5 ANALYSIS OF 3.5 MB REGION ON 20q12-13.2 USING ChIP-CHIP

Gene regulation is a complex process, requiring a large number of proteins acting cooperatively on specific DNA sequences, which are wrapped with histone proteins. Although instructions for the recruitment of the correct factors onto transcription initiation sites are stored in the DNA sequence, it should be decorated with the correct epigenetic markers to permit functional interactions between the trans acting factors and the nucleosome-bound DNA. Until recently, gene regulation was studied *in vitro* with purified proteins and naked DNA sequences (Kim and Ren, 2006), but such methods cannot mimic *in vivo* the genomic environment of regulatory sequences. Gene reporter assays offer the possibility to investigate regulatory sequences within cells where trans-acting factors are present. However, this approach also lacks the ability to investigate a regulatory sequence in the context of its chromatin environment as well as the effect of possible distal cis-regulatory elements acting on it.

A method developed in 1985 by Solomon et al covalently crosslinks DNA to the proteins bound to it in intact cells using formaldehyde, HCHO, (FA) without damaging the DNA (Solomon and Varshavsky, 1985). FA generates DNA-protein (Varshavsky et al., 1979), RNA-protein (Moller et al., 1977) and protein-protein (Jackson, 1978) crosslinks. In *in vivo* crosslinking, unbound proteins (i.e. not bound to DNA) are not crosslinked to the DNA (Solomon and Varshavsky, 1985). This method is very powerful in understanding the cellular nature of DNA-protein interactions, since it enables to take a snapshot of any given region in the genome together with its associated factors. Solomon et al. used *in vivo* FA crosslinking to provide the first evidence that actively transcribed genes are not free of histone proteins in *D.*

melanogaster (Solomon et al., 1988). Since then, this method rapidly became the method of choice to map proteins onto their genomic targets (Orlando and Paro, 1993) (Gohring and Fackelmayer, 1997) (de Belle et al., 1998) (de Belle et al., 2000) (Wells and Farnham, 2002). Chromatin Immunoprecipitation (ChIP) involves two fundamental steps (i) *in vivo* crosslinking of intact cells, followed by (ii) selective immunoprecipitation of the cross-linked DNA:protein complexes using an antibody directed against the DNA-binding protein of interest (Kuo and Allis, 1999).

In 2000, Ren et al combined *in vivo* FA-crosslinking with DNA microarray technology (ChIP on chip) which emerged as an extremely powerful system to perform large-scale DNA-protein interaction studies (Ren et al., 2000). A number of studies applied this methodology to map common transcription factor binding sites on either a collection of selected sites (Li et al., 2003; Rada-Iglesias et al., 2005) or on a particular genomic region or chromosome (Mao et al., 2003) (Martone et al., 2003) (Horak et al., 2002) or on the entire genome (Kim et al., 2005).

In parallel, *in vivo* FA-crosslinking has been extensively utilized to understand the chromatin context of functional genomic sites. It is well established that histone tails can be covalently modified by acetylation, phosphorylation, sumoylation, ubiquitination and methylation (section 1.4). In 2001, it was shown that in humans 5' ends of transcriptionally active genes are enriched in tri-methylated histone H3 at K4 (H3K4me3) (Litt et al., 2001). Transcriptionally repressed chromatin regions are enriched in tri-methylated histone H3 at K9 (H3K9me3), tri-methylated histone 3 at lysine 27 (H3K27me3) and acetylated histone H4 at lysine 20 (H4K20ac) (Shilatifard, 2006) (Fraga et al., 2005). Moreover, phosphorylated histone H3 at serine 10 was linked to both transcriptional activation and chromosome condensation during mitosis(Cheung et al., 2000), whereas ubiquitinylated histone 2A at lysine 119 and tri-methylated histone 3 at K27 are required for both gene silencing and X-chromosome

inactivation (Cao et al., 2002; Plath et al., 2003). The above are examples common histone modifications involved in gene regulation though there are several histone modifications implicated in other cellular processes (Strahl and Allis, 2000).

In this study, I explored the transcriptional activity on the region of interest using ChIP with an antibody against RNA polymerase II. I then investigated the histone code of a 3.5 Mb sub-region of 20q12-13.2 using *in vivo* FA-crosslinking or chromatin immunoprecipitation (ChIP) with antibodies against seven posttranslationally modified histones. Additionally, I employed ChIP with an antibody against CTCF (CCCTC-binding factor) to locate possible insulators within the region. The antibodies used are listed in Table 5.1 in detail. These antibodies were employed successfully in ChIP experiments in this study and will be discussed in detail.

Antibody	Abbreviation	Expected Genomic Regions	Cross Ref.
CTD of RNA polymerase II (phosphorylated at Serine 5)	polII	Actively transcribed regions	1.3
mono-methylated Histone H3 at lysine 4	H3K4me	?	1.4.5
di-methylated Histone H3 at lysine 4	H3K4me2	active promoters	1.4.5
tri-methylated Histone H3 at lysine 4	H3K4me3	active promoters	1.4.5
di-methylated Histone H3 at lysine 9	H3K9me2	Heterochromatin?	1.4.5
tri-methylated Histone H3 at lysine 27	H3K27me3	Silenced genes and heterochromatin?	1.4.5
acetylated Histone H3 at lysine 9 and 14	H3Ac	active promoters	1.4.4
acetylated Histone H4 at Lysines 5,8,12 and 16	H4Ac	?	1.4.4
CTCF	CTCF	mainly insulator elements	1.3.2.1

Table 5.1. Working antibodies used in this study.

5.1 Unsuccessful Antibodies

My strategy was to employ a battery of antibodies to characterise promoter elements in depth. However, many of the antibodies proved difficult to work with and lead to unsuccessful results. Note that there is little or no evidence in the literature for most of these antibodies to have been successfully used in ChIP studies in humans. ChIP

experiments were performed with antibodies recognizing TBP (Upstate, #06-241; Santa Cruz Biotechnology, #sc-273; Abcam, #ab818), TAF1 (Santa Cruz Biotechnology, #sc-735; Abcam, #ab14211) and TAF5 (Santa Cruz Biotechnology, #sc-743) proteins. Their use could have helped to differentiate different transcription initiation complexes assembled on promoter sequences (section 1.3.1) and correlate this with the activity profile and sequence characteristics of promoters. Likewise, an antibody recognising YY1 protein (Santa Cruz Biotechnology, (H-10), #sc-7341) was tested. YY1 can bind to initiator elements on TATA-less promoters (Usheva and Shenk, 1996) and have the ability to produce sharp DNA bends like TBP (Kim and Shapiro, 1996), which makes it a good candidate for a TBP substitute in TBP-free initiation complexes (TFTC). I also used an antibody recognizing Sp1 transcription factor (Upstate, #07-645) (section 1.3.2.1). The rationale was to examine sequence features of promoters (GC% or presence of a CpG island) activated by Sp1.

Antibodies recognizing H3K9me and H3K9me3 were employed in ChIP experiments to detect heterochromatic regions in 20q12-13.2. These experimental failures could be due to badly designed antibodies not being able to bind to epitopes. Alternatively, the region of interest might lack true heterochromatic segments, which may be the case since 20q12-13.2 has the highest gene density along chromosome 20 (see Figure 1.22). Furthermore, the fact that the region does not seem to contain any annotated developmental genes which are often in regions needing silencing also supports this alternative.

5.2 ChIP

Crosslinking is the process of chemically joining two or more molecules by a covalent bond. Crosslinking reagents contain reactive ends to form covalent bonds with specific functional groups (primary amines, sulfhydryls, etc.) on proteins or other

molecules. Crosslinking has many useful applications such as solid-phase immobilization, preparing antibody-enzyme conjugates, immunotoxins or other labelled protein reagents. It is also used for modification of nucleic acids, drugs and solid surfaces.

FA (formaldehyde) is a high resolution crosslinking agent since it can bridge distances of 2 Å (Jackson, 1978). It is a reactive dipolar compound where its carbon atom is the nucleophilic centre (see Figure 5.1).

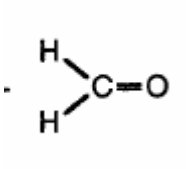


Figure 5.1 Chemical structure of formaldehyde.

Amino and imino groups of proteins (e.g. the side chains of lysine and arginine) and of nucleic acids (e.g. cytosine) react with FA forming a Schiff base (reaction I in Figure 5.2). This intermediate can then react with a second amino group (reaction II in Figure 5.2) and condenses (Orlando et al., 1997). A simple heat treatment is sufficient to reverse the reaction equilibrium toward de-crosslinking (Solomon and Varshavsky, 1985).

FA-crosslinking was the method of choice in this study although there are other crosslinking agents available such as ultra-violet light or laser which have been applied successfully in other studies (Lejnine et al., 1999). FA-crosslinking is commonly used since its crosslinking effect can be fully reversible as well as the chromatin structure being faithfully preserved during and after its application (Orlando, 2000). FA is also proven to be particularly effective in crosslinking lysine and arginine rich histone proteins to DNA (Kuo and Allis, 1999).

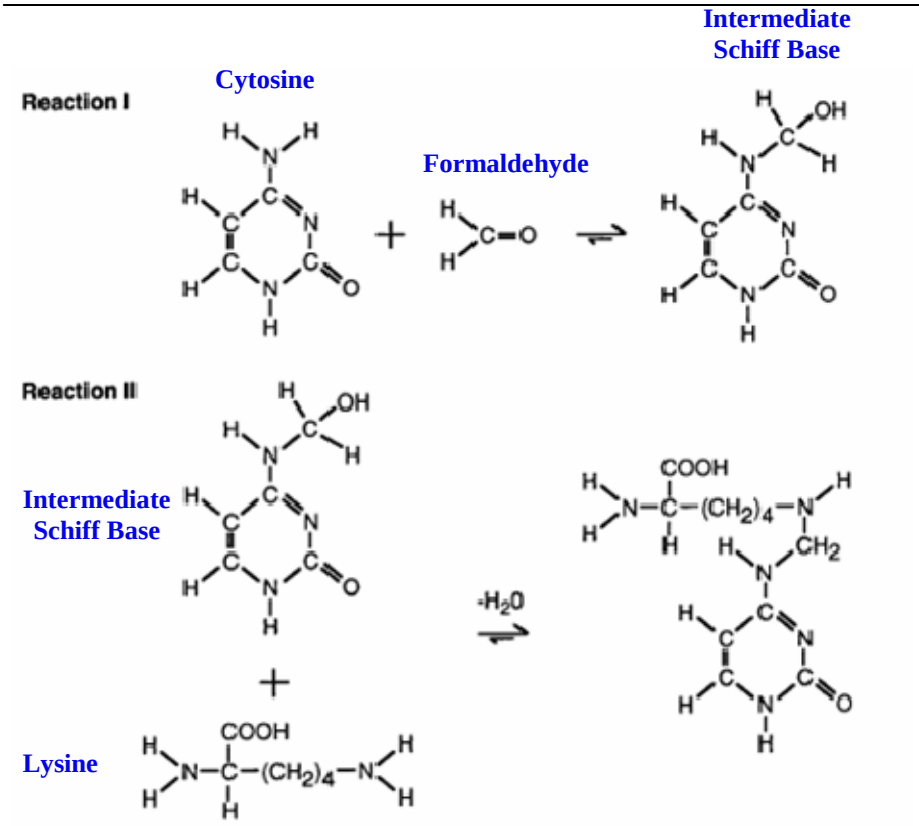


Figure 5.2. Crosslinking of Cytosine to a Lysine by formaldehyde. This figure is reproduced from reference (Orlando et al., 1997).

