expressed at significant levels in the epithelial cells of human adult kidney (Kocher et al. 1996). However, downstream of the 3' end of the MAP17 gene is the cytochrome p450 gene cluster - the genes in this cluster are not orthologous in human and mouse (Gottgens et al. 2001) and are mainly expressed in liver and kidney (Henderson et al. 1994).

The majority of the studies undertaken to understand the regulation of SCL expression have been carried out in mice using mouse cell lines, transfection assays and transgenic models (Gottgens et al. 1997; Sanchez et al. 1999; Sinclair et al. 1999). Similar studies in humans have utilized human haematopoietic cell lines and tissues to understand biological function of SCL in blood development (Bernard et al. 1992; Leroy-Viard et al. 1994). Since many of the human cell lines were originally established from patients with a variety of myeloid and lymphoid leukaemias, detailed characterizations have reported a number of genomic imbalances involving various chromosomes with amplifications, deletions and translocations (see Table 3.2). Despite having various structural and numerical genomic imbalances, these cell lines have served as excellent models to study functions of genes such as SCL, that are important in blood development and haematopoietic differentiation (Green et al. 1991; Leroy-Viard et al. 1994; Shimamoto et al. 1995).

Some of the common human cell lines used in widespread haematopoietic studies include K562, Jurkat, HL60, HPB-ALL, U937 and HEL (Delabesse et al. 2005). Based on their SCL expression, the cell lines can be broadly classified as described below.

1. **SCL expressing cell lines***:* Cell lines exhibiting either normal or inappropriate SCL expression include:

   a) K562: This is a myelogenous cell line in which the predominant cell-type present has been described as a highly undifferentiated granulocytic cell (Lozzio and Lozzio.1977). The cell line can be induced to differentiate down the granulocytic or erythroid lineages and therefore has been used to study the biological role of SCL in cell-proliferation and differentiation in these lineages (Green et al. 1991; Green et al. 1993).

   b) Jurkat: This is a lymphoblastic T-cell line (Schneider et al. 1977). SCL expression is down-regulated in common lymphoid progenitors and is subsequently silenced in

progenitors and terminally differentiated cells of T- and B- lineages. Although Jurkat is a T-cell line, SCL is inappropriately expressed and the activation of SCL has not been linked to any genomic rearrangements at the SCL locus (Leroy-Viard et al. 1994).

2. **SCL non-expressing cell lines:** These include:

c) HL60: This cell line is promyelocytic (Collins et al. 1977) and can be induced to differentiate down the monocytic lineage. SCL expression is turned off in monocyte precursors and, thus, SCL is not expressed in HL60.

d) HPB-ALL is another T-cell line (Morikawa et al. 1978) which, unlike Jurkat, does not exhibit SCL expression.

Identification of chromosomal imbalances in cells using conventional cytogenetic analysis, fluorescence in situ hybridisation (FISH) and comparative genomic hybridisation (CGH) on metaphase chromosomes have a limited resolution of 3 to 5 Mb that is defined by the use of metaphase chromosomes. Development of genomic microarrays, which can measure quantitative changes in genomic copy number, has helped to resolve this problem to a great extent by using cloned DNA segments or PCR-generated sequences instead of metaphase chromosomes as targets for hybridisation (Solinas-Toldo et al. 1997; Pinkel et al. 1998; Albertson et al. 2000; Fiegler et al. 2003). These arrays can detect chromosomal anomalies which are undetectable by FISH, including some involving DNA sequences 30-50 kb in length (Albertson and Pinkel 2003; Mantripragada et al. 2004). A recent study reported the development of a very high resolution array-CGH platform that is highly sensitive and can measure copy-number changes accurately at the resolution of single exons (Dhami et al. 2005). However, even array-CGH methods are not without their limitations, for example array-CGH, at any resolution, will not detect balanced translocations and thus karyotype analysis is required in such cases (Shaffer and Bejjani 2004). Therefore, ideally, the combination of conventional cytogenetic techniques and high resolution array-CGH methods could lead to the identification of the molecular basis of many chromosomal imbalances involved in disease and in cell lines used as experimental models.

3.2. Aims of this chapter

One of the overall aims of the study presented in this thesis was to develop a robust and reproducible array-based platform which would be as sensitive as real-time PCR and could be used in ChIP-chip studies across the SCL locus. To this end, the aims of the work described in this chapter were:

1. To construct and validate sensitive genomic tiling path microarrays containing the human and mouse SCL loci at a resolution of approximately 400-500 bp.
2. To characterize the human haematopoietic cell lines that were selected to be used in ChIP-chip assays across the SCL locus. This would determine the genomic integrity of the SCL locus, which would aid in the interpretation of the data obtained from the ChIP-chip studies presented in subsequent chapters of this thesis.

3.3. Construction of the SCL Genomic Tiling Path Microarray

3.3.1. The array chemistry

As discussed in section 3.1, studies have demonstrated that SCL regulatory elements are located quite distant to SCL, thereby suggesting that it was important to interrogate larger genomic regions to elucidate all key regulatory interactions that could play a role in SCL regulation (Gottgens et al. 2000; Delabesse et al. 2005). Therefore, in order to increase the likelihood that, any novel regulatory elements distant from SCL could be identified, the SCL tiling path arrays were constructed across a larger genomic region than was previously analyzed (90 kb in Delabesse et al. 2005). The construction of such a sensitive array platform for the SCL region was made possible by using the 5'-aminolink array surface chemistry developed at the Sanger Institute, which allows single-strands of DNA derived from double-stranded PCR products to be retained on the surface of the microarray slide (Dhami et al. 2005). This involves incorporation of a 5'-(C6) amino-link modification at the end of one strand of a double-stranded PCR product.

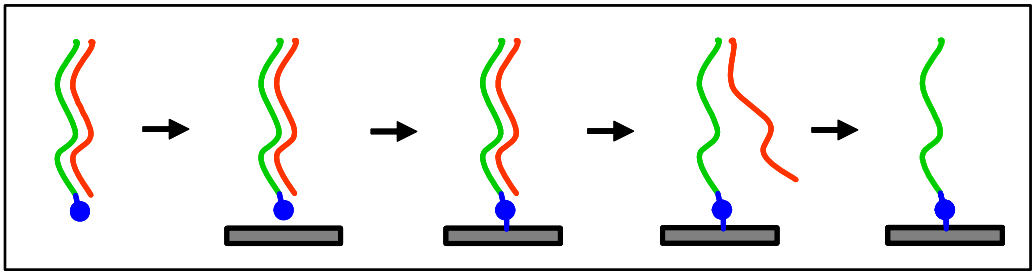


Figure 3.1: Schematic diagram shows the approach adopted to make single-stranded PCR products. Double-stranded PCR products (red/green denote strands) containing a 5'-(C6) amino-linked modification on one strand (blue circle on the green strand) are arrayed onto the surface of the slide (grey bar). Covalent attachment occurs via the 5' amino-link (blue line) and the slide surface. Denaturation of the PCR product renders them single-stranded.

This 5'-aminolink modification drives a covalent attachment between the modified strand and the surface of the slide (Figure 3.1). Upon slide processing, which involves physical and chemical denaturation, the strand attached to the slide is retained, whereas, the unmodified strand is removed. As a result, the single-stranded DNA molecules attached

at one end to the surface of the slide provide an ideal substrate to hybridise with the labeled DNA sample. The sensitivity of this array platform has already been proved by testing its ability to accurately report copy-number changes of individual exons (from 571 bp to 139 bp) in the human genome (Dhami et al. 2005). Thus, utilizing this type of array platform would be ideal to construct a sensitive, quantitative and high resolution genomic tiling path microarray representing the SCL loci in human and mouse.

3.3.2. Defining the genomic regions to be represented on the SCL arrays

The SCL gene in human is flanked upstream by the human SIL and KCY genes and downstream by the MAP17, CYP4A22 and CYP4Z1 genes. Similarly in mouse, SCL is flanked upstream by the mouse SIL and KCY genes and downstream by the MAP17 and Cyp4x1 genes. Figure 3.2 shows a schematic of the human and mouse SCL loci and the genomic regions contained on the SCL tiling arrays.

The construction of the SCL genomic tiling path array was carried out in two phases (1st generation array and 2nd generation or final array) owing to the availability of the finished genome sequence at the human and mouse SCL loci at the time. In the 1st generation array, approximately 193 kb and 199 kb of the genomic regions were chosen in human and mouse respectively to generate contiguous, non-overlapping amplicons. However, upon the release of a subsequent build of the finished human genome sequence (NCBI build 35), it was apparent that regions of the first tiled array were no longer contiguous and the genomic region across the SIL gene had been reorganised in the finished sequence. Thus, it was necessary to fill all of the gaps which resulted in extending the human array to include the 5' end of the KCY gene; the genomic region encompassed on the final human array was approximately 256 kb. In the case of final mouse array, the genomic region was extended further to approximately 207 kb but did not include the 5' end of the KCY gene.

3.3.3. Primer design and PCR amplification

The primers for the human and mouse arrays were designed as described in chapter 2 to generate amplicons that were not more than 600 bp in size; some amplicons contained short interspersed and low complexity repeats. All human and mouse primer pair sequences and their respective genomic coordinates are listed in Appendix 1 and Appendix 2 respectively. Table 3.1 summarizes the characteristics of the human and mouse tiling arrays with respect to their construction.