In this study, HeLa S3 and NTERA-D1 cells were incubated with various concentrations of FA and time courses depending on the nature of the protein targeted to crosslink to the DNA. Although the methodology for crosslinking histone proteins to DNA is well established in the literature (Schubeler et al., 2000) (Bernstein et al., 2005) (Nguyen et al., 2001), crosslinking other DNA binding proteins such as transcription factors requires a great deal of optimization (see section 5.2.1). This may be due to the transient and/or weaker interactions transcription factors have with DNA. Their relative positions within the DNA-protein interaction complex *in vivo* is an additional factor. Finally, the design of antibody epitopes against the protein of interest is also crucial. If the region of the protein where the antibody is supposed to bind (epitope) is occupied by another protein *in vivo*, one cannot precipitate any desired DNA-protein complex with such an antibody irrespective of the efficiency of crosslinking.

A schematic description of ChIP methodology is given in Figure 5.3.

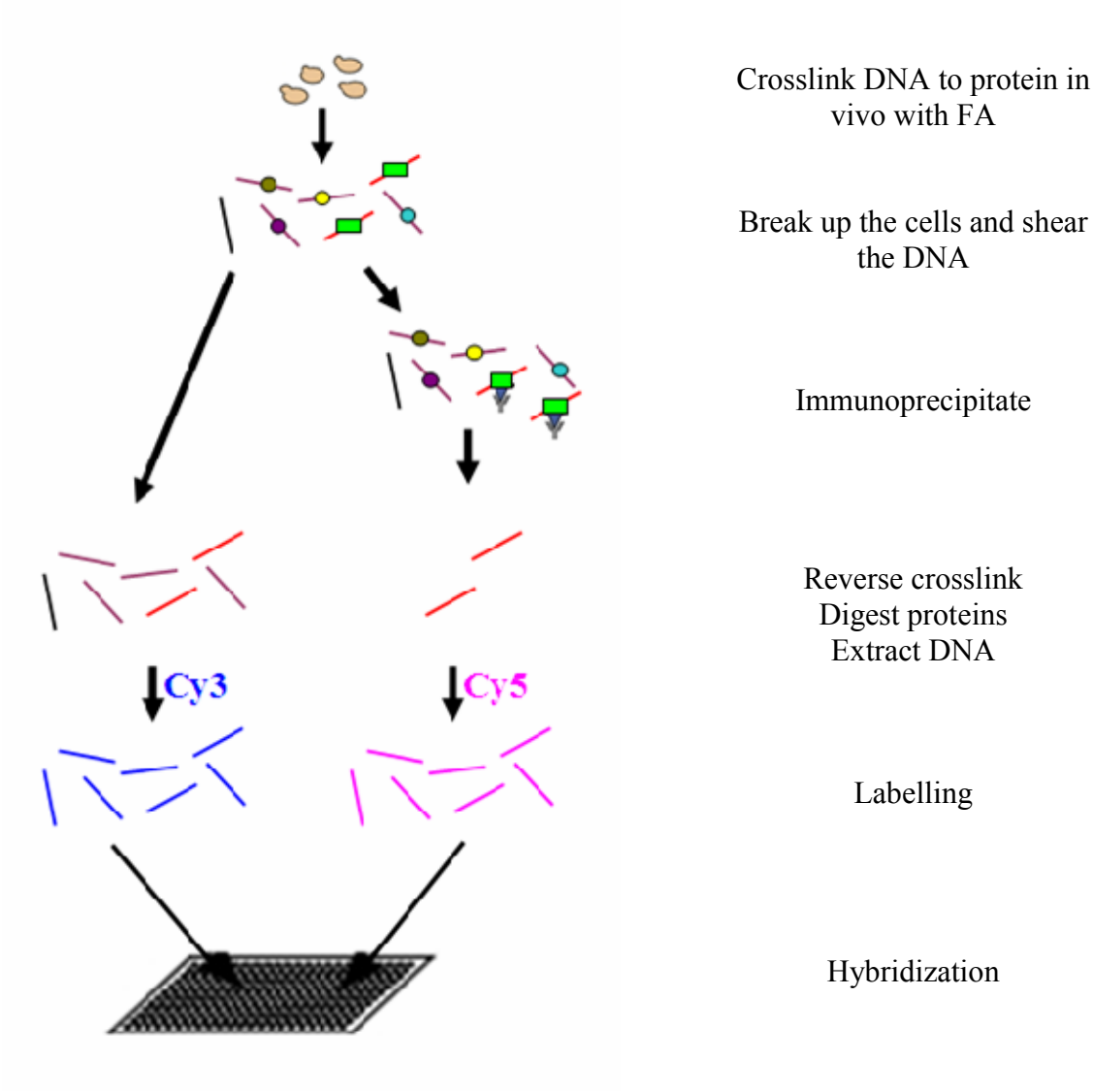


Figure 5.3. Schematic description of Chromatin Immunoprecipitation coupled with DNA microarrays

5.2.1 Optimization of Crosslinking Conditions

An already optimized crosslinking protocol, kindly provided by David Vetrie Lab (WTSI), was applied to crosslink histone proteins to DNA using ChIP, where cells were incubated with 0.37% FA solution for 15 min to crosslink histone proteins to DNA.

Condition Index	Crosslinking Agent	Incubation Time (min)	Crosslinking Agent	Incubation Time (min)
1	1% FA	15	n/a	
2		30		
3		45		
4		60		
5	DMA	30	1% FA	15
6		45		
7		60		

Table 5.2 The concentrations of crosslinking agents and corresponding incubation times to crosslink transcription factors to DNA.

To crosslink RNA polymerase II (polII) to DNA, I incubated cells with different concentrations of FA using different timings. I also used additional crosslinking agents such as Dimethyl Adipimidate.2HCl (DMA) (see section 2.5.2). The crosslinking conditions and timings used in the optimization process are listed in Table 5.2. The crosslinking efficiency was checked using real-time PCR with the primers aligning to the sequence of an active promoter (C20orf121) in 20q12-13.2 (see section 2.6). ChIP experiments were performed as described in section 2.5, then the ChIP sample, which should contain the DNA fragments where the protein of interest is bound *in vivo* (determined by the antibody used to co-immunoprecipitate crosslinked DNA-protein complexes), was used as a template for real-time PCR. The crosslinking conditions that gave the highest amplification with the real-time PCR were chosen for further experiments. Cells that were crosslinked with 1% FA for 60 min (condition 4) showed higher amplification levels than all other timings used. However, the amplification rates obtained from ChIP experiments with polII antibody were ~ten fold lower than those obtained with histone antibodies. For an attempt to improve crosslinking efficiency, I tried an additional crosslinking agent, DMA, which is a strong protein-protein crosslinker, bridging distances of 7.4 Å, and has been successfully applied to crosslink a number of transcription factors with the help of FA (Kurdistani et al., 2002; McCabe and Innis, 2005). DMA is particularly useful if the

protein of interest is not directly bound to DNA, most likely because it specifically crosslinks proteins using their primary amine groups (Green et al., 2001). The polII antibody used in this study recognizes the carboxyl terminal domain of largest subunit of polII, but it is known that polII initiation complex *in vivo* achieves its direct contact with the promoter DNA via its TFIID subunit (see section 1.3.1). Therefore, I used DMA to first crosslink protein complexes and then FA to crosslink proteins to DNA to increase crosslinking efficiency. Crosslinking with DMA for 60 min followed by 15 min crosslinking with 1% FA (condition 7; Table 5.2) showed the highest amplification, but similar to those of condition 4. Crosslinking agents such DSG and DSS, which can bridge distances of 6.2 and 8.8 Å respectively, were also tested instead of DMA, but both of them failed to give any amplification. Therefore, condition 4 was used as the simplest most efficient option for crosslinking polII to DNA for both cell lines.

The crosslinking condition for CTCF was kindly provided by David Vetrie Lab, where cells were incubated with 1% FA solution for 15 min for crosslinking.

5.2.2 Immunoprecipitation and subsequent steps in ChIP on chip

After crosslinking, the cell lysate was sonicated to obtain fragments of around 300-500 bp in size. At this point, a small amount of lysate (input chromatin) was taken and stored at -20 ˚C. This sample contains the full genomic content of a cell and serves as reference to estimate levels of enrichment of DNA sequences in later stages of the experiment. Following sonication, the lysate was incubated with the appropriate antibody for each protein to be investigated. Following overnight incubation of the cell lysate with an antibody, DNA-protein-antibody complexes were immunoprecipitated using protein G-coupled agarose beads, where protein G recognizes the immunoglobulin type of the antibody (Rabbit IgG or Goat IgG) used in

this study. DNA-protein-antibody complexes were then eluted and crosslinking was reversed by incubation at 65 ˚C. From this point on, input chromatin was also included. For a given antibody, the immunoprecipitated samples and the corresponding input chromatin were first treated overnight with Proteinase K to remove proteins and then DNA was extracted using phenol-chloroform; DNA should include most, if not all, of the sites where the protein of interest was already bound during the crosslinking treatment. After this step, there are several ways to analyse ChIP samples. Real time PCR or quantitative PCR are commonly used methodologies where the end product of ChIP and input chromatin are used as template to amplify targeted regions of interest (Johnson and Bresnick, 2002) (Weinmann et al., 2001) (Wang, Derynck et al., 2004). Although PCR methods are shown to be both useful and sensitive, they are too laborious to scale up for whole genome and limited by the amplification efficiency of the regions of interest (Shieh and Li, 2004). An alternative methodology is the utilization of DNA microarrays to analyse ChIP material. DNA arrays can have either spotted overlapping DNA fragments or long oligonucleotides to cover the stretch of DNA under investigation. DNA arrays overcome the limitations of the PCR-based methodologies, as with current array spotting densities several megabases of DNA can be accommodated on a single chip. Thus, this approach was chosen for analysing 3.5 Mb of the 20q12-13.2 region. The construction of the custom-made DNA arrays is described in section 2.4. Input chromatin was labelled with Cyanine 3 coupled cytosine nucleotide (Cy3-dCTP) analogue, and ChIP material was labelled with Cyanine 5 (Cy5-dCTP) coupled cytosine nucleotide analogue. Approximately 500 ng of input chromatin and 4/5 of a ChIP sample (containing 800 to up to 4000 ng of DNA depending on the antibody used) was used for labelling. After removing unlabelled nucleotides from the samples, the labelled ChIP sample and input chromatin were combined and ethanol precipitated together with Human

Cot-1 DNA. Human Cot-1 DNA is a fraction of DNA consisting largely of highly repetitive sequences obtained from total genomic DNA by selecting for rapidly re-associating DNA sequences during renaturation of DNA (Strachan, 2003). The DNA mixture (containing ChIP material, input chromatin and Human Cot-1 DNA) was resuspended and then denatured. Upon denaturation, unlabelled human Cot-1 DNA will readily associate with complementary strands of the repetitive sequences within the labelled probe, thereby effectively blocking the repetitive parts of the labelled sequences. This step is crucial as highly-represented labelled repeat sequences can cause unreliable data (Mantripragada et al., 2004). This repeat masked material containing input chromatin and ChIP material was then hybridized for 48 h to the custom-made DNA microarray described in section 5.3. Hybridized arrays were then scanned using appropriate lasers for Cy3 and Cy5 dyes and light intensity on each spot was digitalized using a software called ProScanArray® Express (Perkin Elmer LAS Inc.). Spots that have a higher signal from Cy5 (ChIP material) channel are the potential binding sites for the protein of interest.

5.3 Custom-made 3.5 Mb Tilepath Array of human 20q12-13.2

As part of the chromosome 20 SNP discovery project at the WTSI (Spencer et al., 2006) resource of the 2 kb plasmid clones across the entire chromosome was available. These clones were generated from flow-sorted chromosome 20 DNA from four different lymphoblastoid cell lines. Each library was sequenced to a two fold depth. The paired reads (forward and reverse) were used to generate a chromosome 20 tilepath (Zemin Ning at the WTSI) of the clones.

The need to pick clone templates manually from the libraries and PCR amplify all selected clones was the main reason for selecting a sub-region of 3.5 Mb rather than the whole 10 Mb region at 20q12-13.2. This sub-region is from 42,274,163 to

45,850,636 bp and was selected to encompass the most gene rich section. In total, 1875 clones were amplified by PCR and 1795 were successful. Further, four primer pairs were designed to close gaps that contained an annotated TSS, and of those, three were successfully amplified. In total, 1798 amplicons were successful and therefore spotted on the array.

The Microarray Facility at the WTSI spotted the above PCR products to generate a custom-made array (see section 2.4). This array contains 1798 spots and has a resolution of approximately 1.8 kb (2067 ± 483 bp). There are 511 gaps of various sizes ranging from 2 bp to 4182 bp. The distribution of gap sizes is shown in Figure 5.4, where 75% of gaps are less than 800 bp long.

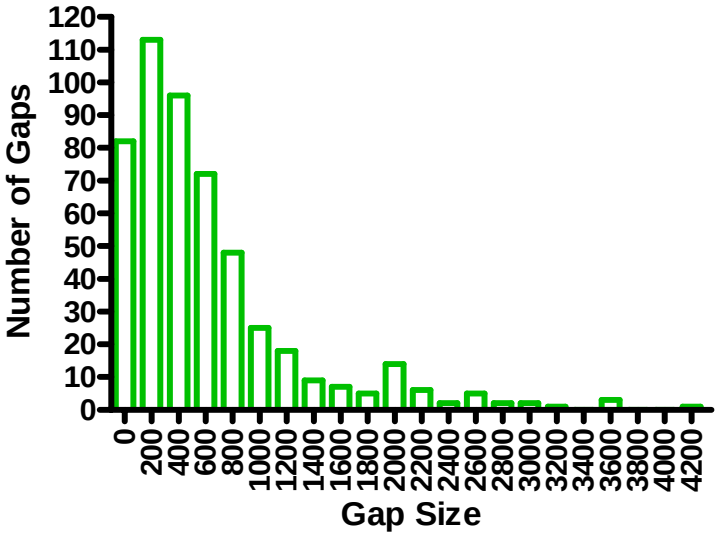


Figure 5.4 Distribution of the gaps in 3.5 Mb tilepath array. The 25% and 75% percentile of the gaps is 157 and 799 bp respectively with a median of 420 bp.

On the spotted array, there are 1723 working spots since 75 spots did not show any hybridization signal. Of these 75 non-working spots, five contain the annotated start site of five different genes (see Table 5.3).

The array contains 72 genes, covering 70% of all annotated genes at 20q12-13.2. Out of the 72 genes, 66 are represented on the array by at least one working spot containing the TSS of one of its coding transcripts. The TSS of 86% (150/175) of the

coding transcripts was mapped to at least one working spot on the array. Out of the 91 non-coding transcripts in the 3.5 Mb sub-region, the 5'ends of 79 (87%) were mapped on at least one working spot. There are 26 processed transcripts (transcripts with ambiguous ORF) with the 5' end of 24 processed transcripts being mapped to at least one working spot. Also, the 5'ends of 16 out of the 18 pseudogenes were also mapped to at least one working spot.

As mentioned earlier, the two cell lines HeLa S3 and NTERA-D1 were analysed. Based on the gene expression analysis described in section 3.4, 33 and 39 genes were expressed in HeLa S3 and NTERA-D1 cell respectively. Table 5.3 lists the genes whose TSS is not represented on the array. Therefore, one expects 28 and 35 spots showing positive signals with the antibodies that detect active TSSs in HeLa S3 and NTERA-D1 cells respectively.

Gene Name	Expression in HeLa S3	Expression in NTERA-D1
ADA	P	P
KCNK15	P	A
SLPI	P	A
DNTTIP1	P	P
SNX21	P	P
CDH22	A	P

Table 5.3 Expression profile of the genes whose TSSs are not contained by any working spot on the array in HeLa S3 and NTERA-D1 cells.

Gene predictions were also taken into consideration while assessing the gene feature content of the spots on the array (see below). There are 103 gene predictions annotated in VEGA Genome Browser (version 19) and the 5' end of 62 gene predictions were included within the boundaries of at least one spot on the array.

Figure A1 (Appendix A) displays three scanned arrays carrying signals obtained with ChIP experiments performed with antibodies recognizing H3K4me3, H3K4me2 and H3K4me.

5.4 Determining Spot Intensities

For each spot on the array, there are two signals; one comes from the cyanine 3 channel produced by the Cyanine 3 labelled DNA sequences (ChIP material) hybridizing on the spot and the other signal is produced by cyanine 5 channel by the Cyanine 5 labelled DNA sequences (input chromatin) hybridizing on the same spot. If a spot carries target sites for the protein of interest, then the cyanine 3 signal intensity should be higher than cyanine 5 since ChIP material should be enriched in fragments with binding sites for this protein. The input chromatin contains the same number of each genomic fragment. Therefore, in order to find out whether a spot is "enriched", the normalized cyanine 3 signal is divided by the cyanine 5 signal, and a high signal indicates the presence of possible binding sites of the protein of interest. This ratio determines the signal of a spot.

There are three replicates of each spot on each array (technical replicates). Each ChIP experiment for a given antibody was done in triplicate (biological replicates). For each replicate the negative control antibody (Rabbit IgG or Goat IgG) carrying the isotype of the main antibody was included. Let x_{ij} be the signal from i^{th} technical replicate of the j^{th} biological replicate, and bg_{ij} be the signal of the negative control antibody (background signal) from i^{th} technical replicate of the j^{th} biological replicate. The mean signal ($\bar{x}_j$) and background signal ($\overline{bg_j}$) obtained from i^{th} technical replicate of j^{th} biological replicate were both calculated as below (Equation 5.1).

$$\overline{a_j} = \left(\sum_{i=1}^{3} a_{ij} \right) \times \frac{1}{3} \textbf{ where } \mathbf{a} \in \{\mathbf{x}, \mathbf{bg}\} \quad \text{Equation 5.1}$$

Also, the standard error of the mean signal (σ_{X_j}) at each technical replicate is given as;

$$\sigma_{a_j} = \sqrt{\frac{1}{2} \times \sum_{i=1}^{3} (\overline{a_j} - a_{ij})^2 \times \frac{1}{3}} \textbf{ where } \mathbf{a} \in \{\mathbf{x}, \mathbf{bg}\} \qquad \text{Equation 5.2}$$

The standard error of the mean background signal (σ_{bg_j}) was calculated according to Equation 5.2

It is assumed that the signal coming from the negative control antibody is also included in the actual measured signal since the antibody and the negative control antibody share the same isotype. Therefore, the actual signal (S_j) is calculated by subtracting the background signal ($\overline{bg_j}$) from the mean signal ($\bar{x}_j$) at j^{th} technical replicate. This subtraction aims to eliminate spots that are enriched in a non-specific manner.

To assess the distribution of the Cy3 signals, the raw Cy5 signals (horizontal axis) are plotted against the raw Cy3 signals obtained from ChIP-chip experiments performed with rabbit IgG (negative control antibody) and tri-methylated K4 of histone H3 (H3K4me3) in NTERA-D1 cells (Figure 5.5). The spots that are on the upper left space in panel B (red circle) correspond to those that have a high Cy3 to Cy5 signal and present sites that are potentially enriched specific to the protein of interest.

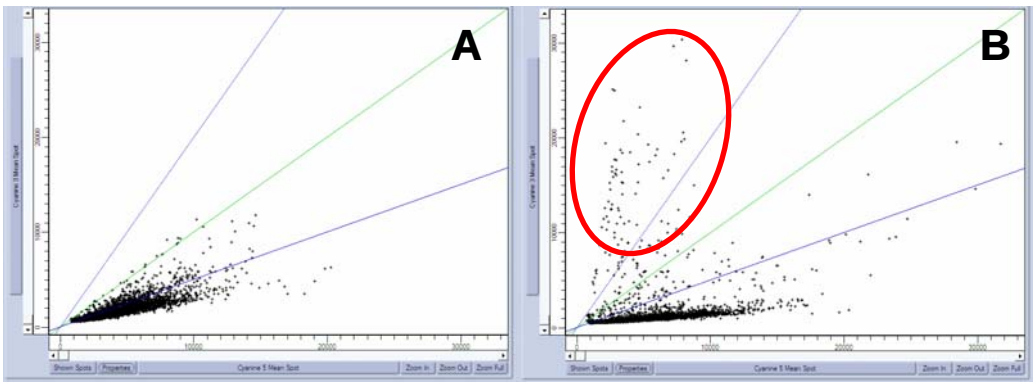


Figure 5.5 The graphs above show raw Cy5 (input chromatin, horizontal axis) relative to the raw Cy3 (antibody, vertical axis) signals for Rabbit IgG (A) and tri-methylated K4 of Histone H3 (B) in NTERA-D1 cells. In graph A, there are no spots on the array, which produce high Cy3 to Cy5 signals, whereas in the graph B, a number of spots (red circle) have high Cy3 to Cy5 signals and are potential biological targets of the protein of interest.