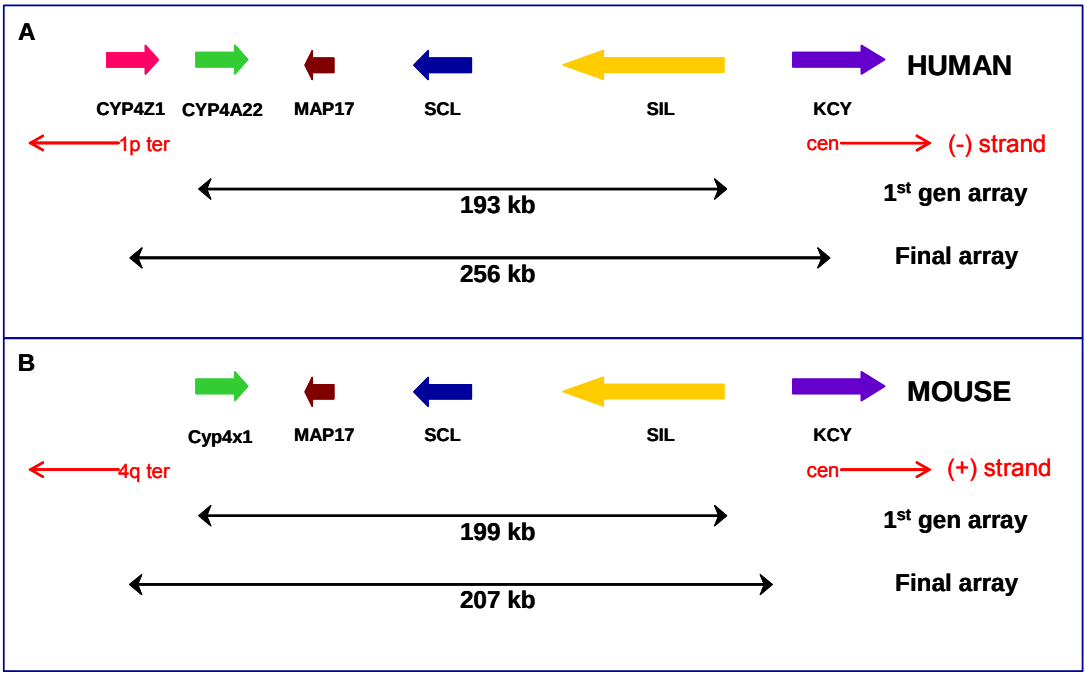


Figure 3.2: A schematic of the genomic regions across the human and mouse SCL loci. The size of the genomic regions included in the 1st generation array and the final array in human and mouse are shown by the black double-headed arrows. The thick, coloured arrows represent the genes. The gene order and the direction of transcription in human and mouse are shown as annotated on the chromosomes 1 and 4 in the respective genomes. The upstream SIL gene and downstream MAP17 gene are highly conserved in both species. The human SCL locus is annotated on the (-) strand of human chromosome 1 whereas the mouse SCL locus is annotated on the (+) strand of mouse chromosome 4. The orientation of the loci with respect to the centromere (cen) and telomere (ter) is shown at the bottom with red arrows.

		HUMAN		MOUSE	
		1st generation array	Final array	1st generation array	Final array
Genomic region included on the array (in kb)		193	256	199	207
Total no. of amplicons designed		171	419	283	530
No. of amplicons passed 1st round PCR		153	407	265	492
No. of amplicons passed 2nd round PCR	Products with correct bands	143	372	250	438
Validation of amplicons by electrophoresis	Missing bands	0	3	1	18
	Multiple bands	4	17	6	21
	Weak bands	6	15	8	12
	Wrong sized products	0	0	0	3
Validation of amplicons by sequencing (amplicons with >90% identity to the expected sequence were considered good matches)	No. of amplicons sequenced	na	180	na	na
	Good matches	na	145	na	na
	Not good matches	na	5*	na	na
	No sequence reads	na	30	na	na
In silico BLAST analysis	Not annotated/homology/repeat	5	5	23	27
Final number of amplicons "passed"		138	367	227	411
Average pass rate (%)		80.7	87.6	80.2	77.5
Number of amplicons in the final array			367		411
Average product size (bp)		420	458	400	443

Table 3.1: Summary of results of the construction of the SCL genomic tiling path array. The tiling path array was constructed in human and mouse in two phases (1st generation and final array). The table lists the sizes of the genomic regions chosen to construct the tiles for human and mouse arrays, the total number of

amplicons designed and the total number of amplicons that finally "passed" after two rounds of PCR amplification. The human and mouse SCL array 2nd round PCR amplicons were electrophoresed on agarose gels and the bands were scored visually. The array elements which were missing, multiple, weak bands or were the wrong-sized products were excluded from the final data sets in all the analyses. In total, 180 randomly selected amplicons from the final human SCL array were sequenced; 145 of the amplicons were confirmed to originate from the expected genomic region; 30 amplicons did not generate successful sequence reads and 5 amplicons were considered not to be a good match to the expected sequences. These 5 (shown with an asterisk) were also products which gave multiple bands as visualised on agarose gels. na = not analysed. The genomic sequences of the human and mouse array elements were mapped against the respective genome sequences using BLASTN (Altschul et al. 1997). Elements showing sequence homology, not mapping to the expected region, not annotated in NCBI build 35 of the human genome or mapping extensively within repeat regions were also excluded.

The array amplicons were generated using a two round PCR amplification procedure (see chapter 2). For the 1st generation array, human and mouse amplicons were successfully amplified in the first round of PCR using total human or total mouse genomic DNA as a template (hereafter called the genomic set). Subsequently, the whole tile for the final array in human and mouse was generated using DNA derived from bacterial PAC and BAC clones as a template in the first round PCR (hereafter called the clone set) in order to increase the success rate of primer pairs. BAC and PAC clones spanning the SCL locus in human (PAC RP1-18D14 and BAC RP11-332M15) and mouse (BACs RP23-242O20, RP23-32K12 and RP-23-246H17) were identified using ENSEMBL Cytoview (http://www.ensembl.org) and used to amplify the appropriate amplicon set.

3.3.4. Controls included on the arrays

A set of PCR products corresponding to human X-linked and autosomal sequences was included as control array elements on the 1st generation SCL tiling path array. The performance of these PCR products was previously tested in competitive hybridisation experiments in order to obtain accurate measurement of copy-number changes (Dhami et al. 2005). Therefore, these PCR products provided an ideal resource to be included as control array elements to test the performance of the SCL array at reporting accurate quantitative measurements of DNA copy number.

Given that the SCL tiling path arrays would be used primarily for ChIP-chip studies (see chapters 4 and 5), sets of sequences were included on the final arrays which had previously been shown to be associated with histone modifications and/or bind sequence-specific transcription factors. The SCL locus had been studied previously using histone modifications (Delabesse et al. 2005) and therefore the SCL tiled amplicons themselves served as a good source of internal control regions for ChIP-chip assays for histone modifications. For ChIP-chip assays for sequence-specific transcription factors, genomic sequences containing known GATA-1 binding sites were

spotted onto the array. In addition to the SCL promoter 1a, which is known to bind GATA-1 in erythroid cells (Bockamp et al. 1995), three array elements (Hb/9BG, Hb/32BG-1, and Hb/32BG-2) containing GATA-1 binding sites at the β-globin locus (Horak et al. 2002) were also included on the array; collectively, these amplicons served as positive controls on the SCL tiling path array – at least for the GATA-1 ChIP-chip experiments described in chapter 4. The primer sequences used to amplify these controls are listed in Appendix 3A.

3.4. Validation of the SCL genomic tiling path microarrays

3.4.1. Assessment of human and mouse SCL array amplicons by electrophoresis

The human and mouse array elements generated after two rounds of PCR amplification were analysed by agarose gel electrophoresis. Each amplicon was assessed visually for the band produced on the gel.

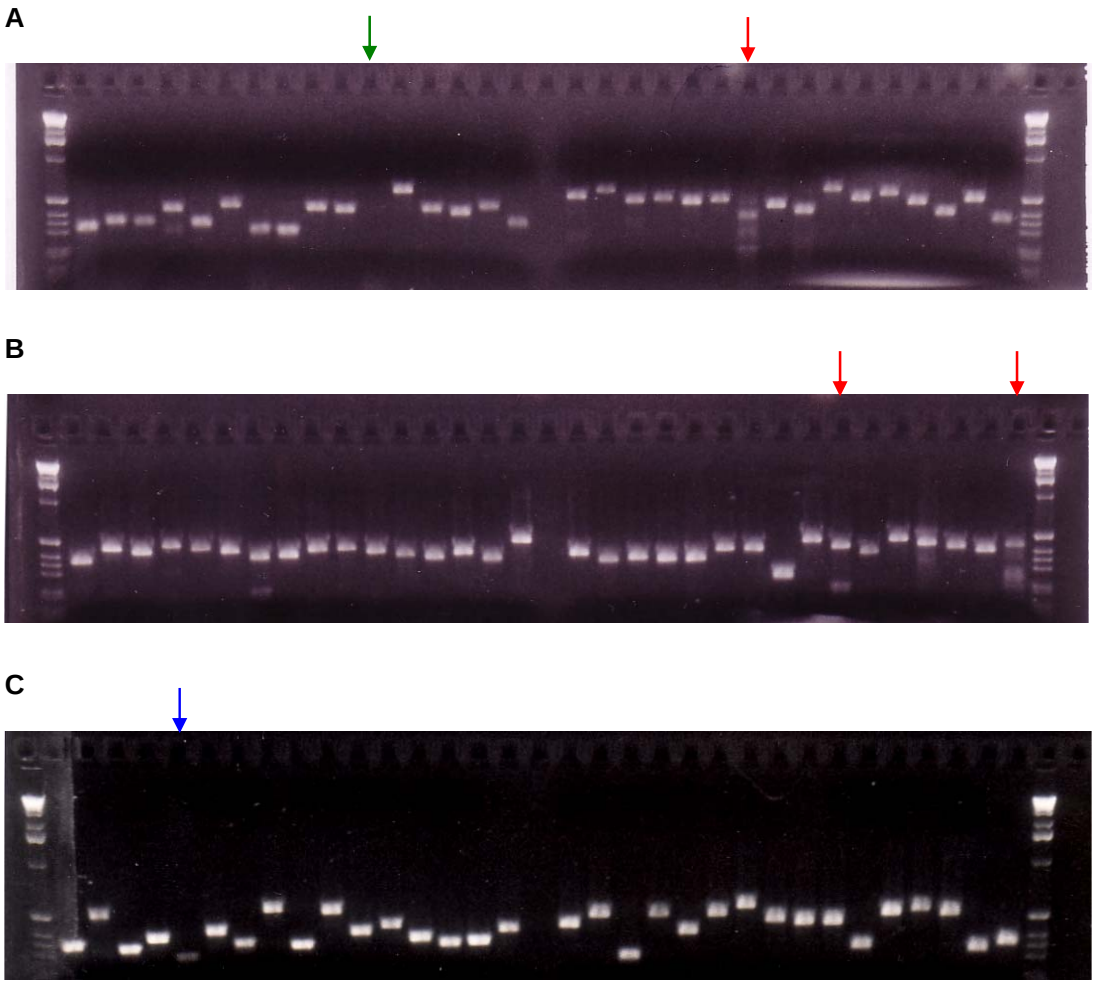


Figure 3.3: Images of agarose gels after electrophoresis of second round PCR amplicons of SCL tiling path array. Panels A and B show PCR amplicons for human and panel C shows PCR amplicons for

mouse. The entire sets of human and mouse array elements were electrophoresed on agarose gels and their bands were scored visually. Lane marked with green arrow highlights a missing band, red arrows highlight multiple bands and the blue arrow highlights a weak band. Products were electrophoresed on 2.5% agarose 1 X TBE gels and visualised with ethidium bromide.