The standard error of the actual signal will be the square root of the sum of the squared standard errors of the measured signal (σ_{X_j}) and the background signal (σ_{bg_j}), since the background (noise) signals were measured independently (Equation 5.3) (Abramowitz, 1972).

$$\sigma_{S_j} = \sqrt{\left(\sigma_{x_j}\right)^2 + \left(\sigma_{bg_j}\right)^2} \qquad \text{Equation 5.3}$$

Then, the average of actual signals obtained from each biological replicate (S_j) was taken to calculate the actual mean signal (S) (Equation 5.4)

$$S = \left(\sum_{j=1}^{3} S_j\right) \times \frac{1}{3} \qquad \text{Equation 5.4}$$

The standard error of the actual signal (σ) propagated over the biological replicates can be calculated according to Equation 5.5 (Abramowitz, 1972).

$$\sigma = \frac{1}{3}\sqrt{\sum_{j=1}^{3}\left(\sigma_{S_j}\right)^2} \qquad \text{Equation 5.5}$$

The above allows associating every spot on the array with a signal and standard error for a given antibody. The signal should give the enrichment level specific to the antibody used since the spot signal coming from the negative control antibody (Rabbit or Goat IgG) is subtracted. Due to low resolution of the array, high signals come from one or at most two adjacent spots. For example, all spots that carry a TSS represent the centre of a high signal whereas neighbouring spots show no such high signal (<2) if they have no sequence overlap with the fragment carrying the TSS. Therefore, no peak finding algorithms was necessary to find "enriched" spots (Kim et al., 2005).

Figure 5.6 shows the background subtracted mean signals of the spots for H3K4me3 in NTERA-D1 cells.

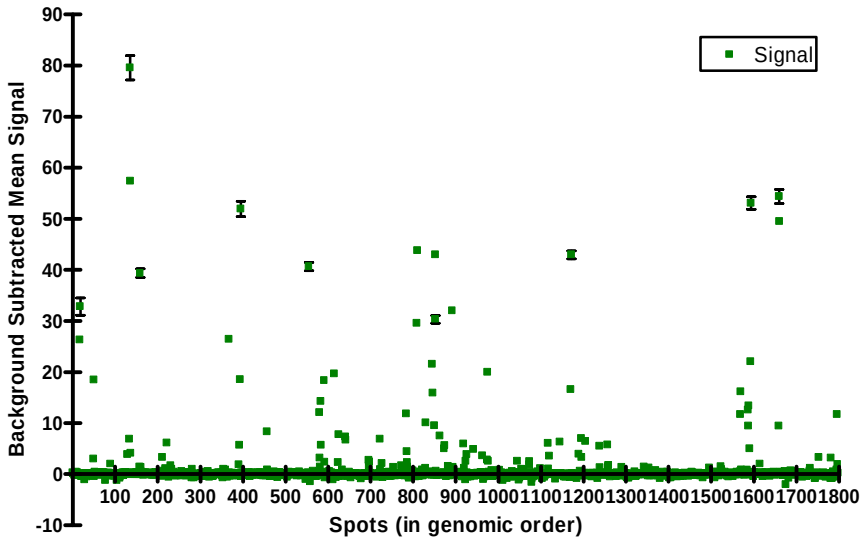


Figure 5.6 The background subtracted mean signals and standard errors obtained from ChIP-chip using antibody recognising H3K4me3 in NTERA-D1 cells. The horizontal axis corresponds to the spots, and are ordered according to the genomic coordinates of the sequences they carry.

The signal does not follow a normal distribution. Therefore statistical analysis valid for datasets with normal distributions could be misleading. The 'outliers' most likely represent the biological genomic targets of the protein of interest. However, since the total number of true positives and true negatives is not known, it is difficult to determine the signals that are the true biological targets. Here, I employ an empirical data analysis method to set a threshold to estimate the number of true positives with a false positive ratio below 5%. Signals obtained from spots that are most likely true positives (for example TSSs for polII and H3K4me3) and true negatives (inter-genic regions with no enrichment with any antibody) were validated using real-time PCR (method described in section 2.6.2). The amplification level in real-time PCR was correlated with the enrichment levels in ChIP on chip.

5.4.1 Validation of ChIP-chip enrichments by Real-time PCR

Promoter regions of three genes (*C20orf121*, *ADA*, *ZNF335*) were selected and primers were designed to amplify around 1000 bp upstream to 300 bp downstream of their annotated TSS. Also, primers were designed to amplify the TSSs of *C20orf142*,

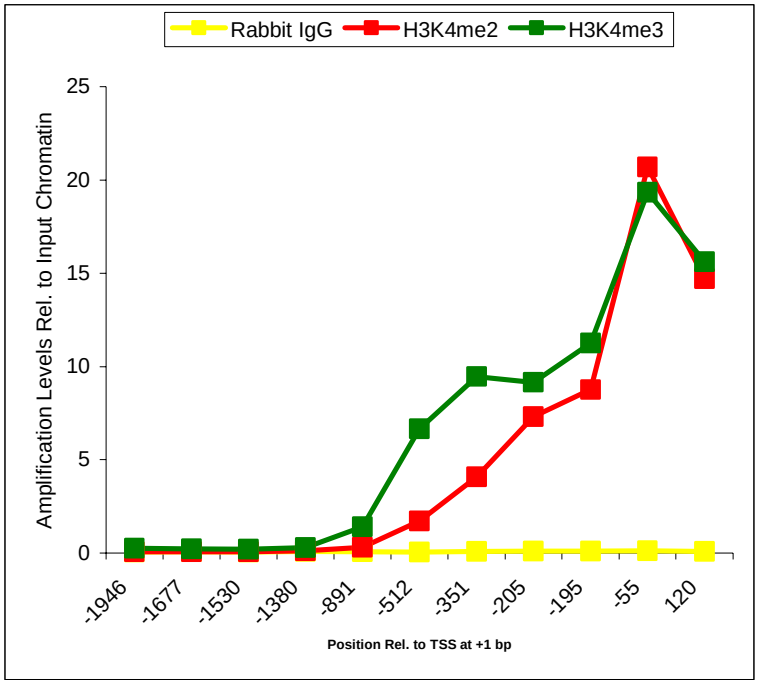
TOMM34, *YWHAB*, *SLPI* and *PIGT*. For H3K27me3 and CTCF antibodies, four regions with a high signal were selected for real-time PCR validations. Finally, 25 inter- and intra-genic assays were designed to amplify regions where no signal was obtained with any antibody used in this study (in either of the cell lines). These will be referred as 'negative real-time PCR controls' All primer sequences are listed in Appendix B. ChIP and input chromatin material obtained from several experiments were quantified using NanoDrop®, and it is estimated that 1:40 diluted input chromatin material contains approximately the same amount of DNA with the ChIP material. Therefore, 0.1 µl of ChIP material or 0.1 µl of diluted input chromatin (1:40) was used for each amplification. Ct values are determined and the amplification level (E) for each antibody is calculated according to the equation given below;

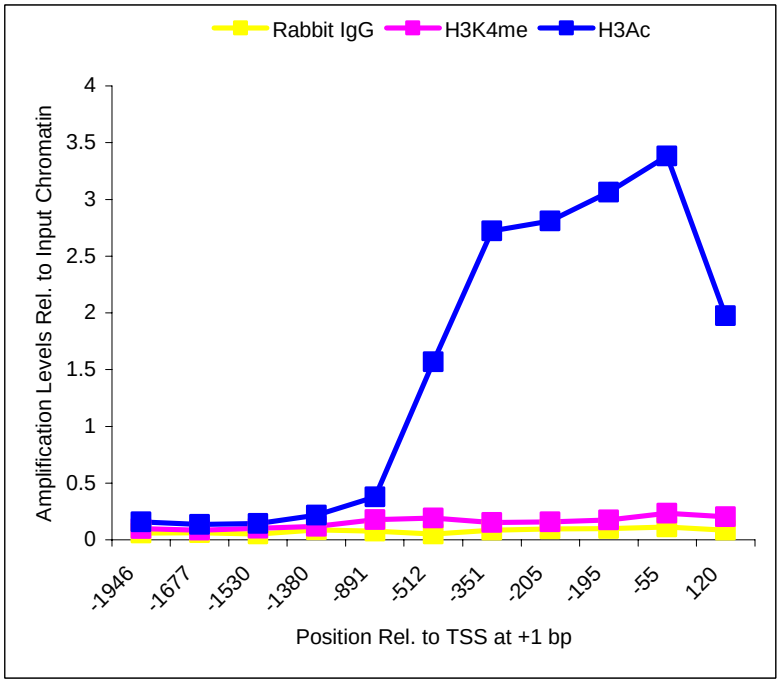
$$E = 2^{(-(Ct(ChIP) - Ct(InputChromatin))}$$

Real-time PCRs were performed with all designed primer pairs (Appendix B) for each ChIP material and the corresponding input chromatin. Promoter region of *C20ORF121* is given as an example for all but H3K27me3 and CTCF antibodies.

Figure 5.7 shows the amplification levels relative to input chromatin using ChIP materials obtained by five antibodies recognizing Rabbit IgG, H3K4me, H3K4me2, H3K4me3 and H3Ac in NTERA-D1 cells respectively. Amplification levels by real-time PCR fully mirror the enrichment level obtained with ChIP on chip, although they are not on the same scale. The spot neighbouring the TSS-carrying spot did not give any signal above background, and no amplification can be seen by real-time PCR.

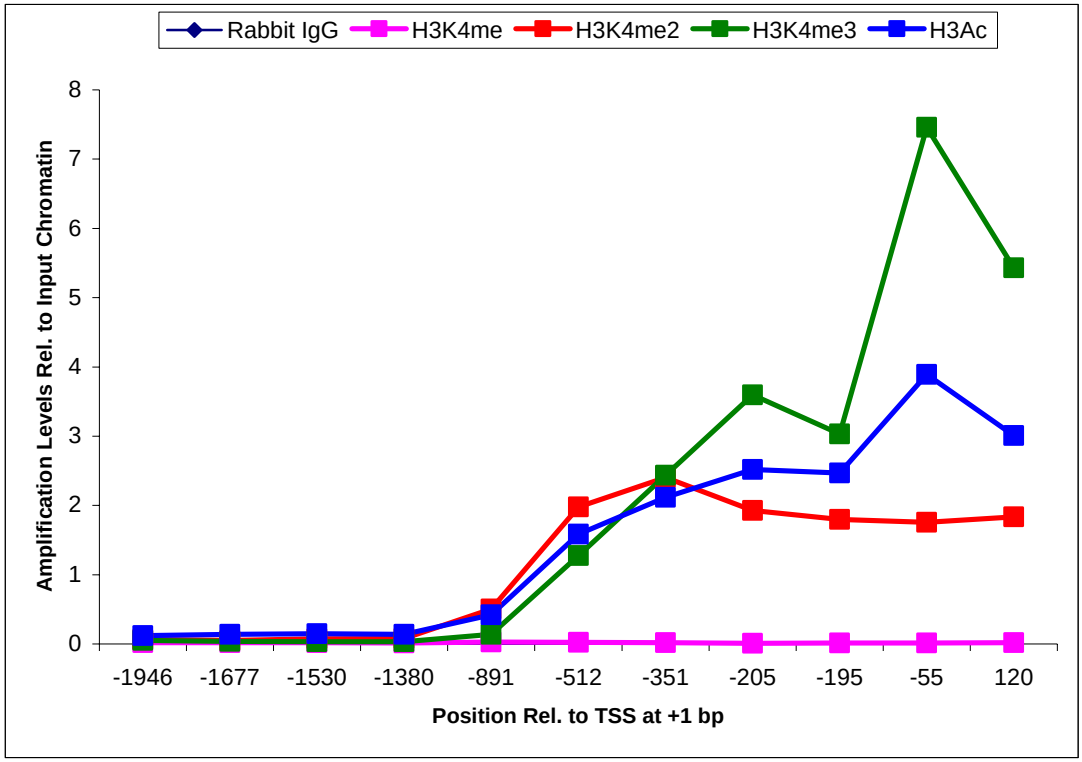
The same analysis was performed with ChIP material obtained with these antibodies in HeLa S3 cells and the amplification levels are shown in Figure 5.8.





Start (bp)	End (bp)	Spot Information	H3K4me	H3K4me2	H3K4me3	H3Ac
-2830	-957		-0.10	0.03	0.21	0.07
-308	435	carries TSS of C20orf121	-0.02	25.77	79.56	7.10

Figure 5.7 Amplification levels of the promoter region of *C20orf121* by real-time PCR using ChIP material obtained by using antibodies listed above and the input chromatin in NTERA-D1 cells. The table below shows the enrichment level of the spot carrying the TSS of the gene and the upstream neighbouring spot in NTERA-D1 cells. The spot coordinates are given according to the TSS.

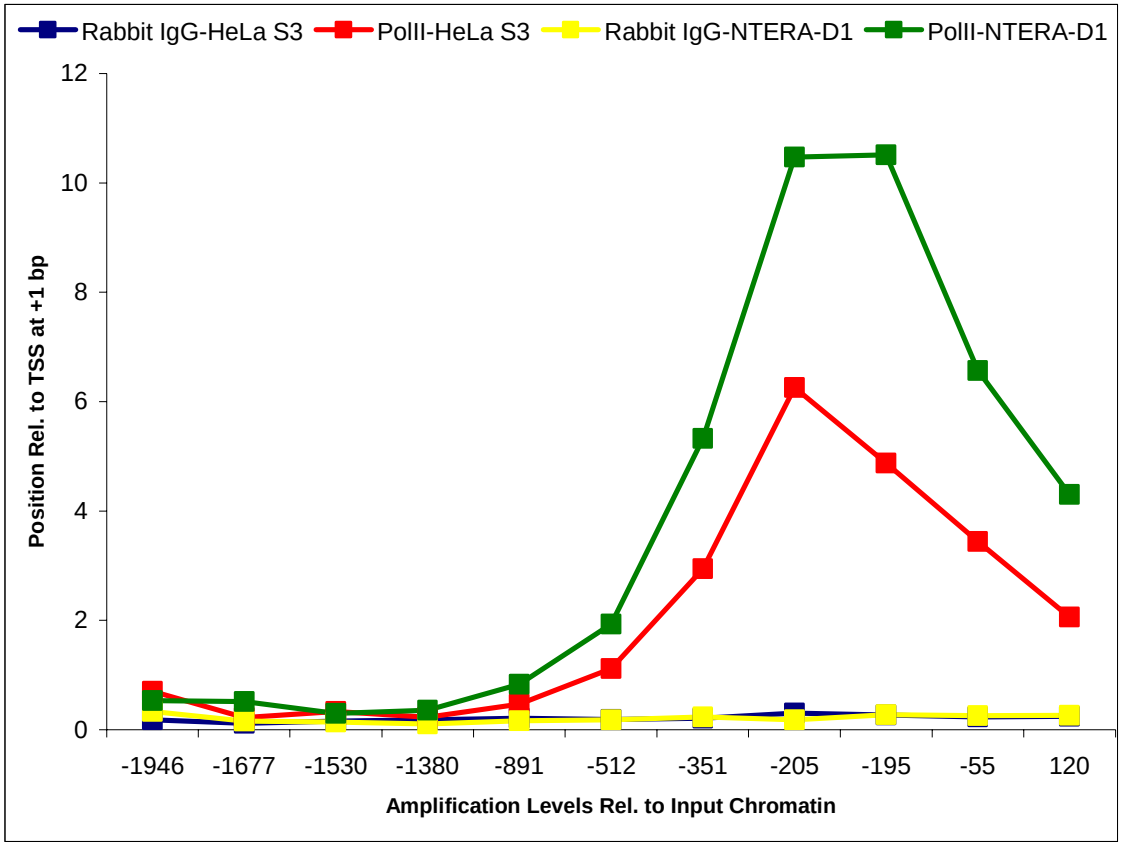


Start (bp)	End (bp)	Spot Information	H3K4me	H3K4me2	H3K4me3	H3Ac
-2830	-957		1.01	0.36	0.40	0.46
-308	435	carries TSS of C20orf121	1.29	17.97	36.72	14.14

Figure 5.8 Amplification levels of the promoter region of *C20orf121* by real-time PCR using ChIP material obtained by using antibodies listed above and the input chromatin in HeLa S3 cells. The table below shows the enrichment level of the spot carrying the TSS of the gene and the upstream neighbouring spot in HeLa S3 cells. The spot coordinates are given according to the TSS.

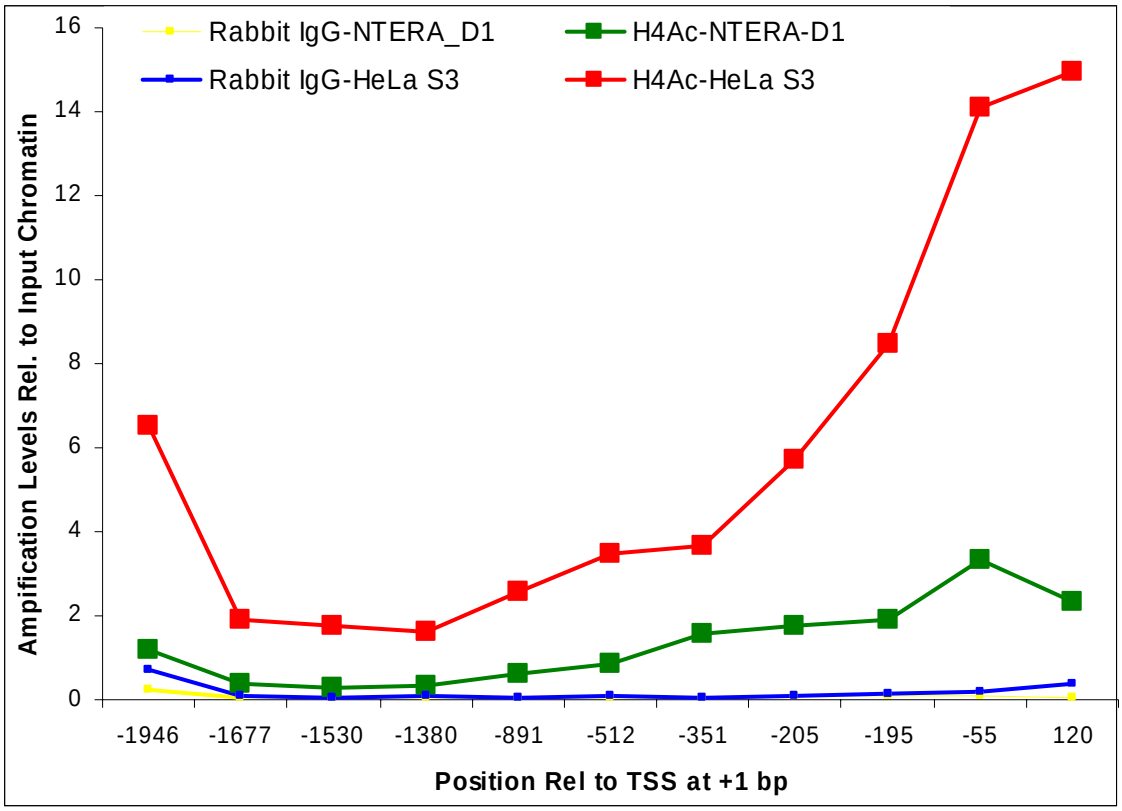
Amplification levels of the ChIP material obtained from HeLa S3 with above listed antibodies relative to input chromatin validated the enrichment levels by ChIP on chip. Note that real-time PCR and ChIP on chip signal intensities follow the same trend in the two cell lines, i.e. higher in NTERA-D1 than in HeLa S3.

The enrichment levels obtained by the antibody recognising RNA polymerase II (polII), H3K4Ac and H3K9me2 were validated with real-time PCR in both HeLa S3 and NTERA-D1 cells. Figure 5.9 – 5.11 show the amplification levels of the *C20orf121* promoter region.



Start (bp)	End (bp)	Spot Information	polII –HeLa S3	PolII – NTERA-D1
-2830	-957		0.10	0.26
-308	435	carries TSS of C20orf121	5.15	7.44

Figure 5.9 Amplification levels of the promoter region of *C20orf121* by real-time PCR using ChIP material obtained by using antibodies recognising Rabbit IgG and PolII, and the input chromatin in HeLa S3 and NTERA-D1 cells. The table below shows the enrichment level of the spot carrying the TSS of the gene and the upstream neighbouring spot. The spot coordinates are given according to the TSS.