Array elements which did not yield clearly visible amplification products of the correct size were scored as (i) missing bands, (ii) multiple bands, (iii) weak bands or (iv) PCR products of the wrong size (see Figure 3.3, panels A, B and C for examples). The results of this analysis for the human and mouse array amplicons generated for the 1st generation and final arrays are summarized in Table 3.1.

3.4.2. Sequence analysis of the SCL amplicons

Human and mouse SCL array amplicons were subjected to a BLASTN *in silico* sequence analysis (Altschul et al. 1997) to determine whether the amplicons had sequence homology with other sequences in their respective genomes (including repeat sequences). These results are summarised in Table 3.1. Furthermore, in order to determine that the human amplicons were actually derived from the expected genomic regions, a significant proportion of these were verified by sequencing. In total, 180 human amplicons represented on the final array were selected at random and submitted for sequencing. The sequences obtained were compared against the expected sequences; sequences showing more than ninety percent identity with the expected sequences were considered as good matches. The results of the sequence analysis of human amplicons are summarized in Table 3.1.

3.4.3. Validation of the human array elements in genomic array assays

The performance of the human array elements was assessed in genomic microarray assays to determine whether each element reported quantitative measures of genomic copy number. Although the tiling arrays were generated for both human and mouse SCL loci, only the human array elements were tested for their performance in the validation assay; the human SCL arrays serve the basis for the majority of the work described in this thesis. The hybridisation conditions were optimised and standardized for the human array elements so as to obtain accurate copy number measurements. A series of hybridisations were performed on both 1st generation and final SCL tiling arrays using male (XY) versus female (XX) genomic DNA comparisons.

3.4.3.1 Validation of the 1st generation SCL genomic tiling path array

At least six male/female DNA hybridisations were performed on the SCL 1st generation genomic tiling path array. Genomic DNA extracted from normal human male (HRC 575) and normal female (HRC 160) lymphoblastoid cell lines (see chapter 2), was differentially

labeled with Cy5 and Cy3 respectively and hybridised onto the array. Mean Cy5:Cy3 ratios, standard deviations (SDs) and coefficients of variation (CVs) were calculated for each array element spotted in quadruplicate. Figure 3.4 shows the data set obtained from one representative male/female DNA hybridisation.

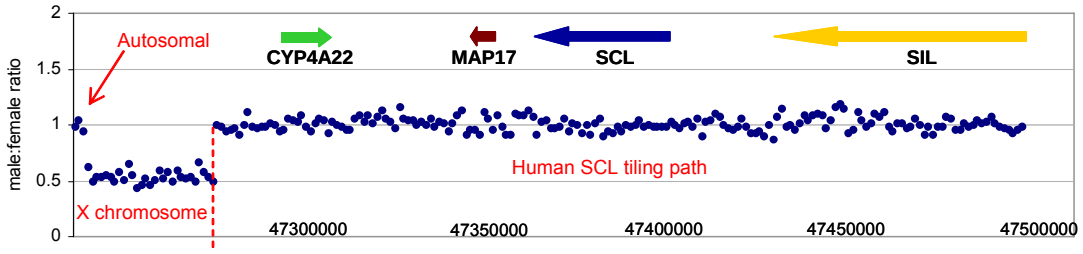


Figure 3.4: A histogram plot of a male versus female hybridisation on the 1st generation human SCL tiling path array. The X chromosome, autosomal and SCL array elements are labelled in red. A representation of the genomic coordinates across the SCL locus is shown along the x-axis, at the bottom of the plotted histogram. The gene order and the direction of transcription are shown at the top of histogram above the SCL array elements. The dotted red line represents a demarcation of the X chromosome array elements from the SCL array elements. The mean male to female ratios for each array element are represented along the y-axis.

The expected male:female ratio for all the human X chromosome array elements is 0.5. The human X chromosome array elements performed consistently and reported accurate copy-number measurements with a mean ratio of 0.56 across the six male/female hybridisations. In the six hybridisations, an average of 96% (28 out of 29) of the X chromosome array elements reported accurate copy number values for analysis and the mean standard deviation (SD) was 0.06 (mean CV = 11%). The three control autosomal array elements (excluding the SCL array elements) reported a mean ratio of 1.02 across the six hybridisations with a mean SD of 0.07 (mean CV = 7%). The SCL array elements collectively reported a mean ratio of 1.01 across the six hybridisations carried out to test their performance and the mean SD was 0.06 (6%). These results demonstrated that the array assays could report accurate copy number changes and that the human SCL array elements were able to report accurate copy number measurements in a genomic array assay. These results were consistent with the assessment of the single-stranded array platform reported previously (Dhami et al. 2005).

3.4.3.2 Validation of the final SCL genomic tiling path array

Three human male/female DNA hybridisations were performed using normal human male and female genomic DNAs. The data analysed for each experiment only included the clone set, as only the clone set of amplicons (i.e., those derived from BAC and PAC templates) included the complete set of array elements representing the human SCL

locus in the final array. In total, 367 human SCL array elements (i.e., the total number of "passed" array elements determined from the analyses summarized in Table 3.1) were included in the final data set. Mean ratios, SDs and CVs were calculated for each array element spotted in duplicate.

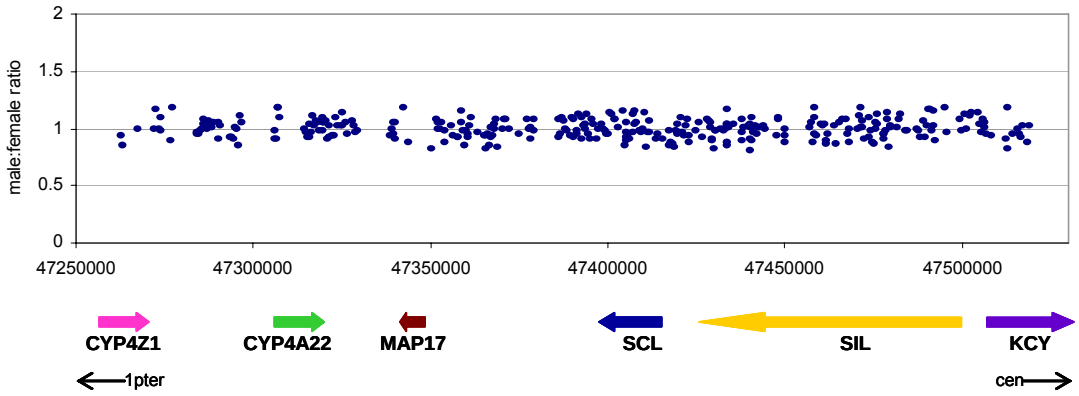


Figure 3.5: A histogram plot of a male versus female hybridisation on the final human SCL tiling path array. The array elements representing only the clone set, i.e. the set which was amplified using DNA derived from bacterial clones, was used for the final analysis as this set included the complete set of array elements representing the human SCL locus. The array elements reporting deviant ratios are not shown on the plot (see text). The array elements are plotted based on their genomic positions and x-axis represents the genomic coordinates along human chromosome 1. The y-axis represents male to female ratio for each array element on the array. The thick coloured arrows represent the gene order and the direction of transcription as shown at the bottom of the histogram plot.

Figure 3.5 shows the final SCL data set of mean ratios for each array element from one representative male/female DNA hybridisation. The mean ratio of the human SCL array elements was 0.99 (the expected ratio was 1) with a mean SD of 0.08 (mean CV = 8%). The final data set reported ratios within 0.2 copy-number units of the theoretical value of 1. Seven array elements (including HSSCL/M93B, HSSCL/M93B3, HSSCL/M93B5, HSSIL/M18B, HSTAL.89, HSTAL.226 and HSTAL.227) consistently reported deviant ratios (either below 0.8 or above 1.2) in the three hybridisation experiments and were excluded from further analyses. Furthermore, no strong correlation was found between the unexpected ratios reported by these elements and their G+C content, length, or repeat content. A lack of correlation between the performance of array elements and sequence features has been reported previously in the detection of copy-number changes at the level of individual exons in the human genome (Dhami et al. 2005). The array elements representing the GATA-1 control amplicons (Hb/BG09, Hb/BG032-1, and Hb/BG032-2) reported mean ratios of 0.91 with mean SD of 0.11 (mean CV = 12%).

Based on the results from all the above described validation experiments, 360 and 411 array elements representing the human and mouse SCL loci respectively were included

in all the array analyses described in this thesis (see Appendix 1 and 2 for the complete lists in human and mouse respectively).