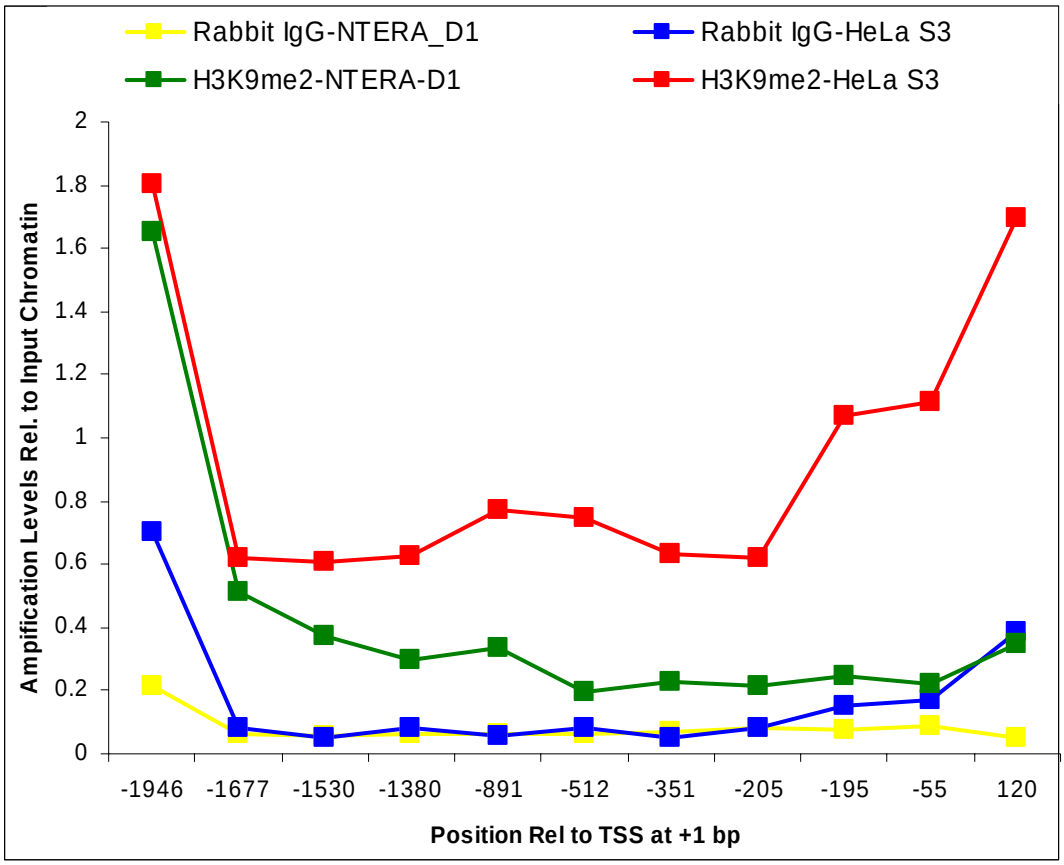


Start (bp)	End (bp)	Spot Information	H4Ac –HeLa S3	H4Ac – NTERA-D1
-2830	-957		0.74	-0.05
-308	435	carries TSS of C20orf121	6.51	0.16

Figure 5.10 Amplification levels of the promoter region of *C20orf121* by real-time PCR using ChIP material obtained by using antibodies recognising Rabbit IgG and H4Ac and the input chromatin in HeLa S3 and NTERA-D1 cells. The table below shows the enrichment level of the spot carrying the TSS of the gene and the upstream neighbouring spot. The spot coordinates are given according to the TSS.

Amplification levels are in concordance with the enrichment levels obtained by ChIP on chip with antibodies above in both cell lines.

In Figure 5.10, ChIP material obtained by H4Ac antibody in NTERA-D1 cells showed ~2 fold amplification relative to the input chromatin whereas no significant enrichment was observed in ChIP on chip experiment. Real-time PCR appears to be more sensitive than ChIP on chip since a low resolution array was used to detect enrichments.



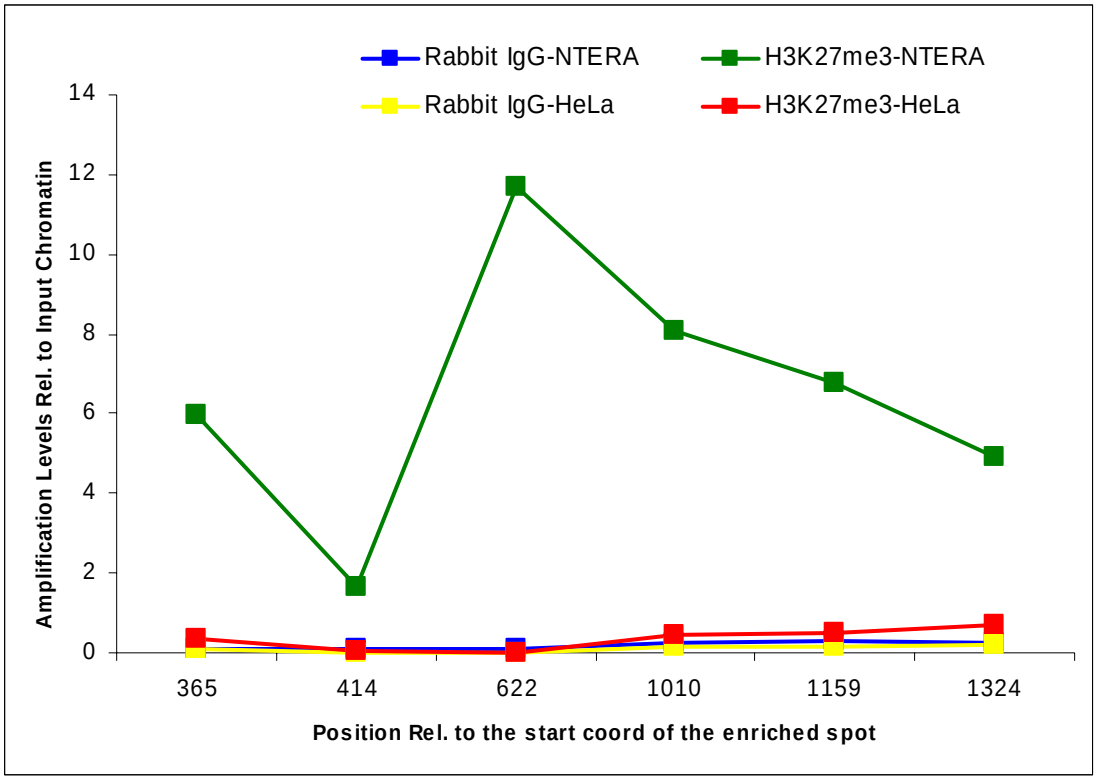
Start (bp)	End (bp)	Spot Information	H3K9me2 – HeLa S3	H3K9me2 – NTERA-D1
-2830	-957		-0.02	0.02
-308	435	carries TSS of C20orf121	-0.52	-0.89

Figure 5.11 Amplification levels of the promoter region of *C20orf121* by real-time PCR using ChIP material obtained by using antibodies recognising Rabbit IgG and H3K9me2 and the input chromatin in HeLa S3 and NTERA-D1 cells. The table below shows the enrichment level of the spot carrying the TSS of the gene and the upstream neighbouring spot. The spot coordinates are given according to the TSS.

Amplification levels using ChIP material obtained by H3K9me2 antibody in HeLa S3 were slightly higher than those of Rabbit IgG, however it is very low compared to the enrichment levels obtained by other antibodies. Real-time PCR reactions performed for regions elsewhere did not show high amplification levels either (<2.5). This antibody did not show high enrichment levels on ChIP on chip experiments in both cell lines.

For H3K27me3 validation, four regions with high signal were used in real-time PCR. Figure 5.12 shows the amplification levels in both cell lines (H3K27me3_Pr21 to

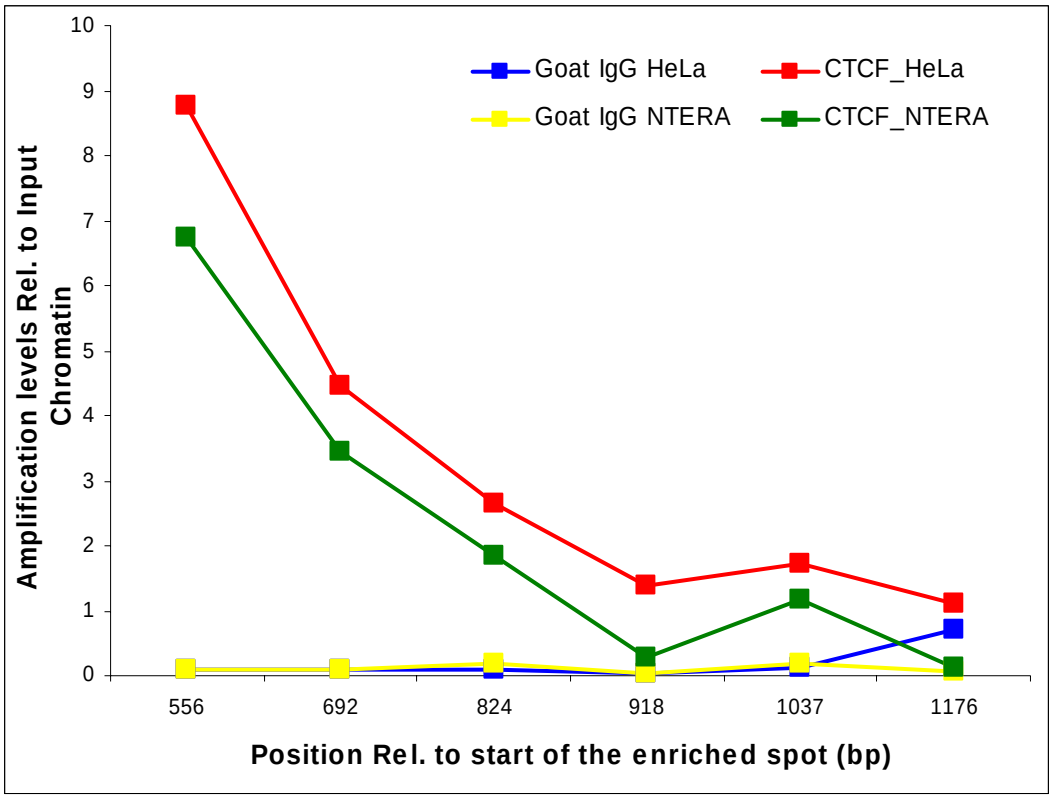
H3K27me3_Pr26, Appendix B) only enriched by H3K27me3 in NTERA-D1 cells. The spots that were enriched in HeLa S3 cells were also validated with real-time PCR (data not shown).



Start (bp)	End (bp)	Spot Information	H3K27me3 –HeLa S3	H327me3 – NTERA-D1
44,094,336	44,095,961		0.65	7.20

Figure 5.12 Amplification levels of an H3K27me3 enriched spot in NTERA-D1 cells by real-time PCR using ChIP material obtained by using antibody recognising Rabbit IgG and H3K27me3 and the input chromatin in HeLa S3 and NTERA-D1 cells. The table below shows the enrichment level of the spot in both cell lines.

Lastly, enrichment levels obtained by the antibody recognising CTCF were validated by designing real-time PCR primers to four regions enriched in CTCF in one or both cell lines. Figure 5.13 shows the amplification levels on a region enriched by CTCF in both cell lines.



Start (bp)	End (bp)	Spot Information	CTCF –HeLa S3	CTCF – NTERA-D1
-1421	+1171		8.67	5.69
+1	+1393		7.49	5.30

Figure 5.13 Amplification levels of an CTCF enriched spot in HeLa S3 and NTERA-D1 cells by real-time PCR using ChIP material obtained by using antibody recognising Goat IgG and CTCF and the input chromatin in HeLa S3 and NTERA-D1 cells. The table below shows the enrichment level of the spot and its upstream neighbouring spot in both cell lines.

Results for ADA, ZNF335 and the TSS-containing amplicons gave comparable results for the corresponding antibodies (data not shown). Finally, none of the 25 negative real-time PCR control regions showed significant amplification (<2-fold) above input chromatin (all antibodies in both cell lines).

After validation using real-time PCR, the following strategy was employed for H3K4me3 and polII to estimate the number of true enriched regions. The background-subtracted mean signals produced by H3K4me3 antibody were considered to select a threshold. First, the spots that gave enrichment more than one-fold over the background were taken. Within this set, the signals coming from the spots carrying a

TSS were considered and the minimum signal among those spots was determined. Then, the spots that gave a more than one-fold enrichment with more than one antibody were taken and the minimum signal among those spots was also determined since such spots are assumed to be more likely to be true positives. A threshold was selected to include the maximum number of "enriched" spots that are either carrying a TSS or are enriched with two or more antibodies while keeping the number of spots that are "enriched" with only one antibody to no more than 5% of the total number of "enriched" spots. I used this as an estimate of the false positive ratio of the experiment.

As an example, Figure 5.14 shows the case of the H3K4me3 antibody in NTERA-D1 cells, in which the signals are plotted in relation with the type of the spot and the signals obtained with other antibodies. For ease of visualisation, only the spots that gave a signal between 1 and 5 were shown. All spots showing signal values higher than 5 either carry or neighbour a TSS, or showed enrichment with more than one antibody. Here, the threshold for including the maximum number of possible true positive spots while keeping false positive ratio at most 5% was 1.5.

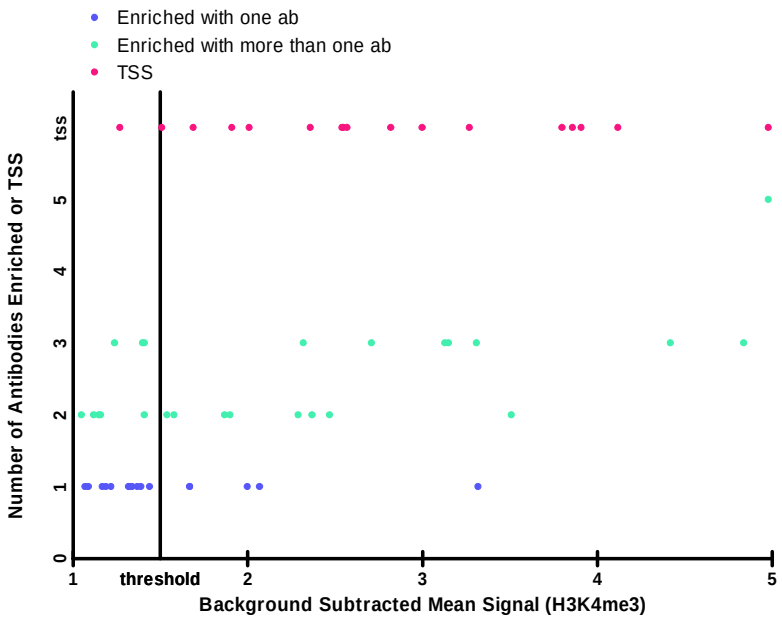


Figure 5.14 The H3K4me3 signals in NTERA-D1 cells. Spots that showed signals between 1 and 5 were shown here for the ease of visualisation. Vertical axis denotes the number of antibodies that a corresponding spot showed "enriched" signal. "tss" denotes that these spots either carry or neighbours a TSS.

The same threshold was determined for H3K4me3 antibody in HeLa S3 cells. The threshold for RNA polymerase II antibody was set to 1.75 in both cell lines to achieve a false positive ratio below 5%.

The same strategy cannot be directly applied to other histone antibodies since we have a limited knowledge of known true targets. Therefore, the threshold 1.5 was also selected for other histone antibodies based on the real-time PCR data and the assumption that enrichment patterns between different histone antibodies will be similar. As only real-time PCR data were available for CTCF antibody, conservative threshold 1.75, was selected.

5.5 Analysis of RNA polymerase II binding sites by ChIP

ChIP experiments were performed to screen for RNA polymerase II (polII) binding sites in HeLa S3 and NTERA-D1 cells. The polII antibody (Abcam) #ab5131 was raised in rabbit and it recognizes the phosphorylated serine found in the amino acid 5 position (serine-5) of the carboxyl terminal domain (CTD) repeat YSPTSPS of the

protein. The unphosphorylated form of CTD is required for the initiation of the transcription and this form also interacts with a wide range of general transcription factors (Proudfoot et al., 2002). However, once the transcription is initiated, serine-5 at CTD is phosphorylated by the basal transcription factor TFIIH at the promoter, and the mRNA capping enzyme is brought to the initiation complex via binding to this modified form of CTD (Kim et al., 2004) (see section 1.3.1). During elongation, serine-5 phosphorylation drops and the capping enzyme dissociates (Komarnitsky et al., 2000). Therefore, this antibody should be able to detect DNA regions where RNA polymerase II assembles itself to initiate transcription.

For each transcript, the spot carrying a TSS plus the spots carrying the right and left flank were searched for an enrichment in signal. This roughly corresponds to 2 kb upstream and downstream of the TSS. Heat maps were generated for these sites. A heat map is a false-coloured graph representation of the signal intensities plotted against the positions of the feature in the genome. Higher signal values are represented as red while lower signal values are represented as white.

In a second round, spots that gave a high signal but not overlapping within ±2 kb of TSS of a coding transcripts were discussed. Such signals were analysed in the context of annotated features such as processed transcripts, pseudogenes or gene predictions including multi-species conservation of the DNA sequence of these spots.

5.5.1 Results in HeLa S3 cells

There were 62 spots which gave a signal intensity higher than the selected threshold (1.75) in HeLa S3 cells using the RNA polymerase II antibody (Abcam, #5131). Of these, 36 enriched spots were located ±2 kb of the TSS of an annotated feature. There were 25 spots within the genes whose TSSs showed polII enrichment. The remaining three spots were not in close proximity of any annotated TSS.

With respect to coding transcripts, the 35 enriched spots near a TSS represent 16 coding transcripts (15 genes). Per signal the intensity of the enriched spot and that of the three adjacent spots on either side were plotted against their relative position to the start site of the coding transcript they contain or neighbour. Figure 5.15 shows the signals of the 16 coding transcripts in the colour coded manner known as a heat map, red representing the highest and white the lowest signal.

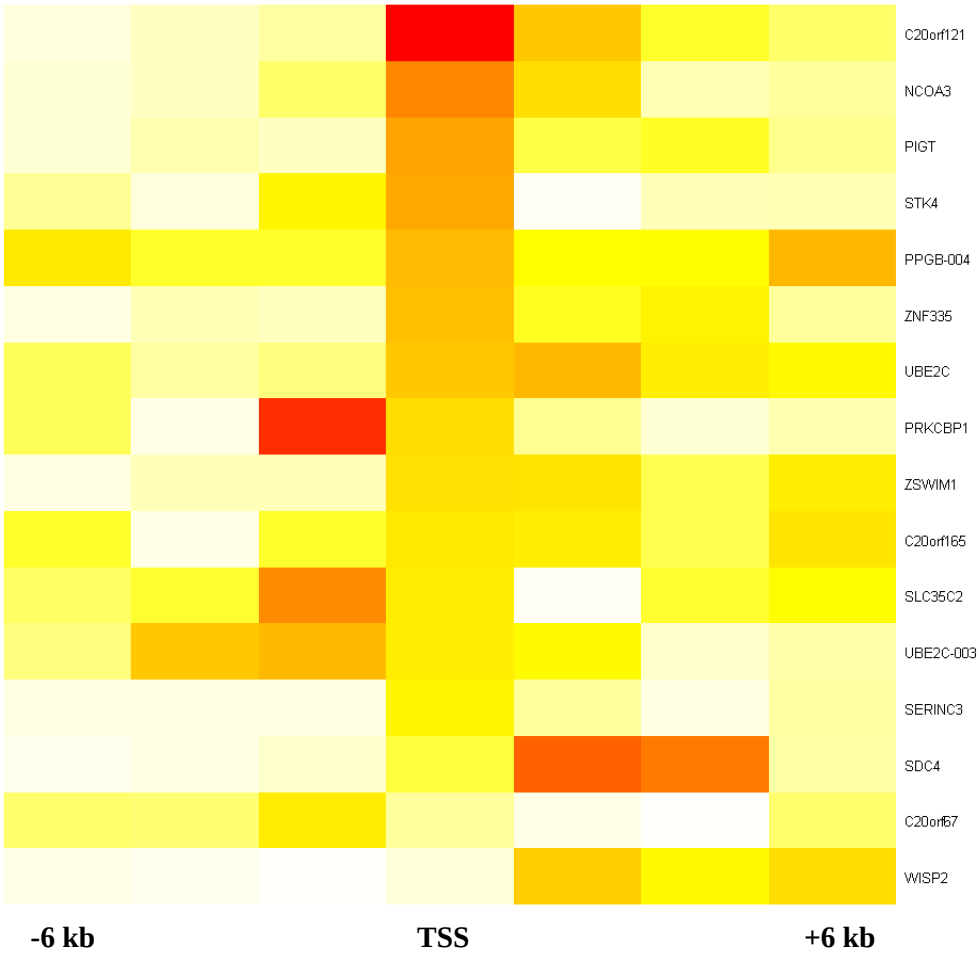


Figure 5.15 Heat map which displays spot intensities within ±6 kb distance of 16 coding transcripts in HeLa S3 cells. Highest signal is shown as red while the lowest signal is denoted as white. Representative transcripts are labelled with gene name only.

The heat map that displays the spot intensities of all 79 transcripts (corresponding to 66 genes) represented on the array can be found in Figure A2 (Appendix A).

Figure 5.15 shows *PPGB* among the positives. Note that the 5' ends of *NEURL2* and *PPGB* were contained in the same spot. Since *NEURL2* is not expressed but *PPGB*

has a very high expression according to the Affymetrix Expression Arrays in HeLa S3 (Table A5, Appendix A), the observed enrichment is attributed to *PPGB* only.

Figure 5.15 shows that the phosphorylated form of RNA polymerase II recruits itself on the annotated TSS of ten transcripts, although there are exceptions at which a higher signal is observed on either upstream (PRKCBP1-007, SLC35C2-006 and UBE2C-003) or downstream (PPGB-001, SDC4-001 and WISP2-002) of the TSS. The higher signals upstream of the UBE2C-003 transcript is the signal of the upstream representative transcript of the gene (UBE2C in Figure 5.15). In the case of PRKCBP1-007 (its TSS is located at the far end of the spot) and SLC35C2-006, the higher signals were produced mostly by the adjacent spots which span 2 kb of the upstream region of these genes. However, in the case of SDC4-001 and WISP2-002, the neighbouring spots which carry the first intron of these genes showed higher signal than the spots carrying the TSS of these genes. This variation might be due to the annotated TSS being imprecise. This possibility will be discussed in later sections by comparing the enrichment levels by different modified histones on these spots.