3.5. Characterization of human SCL expressing and non-expressing cell lines

As stated in section 3.1, numerous human and mouse haematopoietic cell lines have been used to understand the regulation of SCL (Bernard et al. 1992; Leroy-Viard et al. 1994; Fordham et al. 1999; Gottgens et al. 1997; Gottgens et al. 2002). Four human haematopoietic cell lines, K562, Jurkat, HL60 and HPB-ALL (also see section 3.1) and two mouse cell lines 416B and E14 ES cell line were selected to perform ChIP-chip experiments across the SCL locus. Between them, these cell lines cover a range of SCL expression patterns from normal (K562, 416B cell lines), inappropriate (Jurkat cell line) to no expression (HL60, HPB-ALL, mouse E14 ES cell lines) and provided excellent experimental systems to understand the transcriptional regulation of SCL. The human cell lines are described in section 3.1 of this chapter. The mouse 416B cell line is a bipotential cell line with majority of the cells identified as undifferentiated blast cells (Dexter et al. 1979). About 95% of the cells possess a diploid chromosome complement, while the remaining cells being true tetraploid or octaploid cells (Dexter et al. 1979). The mouse E14 ES cell line was derived from strain 129/Ola mouse blastocysts and the cells have the potential to differentiate into multiple cell types (Hooper et al. 1987).

Since the four human haematopoietic cell lines were originally derived from leukaemic patients, detailed characterizations using G-banding, M-FISH and conventional CGH analysis on metaphase chromosomes had reported numerous structural and numerical genomic imbalances in these cell lines (see Table 3.2). From all of the reported studies so far, it is evident that different versions of the same cell lines can exist with varied cytogenetic and cellular characteristics. Furthermore, it is widely believed that *in vitro* culturing of cell lines can also introduce additional chromosomal rearrangements, changes in gene expression patterns and changes in DNA methylation (Drexler et al. 2000). To establish the level of SCL expression, extent of genomic imbalances and any chromosomal rearrangements that may affect the SCL locus, the four human cell lines were fully evaluated using four different approaches as detailed below.

Cell-line	Cell-type	SCL expression	Cytogenetic characteristics	Chromosomes involved in genomic imbalances	Previously unreported
K562[a,b]	myelogenous	Yes[f]	Diploid (46, XX) Hyperdiploid (50-52, XX) Near-triploid (69-73, XX) or (62-69, XX) Near-tetraploid (90-96, XX)	1, 3, 5, 6, 7, 9, 10, 13, 14, 17, 18, 20, 21, 22	4, 8, 12, 15, 16
Jurkat[c,d]	T-cell	Yes[f]	Diploid (45-48, XY) Hypotetraploid	2, 3, 4, 5, 8, 9, 18, 20, X, Y	10, 15, 22
HL60[d]	Promyelocytic	No[f]	Diploid (44-47, XX)	5, 8, 9, 11, 13, 14, 15, 18	4, 6, 7, 10, 16,17, 22
HPB-ALL[e]	T-cell	No[f]	Diploid (46, XY) Tetraploid (94, XY)	1, 2, 3, 5, 14, 16, 21	7, 8, 10, 17

Table 3.2: Cytogenetic characteristics and SCL expression patterns in the cell lines: K562, Jurkat, HL60 and HPB-ALL. Note: The cytogenetic and expression data reported in this table has been compiled from other published studies: [a] Gribble et al. 2000; [b] Naumann et al. 2001; [c] Snow et al. 1987; [d] Cottier et al. 2004; [e] MacLeod et al. 2003; [f] Delabesse et al. 2005. The last column (previously unreported) shows the results obtained with 1 Mb array-CGH performed in the study presented for this thesis.

3.5.1. Real-time PCR analysis of the SCL expression in the four cell lines

To investigate the levels of SCL expression in the four cell lines used in this study, quantitative real-time PCR (RT-PCR) using SYBR Green was performed using a 7000 sequence detection system (Applied Biosystems). The SIL gene (upstream of the SCL gene) is ubiquitously expressed in all four cell lines and thus represented a positive control in the experiment. Primer pairs mapping to the 3'-end of the SCL (within exon 6), SIL and β-actin transcripts were designed and are listed in Appendix 3B. The expression level of β-actin was used as an endogenous reference gene against which SCL and SIL expression levels were normalised.

Figure 3.6 illustrates the relative expression levels of the SCL and SIL genes in K562, HL60, HPB-ALL, and Jurkat. The reported results were as expected with the previously published results; SCL gene was expressed in K562 and Jurkat but was not expressed in HL60 and HPB-ALL. As expected, SIL expression was observed in all of the cell lines.

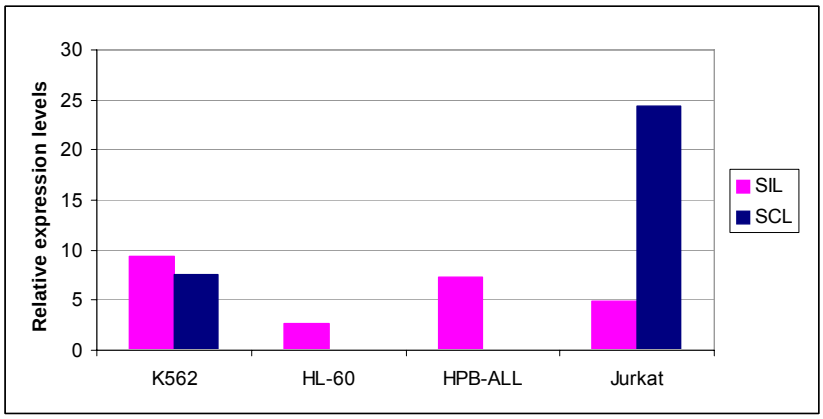


Figure 3.6: Analysis of SCL expression by real-time PCR. Relative levels of SCL and SIL expression in the four cell lines as measured by quantitative real-time PCR. SIL gene is ubiquitously expressed in all cell lines and served as a positive control. The expression level of β-actin was used as an endogenous reference gene against which SCL and SIL expression levels were normalised. The cell lines are represented at the bottom of the plot and the y-axis represents the relative levels of SIL and SCL transcripts.

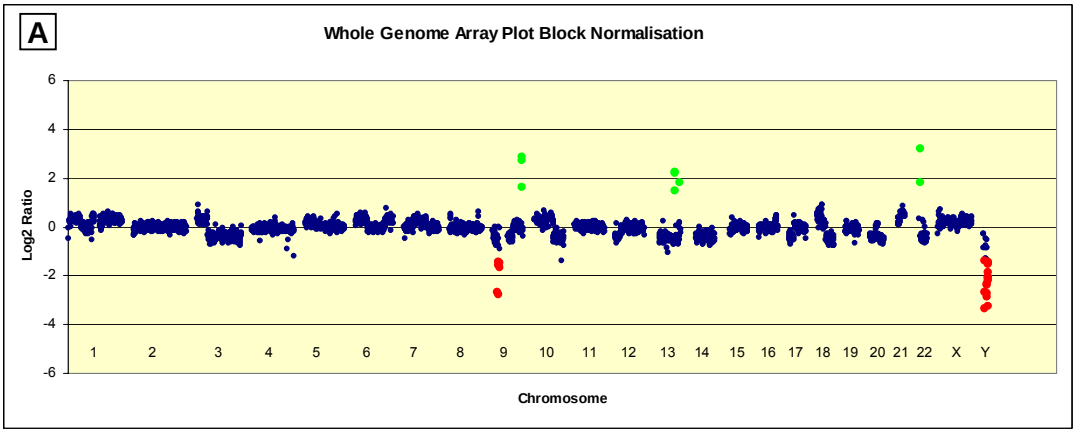
3.5.2. 1 Mb genomic microarray-CGH analysis of K562, Jurkat, HL60 and HPB-ALL

Array CGH hybridisations were performed on a genomic microarray that contains 3500 large insert BAC and PAC clones containing sequences from across the human genome (Fiegler et al. 2003). The purpose of the array-CGH analysis was to identify genomic imbalances in the four cell lines, in addition to the ones previously reported (Table 3.2), and to detect any rearrangements affecting the SCL locus (on chromosome 1) in particular. Since the resolution of the genomic array is one clone per 1 Mb of the human genome, none of the clones represented on the array span the genomic region containing the SCL locus. Two clones, RP11-8J9 at 46.9 Mb and RP11-330M19 at 48.0 Mb positions, map to either sides of the genomic region containing the SCL locus on chromosome 1. A male pool DNA sample that was used as the reference DNA in the array hybridisations was derived from combining peripheral blood DNA from 20 normal males.

The arrays were analysed using a Microsoft Excel spreadsheet specially written for the 1 Mb array. The ratios (median intensity – background) of Cy3:Cy5 intensities for each spot were normalised to the median raw ratio of all the autosomal spots on the array (global normalization). The normalised ratios for each spot in a sub-grid (block) of the array were further normalised to the median ratio of all the spots in that block (block normalization). After the global and block normalizations, the mean ratios, standard deviations and CVs were calculated for duplicate clones. Duplicates reporting difference of more than 20% were excluded from the data set. The mean linear ratios of the accepted clones were converted into $\log_2$ ratios and plotted on a histogram against the

genomic position of the clone in the genome according to build “35” NCBI. In order to detect genomic copy number changes such as gains and deletions, a “significant cut-off” value calculated for individual hybridisations was used. The significant cut-off value was set at five times the calculated standard deviation of 95% ratios in each experiment. This means that any clone reporting a ratio greater or less five times the standard deviation of the majority of the clones was considered significant to indicate copy number gain or deletion respectively. Known polymorphisms in the human genome were also flagged. The results obtained in all the four cell lines are described below.

1. The K562 cell line is known to exhibit multiple/complex genomic imbalances involving various chromosomes with copy number gains (including high-level amplifications), deletions and translocations (Table 3.2). Array-CGH analysis reported imbalances in almost all of the chromosomes (Figure 3.7, A). Copy number gains (including some amplifications) and losses (deletions) were observed in chromosomes 1, 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 14, 15, 16, 17, 18, 20, 21 and 22. From this, it was apparent that the K562 cell line used in this study displayed additional genomic rearrangements than those previously reported. Genomic imbalances had not been previously reported in chromosomes 4, 8, 12, 15 and 16 (Table 3.2).



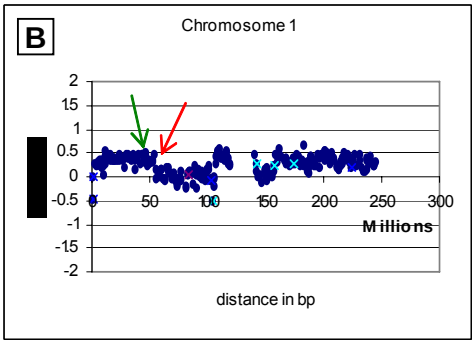
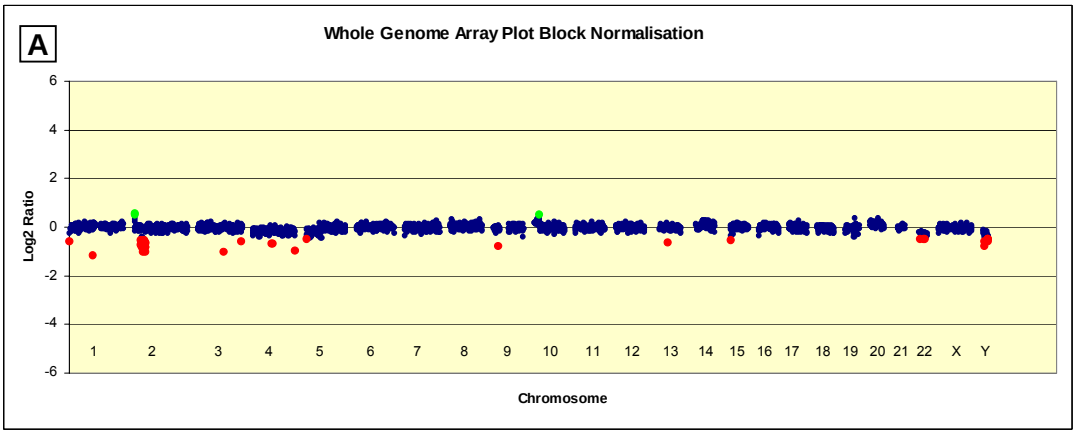


Figure 3.7: Array-CGH analysis of K562 cell line using the 1 Mb genomic array. Panels A and B show the genome plot and the chromosome 1 plot respectively. The red and green dots in the panel A represent some of the clones reporting deletions and copy number gains respectively. The green arrow in the chromosome 1 plot points to the position of the SCL gene and the red arrow points to the chromosomal breakpoint on 1p. The distal arm of 1p containing the SCL locus shows a copy number gain from 12.4 Mb to 53.8 Mb. The clones excluded (based on analysis criteria) in panel B are shown with Xs.

It had been reported previously using FISH analysis that a translocation breakpoint on chromosome 1p (p32-pter) in K562 did not structurally disrupt the SCL locus (Gribble et al. 2000). The array-CGH analysis reported a copy number gain of chromosome 1p from 12.4 Mb to 53.8 Mb which has not been previously reported by conventional CGH analysis on metaphases. This genomic region on the distal arm of chromosome 1p contained the SCL locus (Figure 3.7, B), which did not appear to be structurally disrupted. Clone RP11-243A18 at 53.8 Mb reported a $\log_2$ ratio of 0.44 whereas the clone RP5-1070D5 at 55.6 Mb reported a $\log_2$ ratio of 0.14, suggesting that this may represent a chromosomal breakpoint associated with the gained region. A gain of 1q had previously been reported (Gribble et al. 2000) which was confirmed here by array-CGH.



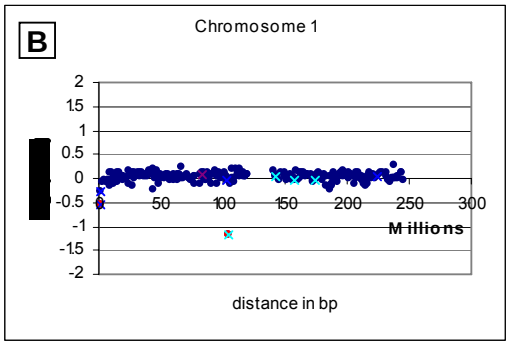


Figure 3.8: Array-CGH analysis of Jurkat using the 1 Mb genomic array. Panels A and B show the genome plot and the chromosome 1 plot respectively. The red and green dots in the panel A represent clones reporting deletions and copy number gains respectively. The chromosomes are listed at the bottom of panel A. No genomic imbalances were reported on chromosome 1p containing the SCL locus. The clones excluded (based on analysis criteria) in panel B are shown with Xs.

2. Array-CGH analysis of the Jurkat cell line determined that various chromosomes were involved in genomic imbalances including some that had previously not been reported (Table 3.2; Figure 3.8, A). A number of clones reported deletions and copy number gains on chromosome 2, deletions on chromosomes 3, 4, 5, 9, 15, 22, and Y, and copy number gains on chromosomes 8, 10, and 21. However, no copy number gains or deletions were seen on chromosome 1 (Figure 3.8, B).

3. Array-CGH analysis of the cell line HL60 did not only detect all the genomic imbalances that had been reported previously, but also detected additional copy number gains and deletions in various chromosomes (Table 3.2; Figure 3.9, A). A number of clones reported significant deletions on regions of chromosomes 4, 5, 9, 10, 14, and 17. Several clones reported copy number gains of regions of chromosomes 6, 7, 8, 13, 18 and 22 including a high-level amplification on the long arm of chromosome 8. Several clones reported significant copy number gains on the short arm of chromosome 16 whereas a number of clones representing the long arm of chromosome 16 reported significant deletions. Clones representing chromosome 1 did not report any significant ratios indicative of copy number gains and/or deletions (Figure 3.9, B).

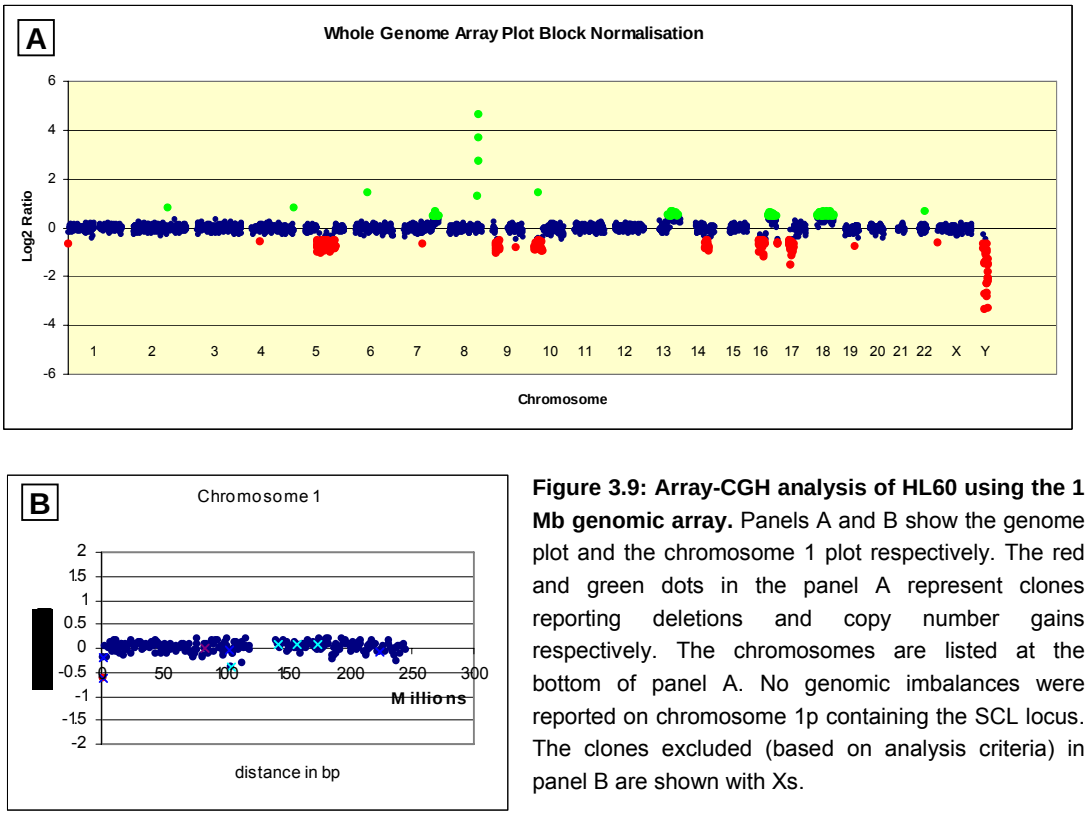
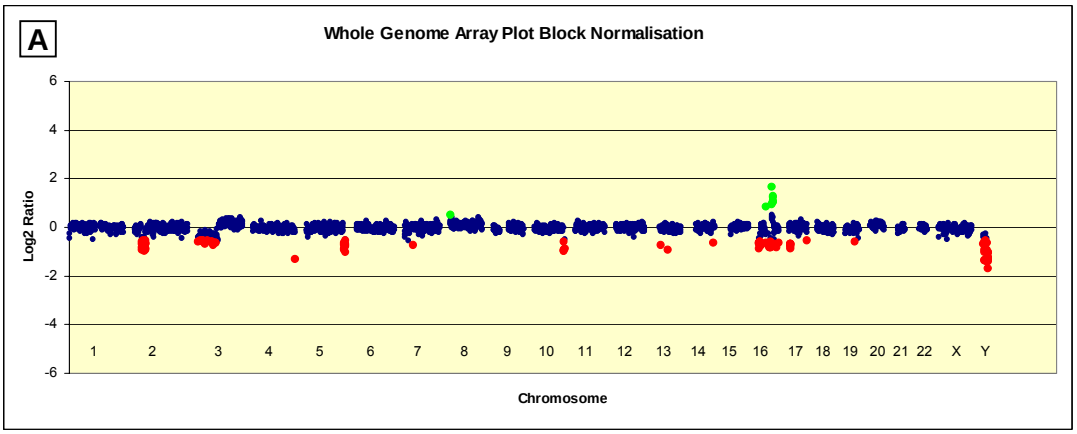


Figure 3.9: Array-CGH analysis of HL60 using the 1 Mb genomic array. Panels A and B show the genome plot and the chromosome 1 plot respectively. The red and green dots in the panel A represent clones reporting deletions and copy number gains respectively. The chromosomes are listed at the bottom of panel A. No genomic imbalances were reported on chromosome 1p containing the SCL locus. The clones excluded (based on analysis criteria) in panel B are shown with Xs.

4. Several chromosomes including chromosome 1 had been reported to be involved in chromosomal translocations and genomic copy number changes in the cell line HPB-ALL (Table 3.2). Array-CGH analysis of HPB-ALL identified additional chromosomal imbalances (Table 3.2; Figure 3.10, A). A number of clones on the chromosomes 2, 3, 5, 7, 10, and 17 reported low ratios indicating deletions whereas one clone on chromosome 8 reported a high ratio, indicating a gain. Several clones on chromosome 16 reported

high levels of amplifications while a number of other clones reported deletions. A clone on chromosome 4 reporting a low ratio was known to be polymorphic.



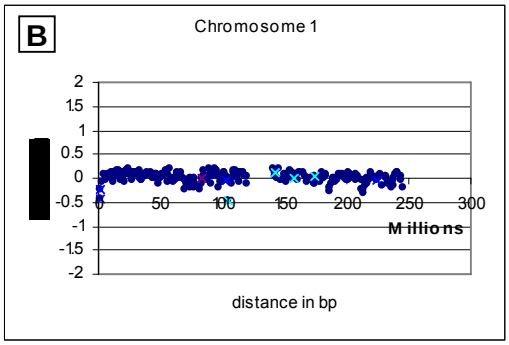


Figure 3.10: Array-CGH analysis of HPB-ALL using the 1 Mb genomic array. A and B represent the genome plot and chromosome 1 plot respectively. The red and green dots in the panel A represent clones reporting deletions and copy number gains respectively. The chromosomes are listed at the bottom of panel A. No genomic imbalances were reported on chromosome 1p containing the SCL locus. The clones excluded (based on analysis criteria) in panel B are shown with Xs.

The involvement of chromosome 1 in a genomic translocation had been reported previously in HPB-ALL (MacLeod et al. 2003). Any chromosomal breakpoint was not obvious with the array-CGH analysis (Figure 3.10, B). It is possible that the translocation could be a balanced rearrangement which would not be detected by array-CGH. However, none of the clones representing chromosome 1 reported any significant deviation from the normal ratios suggesting no obvious copy number gains or deletions were present on chromosome 1.

3.5.3. Array-CGH analysis of K562, Jurkat, HL60 and HPB-ALL using the SCL genomic tiling path array

To investigate the SCL genomic region at a higher resolution (approximately 400-500 bp), array-CGH experiments were performed using the final SCL genomic tiling path array. Hybridisation experiments with DNA samples extracted from the four cell lines against a reference DNA on the tiling path array would report any small gains or losses in the SCL genomic region which could not be detected by array-CGH analysis using the 1 Mb BAC array.

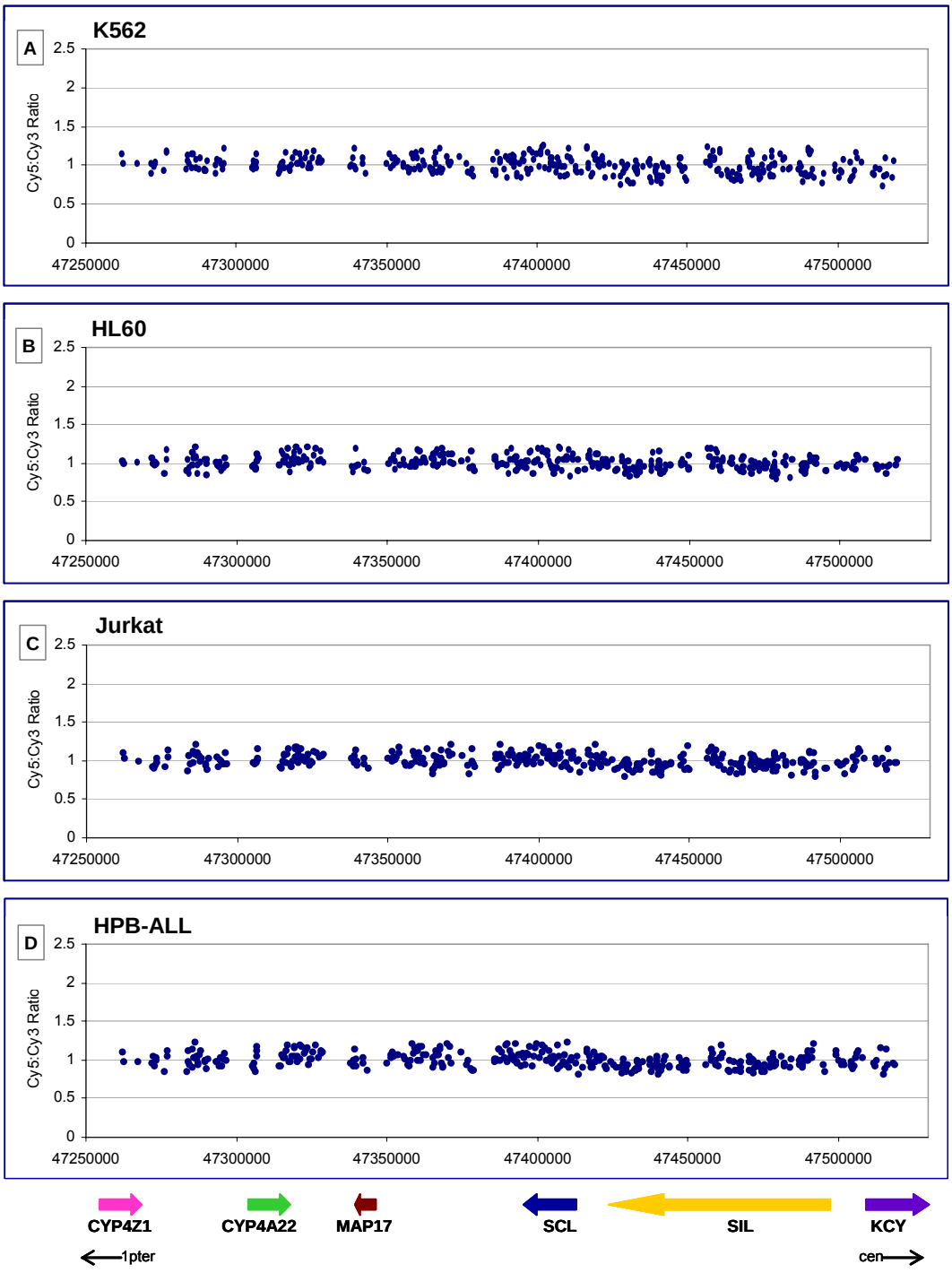


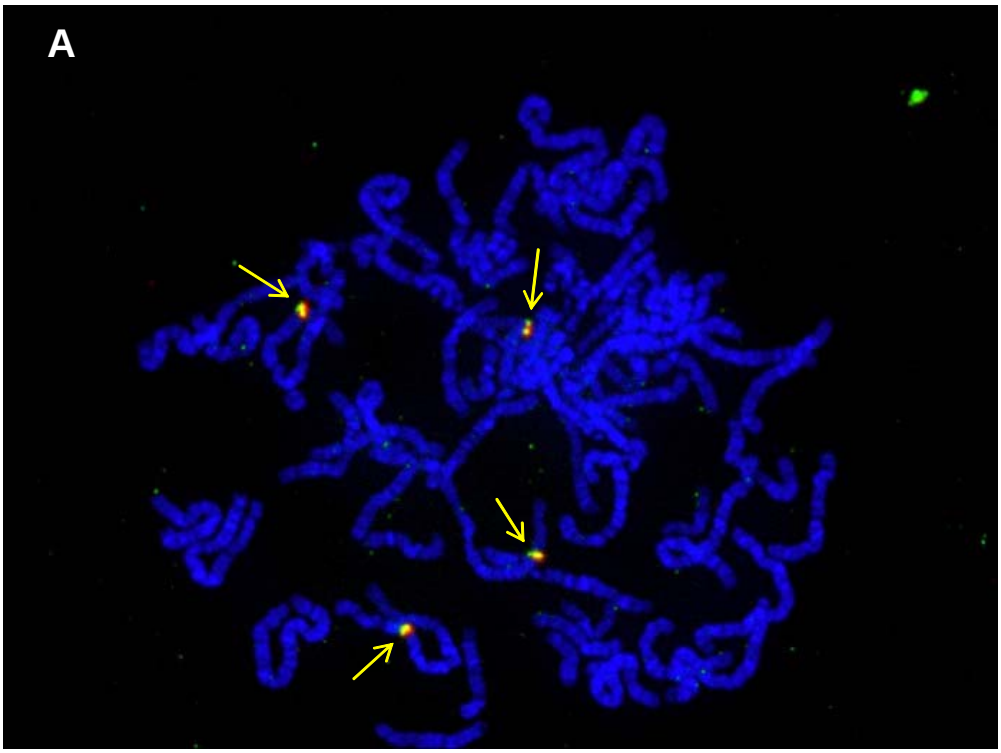
Figure 3.11: Array-CGH analysis of the four cell lines using the high resolution SCL genomic tiling path array. Panels A-D show histogram plots of K562, HL60, Jurkat and HPB-ALL respectively. The Cy5:Cy3 ratios are plotted on the y-axes against genomic position of the array elements along chromosome 1 on the x-axes. Mean SDs calculated for each data set were 0.10 for the K562, 0.08 for the HL60, 0.08 for the Jurkat and 0.09 for HPB-ALL. The thick coloured arrows represent the gene order and the direction of transcription as shown at the bottom of the figure. The orientation of the locus with respect to the centromere (cen) and telomere (ter) on human chromosome 1 is shown by the thin black arrows at the bottom of the figure.

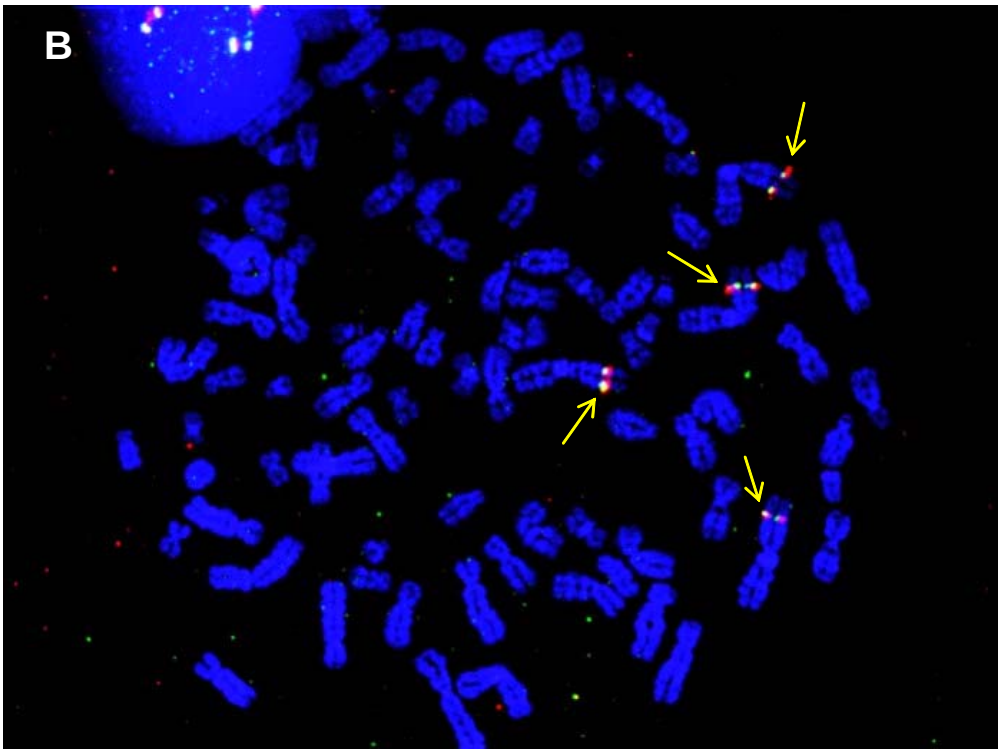
Figure 3.11 shows the histogram plots for high resolution array-CGH analysis across the SCL region for all the four cell lines. The expected Cy5:Cy3 ratio for all the array elements on the SCL tiling path array was 1, as all the array elements correspond to autosomal sequences. No changes in genomic copy number which affected the SCL region were reported by the SCL array elements in any of the four cell lines. This suggested that no genomic imbalances such as copy number gains or deletions lay within the 250 kb genomic region containing the SCL gene which could structurally disrupt the SCL locus, thus affecting its regulation. Although it was known that the cell line K562 had a copy number gain of chromosome 1 encompassing the SCL locus (see Figure 3.7, B), the high resolution SCL array did not detect it, since all of the array elements on the array were affected by this copy number gain in the same way. In other words, copy number differences can only be reported if some of the elements report the normal (modal) values.

3.5.4. FISH analysis of K562, Jurkat, HL60 and HPB-ALL with two DNA clones spanning the human SCL locus

To investigate the possibility of a balanced rearrangement affecting the SCL locus, FISH analysis was performed using two overlapping bacterial clones spanning the SCL locus in human. The clones RP1-18D14 and RP11-332M15 were used (section 3.3.3) for this analysis since these clones contained the genomic region represented on the SCL genomic tile path array. The DNA from both clones was differentially labelled by nick-translation, mixed and hybridised to metaphase chromosome spreads from each cell line (see chapter 2 for methods). After detection, slides were scanned and at least ten metaphase spreads were analysed per cell line. Any translocations affecting the SCL locus would be visible by the appearance of FISH signals for either clone which did not map to the same chromosomal location. The FISH analysis of K562, Jurkat, HL60 and HPB-ALL is shown in Figure 3.12. The SCL locus was not structural disrupted in any of the four cell lines. All four cell lines were mostly polyploid as described in the legend of Figure 3.12.

.


A


B

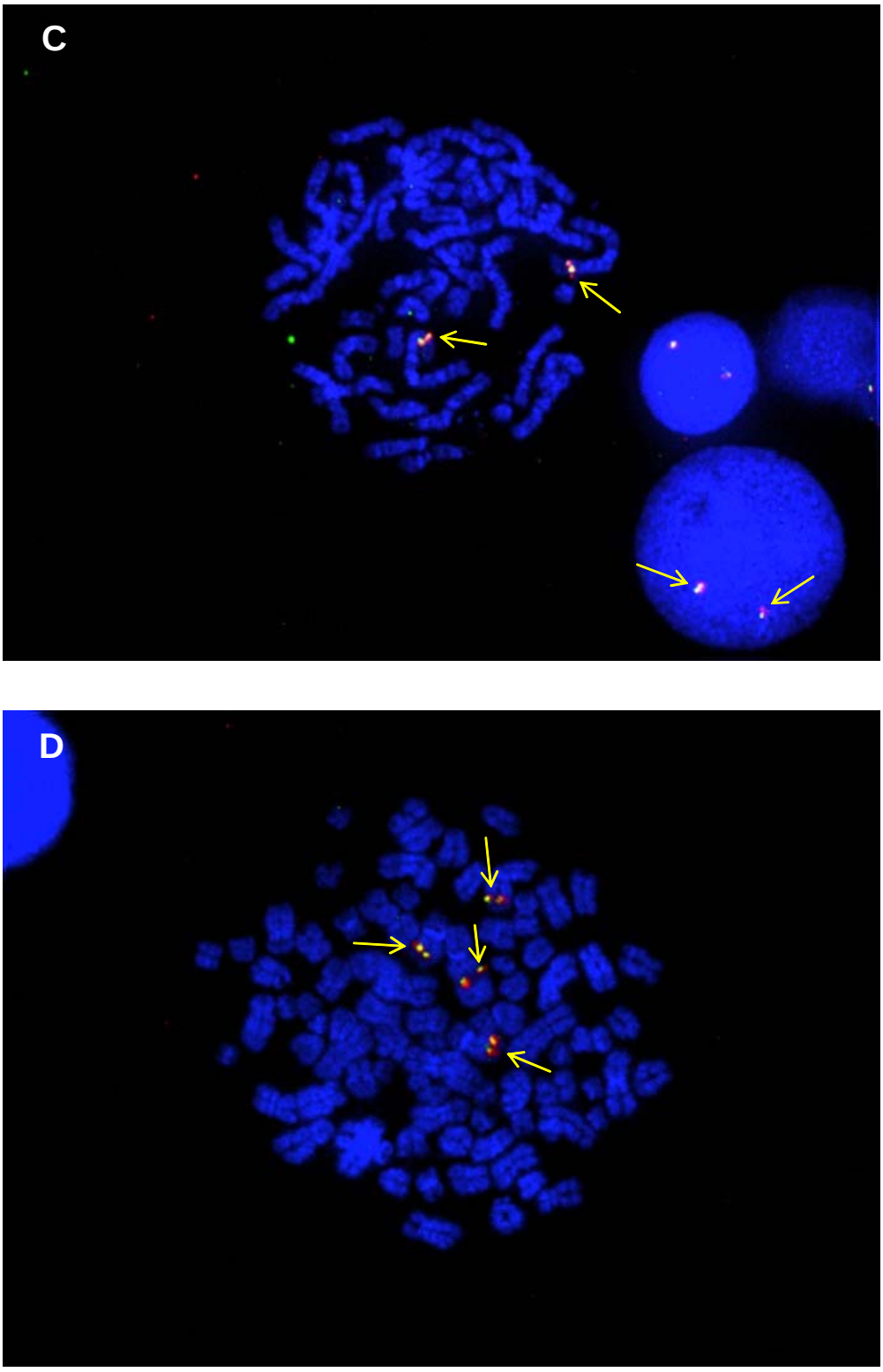


Figure 3.12: FISH images of metaphase spreads from the four human cell lines. A: K562, B: Jurkat, C: HL60 and D: HPB-ALL. FISH analysis was performed by hybridising two clones (RP1-18D14 and RP11-332M15) spanning the SCL locus to metaphase chromosomes prepared from each of the cell lines. Yellow arrows in each image point to the co-localized FISH signals on chromosome 1p. The red signals represent clone RP1-18D14 and the green signals represent clone RP11-332M15. The K562 cell line was observed to

be near-triploid with chromosome number ranging from 62 to 75. In K562, the majority of the cells had four bright co-localized signals overlapping each other on the short arm of chromosome 1 (panel A), while other cells showed two bright signals. The presence of four co-localized signals indicated that the genomic region spanned by the two clones i.e. 1p33 was present in four copies and was not structurally disrupted anywhere in the genome of K562 cell line. 1 Mb array-CGH analysis reported amplification of 1p32.3-1p36.33 region containing the SCL locus (at 1p33) which was also confirmed by FISH analysis. The presence of four co-localized FISH signals in a near-triploid cell line suggested that an extra copy of the chromosome 1p33 region existed in K562. At least, two normal chromosomes 1 could be identified in the majority of the cells; the others were possibly marker chromosomes with genomic material from the short arm of chromosome 1. Similarly, in all the other cell lines, Jurkat, HL60, and HPB-ALL, the signals were always co-localized indicating no disruption in the SCL region. The majority of the cells scanned in Jurkat had four co-localized signals and the cell line was observed to be tetraploid with chromosome number 90 to 92 (panel B). HL60, on the other hand, was observed to be diploid with two co-localized signals per cell (panel C). Four co-localized FISH signals were seen in majority of the cells scanned in HPB-ALL with chromosome number ranging from 80 to 85 making it hypo-tetraploid (panel D).

3.6. Discussion

3.6.1. SCL genomic tiling path array: sensitivity and resolution issues

Genomic microarrays are increasingly being used to unravel and understand aspects of genome biology. Depending on how genomic arrays are used in various applications in various species, the requirements for sensitivity can vary. For example, genomic arrays used for array-CGH are required to report accurate quantitative measurements of single copy number changes; this is even more challenging within highly complex genomes, such as the human genome. The quantification and sensitivity requirements for array-CGH were used as benchmarks to develop the highly sensitive SCL genomic arrays described in this chapter. This work was preceded by the development of an array-CGH platform which had the ability to detect single copy number change at the resolution of a single exon (139-571 bp) in the human genome (Dhami et al. 2005). The development of this “exon” array-CGH platform was shown to be two orders of magnitude more sensitive than all other array-CGH technologies in terms of its ability to detect copy number changes at high resolution in the human genome. Thus, the single-stranded array chemistry which was used to develop the “exon” array-CGH platform was used to construct the SCL arrays presented in this chapter. As a result, the SCL array platform has been shown to be sensitive enough to accurately measure genomic copy number changes within the human genome.

The SCL tiling array platform would be ideal to use in any array-based assay in which accurate quantitative measurements of DNA is required, including mapping DNA-protein interactions at the SCL locus. In mammalian systems, many published studies have made use of genomic tiling arrays to map DNA-protein interactions across genomic region or even genome-wide scale (Horak et al. 2003; Cawley et al. 2004; Martone et al.

2003; Euskirchen et al. 2004; Kim et al. 2005). However, none of the published ChIP-chip studies have tested the ability of array platforms to perform accurately within the dynamic range required for array-CGH. Yet, the ability to report accurate quantitative measurements is highly desirable for ChIP-chip platforms as well - using an array platform which could perform to this level would prove to be a great asset in measuring subtle ChIP enrichments or low-affinity binding sites of transcription factors. Subsequent work in this thesis will demonstrate how this level of sensitivity is an absolute requirement for ChIP-chip studies, in order to detect the full range of regulatory interactions at the SCL locus.

In the published studies mentioned above, genomic tiling arrays were constructed by using either tiled PCR products (Horak et al. 2002; Schübeler et al. 2004; Martone et al. 2003; Euskirchen et al. 2004) or tiled oligonucleotides (Cawley et al. 2004; Pokholok et al. 2005; Kim et al. 2005). The present study used PCR products to construct the SCL arrays. Although using long or short oligonucleotides appears to be an attractive possibility to increase the resolution of mapping DNA-protein interactions, it is notable that oligonucleotides for use in ChIP-chip studies have two potential limitations. Firstly, it is not only the array elements which confer the resolution of the platform, but also the size of the sheared chromatin fragments arising from ChIP which are used in array hybridisations. The average size of these sheared DNA samples is usually in the range of 300-1000 bp, thus lowering the effective array resolution substantially when hybridised onto oligonucleotides. Secondly, analysis of oligonucleotide arrays relies on averaging measurements taken from multiple array elements, which effectively decreases their resolution to approximately 500 bp. Given these issues, it has not yet been empirically determined which type of array platform is better for ChIP-chip studies given that a direct and comprehensive comparison of PCR-spotted arrays or oligonucleotide arrays for ChIP-chip experiments has not yet been published. Furthermore, the effective resolution of oligonucleotide ChIP-chip arrays described above is similar to that for the SCL arrays described in this chapter, suggesting that oligonucleotide arrays would not have conferred a resolution advantage for the work reported later in this thesis.

It should also be noted that the validation experiments of the human SCL array described in this chapter, represent an initial characterization of this platform, but are by no means an exhaustive one. Additional work on the validation of the array for its use in ChIP-chip assays is described in chapter 4 of this thesis. Furthermore, although an SCL tiling path array was also constructed across the mouse SCL locus, the mouse array elements were not assessed for their performance in array-CGH experiments in the same manner as the human array elements. However, the array elements representing

the mouse SCL locus were assessed for their performance in ChIP-chip based assays for histone H3 acetylation as described in chapter 4.

3.6.2. Characterizing human haematopoietic cell lines

All of the human haematopoietic cell lines selected for the study presented for this thesis were derived from leukaemic patients; previous characterizations have reported the existence of numerous genomic imbalances in these cell lines (Gribble et al. 2000; Naumann et al. 2001; Snow et al. 1987; Cottier et al. 2004; Macleod et al. 2003). Performing array CGH analysis using the 1 Mb genomic microarray (Fiegler et al. 2003), identified the previously reported genomic imbalances along with the detection of additional copy number gains and deletions in various chromosomes. The detection of previously unidentified genomic imbalances in these cell lines using array-CGH suggests that (i) these imbalances were undetectable with previous low resolution characterization of these cell lines using CGH on metaphase chromosomes, (ii) the cell lines used in the present study represent outgrowth of cryptic sub-clones already present when these cell lines were originally established (Drexler et al. 2003), or (iii) additional genomic imbalances have occurred during the many years since these cell lines were originally established.

Data presented in this chapter show that the SCL region represented on the human tiling path array is structurally intact in the four human cell lines used for the work of this thesis. Of these four cell lines, only K562 showed chromosomal imbalances on chromosome 1. In a previously published study, a translocation breakpoint was detected on chromosome 1p in K562 which did not structurally affect the SCL gene (Gribble et al. 2000). FISH analysis described here using two overlapping bacterial clones confirmed that the SCL locus was not involved in this translocation. Furthermore, array CGH performed at a resolution of 1 Mb in K562 reported a copy number gain of the distal arm of 1p which encompassed the entirety of SCL, suggesting that the SCL locus was structurally intact. Collectively, these results suggest that SCL is likely to be regulated by its own regulatory elements in K562 cell line, although the effect of very distant chromosome 1p genomic imbalances (i.e., further away than the region contained on the SCL tiling array) on SCL expression is not known.

It is possible that genomic imbalances in other parts of the human genome could impinge on the expression of SCL in any of the cell lines studied here. It is known that transcription factors including GATA-1, GATA-2, Elf-1 and Fli-1 play important roles in the regulation of SCL at various stages of development (Aplan et al. 1990; Lecointe et al. 1994; Gottgens et al. 2002; Gottgens et al. 2004). In K562, the genomic regions

containing GATA-2 and Elf-1 on chromosomes 3 and 13 respectively show single copy deletions which could affect the expression of these transcription factors (Figure 3.7, A). Whether these genomic imbalances affect SCL regulation in K562, is not known.

An interesting feature of the human leukaemic cell lines studied here is that they are mostly polyploid in addition to the existence of genomic imbalances of discrete regions of various chromosomes. For instance, four co-localized FISH signals for the SCL locus were seen in most of the cells analyzed in K562 and Jurkat, both of which exhibit SCL expression. This suggests that at least four copies of the SCL gene are present in these cell lines. SCL expression in Jurkat has previously been found to be mono-allelic (Leroy-Viard et al. 1994), suggesting that all of the other SCL alleles in Jurkat are not expressed. Similarly, it is not known whether SCL is transcribed from all of the copies in K562. Thus, the fact that both these cell lines carry extra copies of the gene should be taken into consideration when analysing ChIP-chip experiments performed in these cell lines. It would be difficult to determine, using ChIP-chip experiments, whether the DNA-protein binding interactions are specific to one locus or represent a composite profile of these interactions at all of the SCL loci in each cell.

Despite the presence of all the genomic imbalances, these haematopoietic cell lines provide excellent cell line based experimental systems to decipher the key regulatory interactions involved in SCL regulation (Aplan et al. 1990; Green et al. 1991; Leroy-Viard et al. 1994; Bernard et al. 1992; Cheng et al. 1993; Delabesse et al. 2005) and are routinely used for such studies. The information gained about regulation of SCL from these is indeed invaluable and the major advantage of using cell lines is the virtually unlimited supply of biological material. However, it has been suggested that some additional cellular characteristics are acquired during establishment of cell lines or during extended *in vitro* culturing (Drexler et al. 2003), thereby, highlighting the pitfalls of using cell lines instead of primary cells or tissues in deducing biological information for living organisms. For ChIP-chip assays, the number of cells required to perform a single assay are quite large – in the range of 10 to 20 million cells (Chapter 4 and 5). Therefore, to obtain such large number of primary cells to carry out an extensive study to elucidate DNA-protein interactions across the SCL locus is not presently feasible. Moreover, under optimal cell culture conditions, haematopoietic cell lines can stably retain the major features of their original cells (Drexler et al. 2003), thus, information about the real biological activities of the cell types can be elucidated.

3.6.3. Conclusions

In conclusion, a high resolution genomic tiling path microarray was constructed across the human and mouse SCL loci. The unique amino-link technology employed for its construction provided the array with the ability to be highly sensitive and quantitative. Furthermore, the biological material i.e. human haematopoietic cell lines selected to perform ChIP-chip assays for subsequent parts of the study were characterized in detail to establish their cytogenetic characteristics and to examine the structural integrity of the SCL locus. Used together, these resources provide an excellent experimental system to obtain detailed profiles of DNA-protein interactions across the SCL locus using ChIP-chip assays.