The annotated TSS of *KCNK15* is not present on the array. However, the spot carrying the neighbouring 2.2 kb fragment which spans the first intron (its genomic coordinates are 42,809,326..42,811,520) gave a 3.3 fold enrichment. Also, the *KCNK15* is expressed in HeLa S3 cells according to the Affymetrix Expression Arrays which supports the fact that *KCNK15* is active in HeLa S3 cells.

There is one polII enriched spot (42,754,161 to 42,755,685 bp), which does not contain any potential TSS in its close proximity. The region does not contain any known micro RNA nor any transcriptionally active regions (TARs) reported by high density oligonucleotide tiling arrays (Bertone et al., 2004). On the other hand, this region was found to be enriched with mono and di-methylated histone H3 at K4 (~4-5

fold). Therefore, it will be discussed in detail in section 5.7.1 for their possible regulatory functions.

The other polII enriched spot (located between 42,813,344 and 42,816,090) contains 3' UTR end of *RIMS4*, which is not expressed and its TSS is not enriched with polII HeLa S3 cells. The same spot also showed an enrichment with mono-methylated K4 of histone H3 (~1.75 fold) and will be discussed in section 5.7.1. The polII spot located between 42,469,627 and 42,471,927 bp lies within the *HNF4A* and will be discussed in 5.10.

In summary, gene initiation activities of 15 genes (16 transcripts) were detected by ChIP experiment using the polII antibody and 13 of them are expressed in HeLa S3 cells according to Affymetrix Expression Arrays. The *ZNF335* and *C20ORF165*, which gave an enriched signal but are not expressed may correspond to transcripts with post-transcriptional regulation. There are 15 more genes that are expressed in HeLa S3 cells but no polII enrichment was detected. This could be simply due to suboptimal working experimental conditions. Alternatively, polII initiation complexes on the TSSs of these genes may be short-lived which inevitably results in lower or no signal from these sites by ChIP. Another possibility would be that the epitope of polII recognized by the antibody is competed *in vivo* by other proteins hindering the detection of initiation complexes with this antibody.

It is important to note that all polII enriched start sites, except that of *SLC35C2*, were also enriched with tri-methylated histone H3 (H3K4me3) at K4 and acetylated histone H3 (H3Ac) as well. Note that these are the epigenetic markers found mainly on the start sites of actively transcribed genes (Bernstein et al., 2005) (see sections 1.4.4 and 1.4.5).

5.5.2 Results in NTERA-D1 cells

ChIP experiments with the polII antibody in NTERA-D1 cells resulted in 89 enriched spots. Of these, 42 were located within ±2kb of the TSS of an annotated coding transcript. While seven of the remaining 47 enriched spots were within intergenic regions, the rest of the enriched spots were contained in the intra-genic regions of active genes.

In NTERA-D1, annotated TSS of 17 coding transcripts (representing 14 genes) gave an enrichment and the heat map of these transcripts is shown in Figure 5.16. Note that all these polII enriched start sites showed an enrichment with H3K4me3 and H3Ac as well. The spot containing *C20orf67* did not produce any signal on a specific batch of the arrays used (marked as white in Figure 5.6), however, the adjacent spot (end of this spot is 400 bp upstream of the TSS of *C20orf67*) gave around 3 fold enrichment which attributed to *C20orf67* promoter activity.

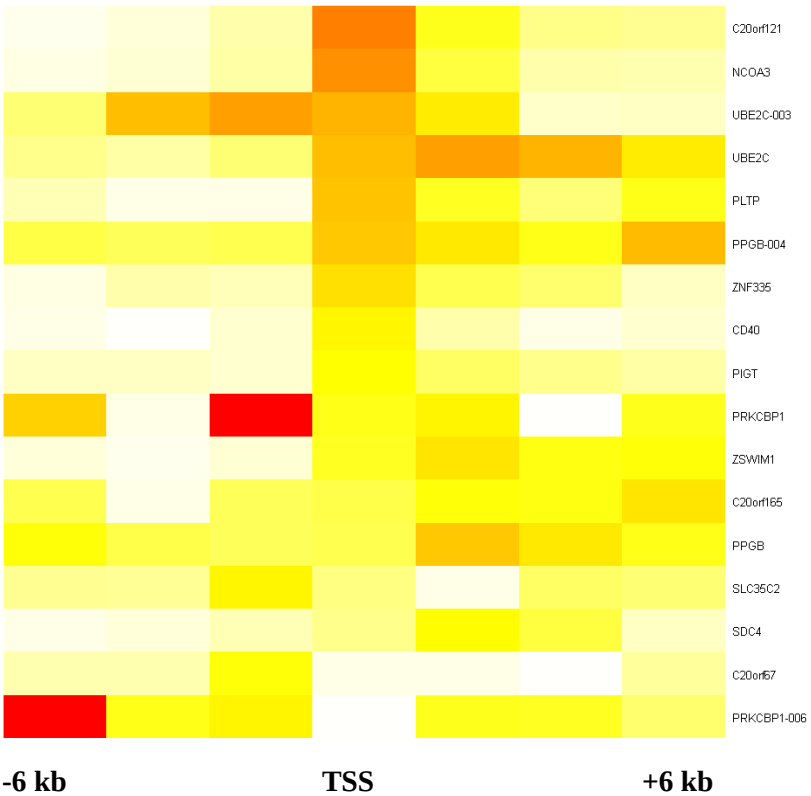


Figure 5.16 Heat map which displays spot intensities within ±6 kb distance of 17 coding transcripts in NTERA-D1 cells. Highest signal is shown as red while the

lowest signal is shown as white. Note that representative transcripts were denoted by gene name only.

The heat map that displays the spot intensities of all 79 transcripts (corresponding to 66 genes) represented on the array is given Figure A2 (Appendix A).

Figure 5.17 shows the enrichment levels of *PRKCBP1* together with its annotation (Ensembl Genome Browser; version 39).

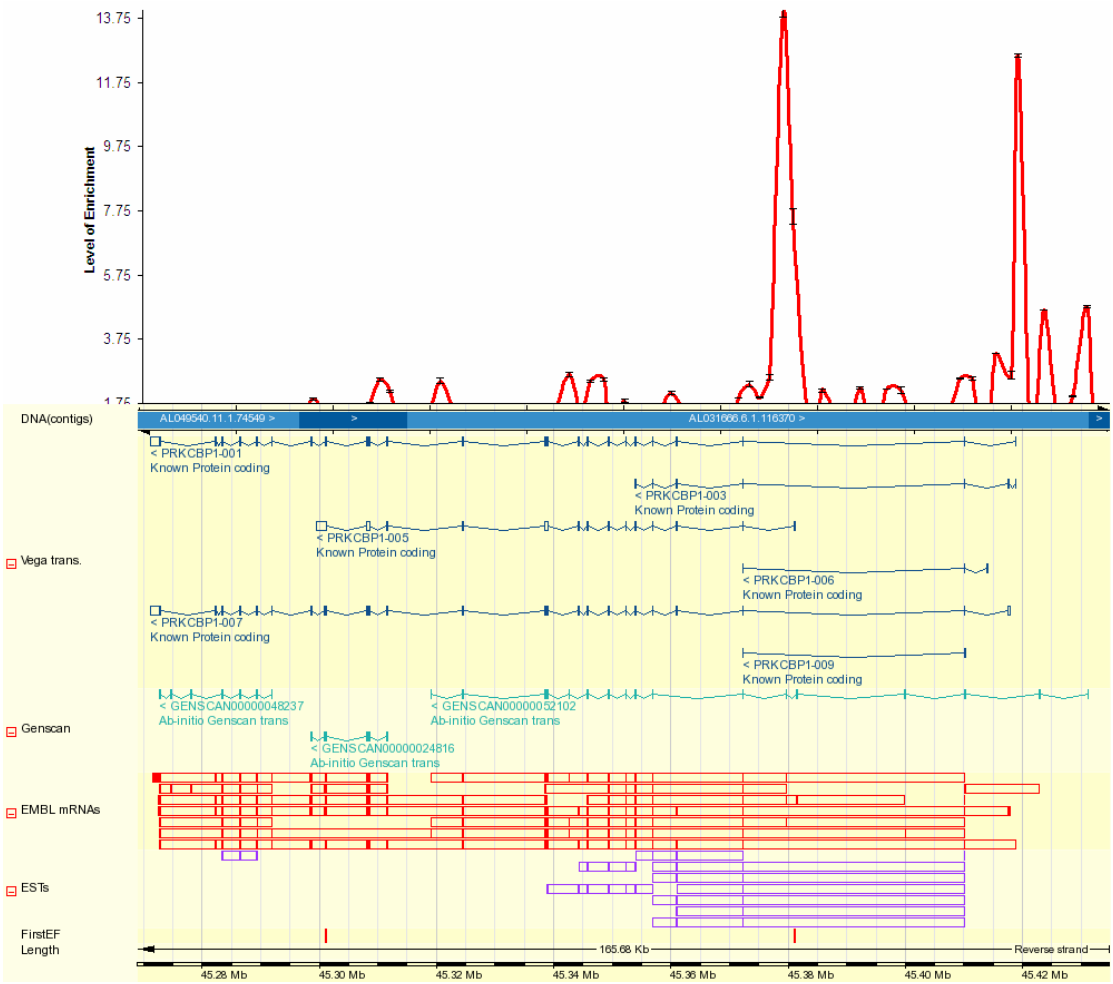


Figure 5.17. Enrichment levels (only signals above threshold were shown) on PRKCBP1 gene shown together with the annotation tracks reproduced from Ensembl Genome Browser. Track information is displayed on the left hand side of the annotation window.

The region downstream of the annotated TSS of the PRKCBP1-005 transcript showed 13.5 fold polII enrichment, while the proximal promoter region of the *PRKCBP1* (representative transcript) showed ~12 fold enrichment. Interestingly, spots carrying the annotated TSSs of these transcripts showed only ~2 fold polII enrichment. Also,

there are two polII-enriched region (4.2 fold) at around 2 kb and 10 kb upstream of the annotated start site of *PRKCBP1* representative transcript. The +10 kb region is also enriched with H3K4me3 (53 fold) and H3Ac (3.8 fold). This region contains the 5' end of a gene prediction (GENSCAN00000035740). Further, there is a human spliced EST (CN335160, UCSC Genome Browser) from embryonic stem cells. The above strongly supports an alternative *PKRCBP1* transcript whose 5' end is 10 kb further upstream of the currently annotated start site. None of these upstream peaks were observed in HeLa S3 cells which suggests that the putative alternative transcript described above is tissue-specific.

All 14 genes except *C20orf121*, *NCOA3* and *CD40* showed polII enrichment across their gene sequence while these three genes specifically showed polII enrichment on their annotated start site.

As mentioned earlier, there are seven polII enriched spots that do not contain by any gene or lie within 2 kb of a start site of a genic feature. Two of them are located upstream of *PKRCBP1* and explained above. Another one is located between 43,354,154 and 43,355,325 bp, and showed a 2.5 fold polII enrichment. This region also showed a 40 fold enrichment with H3K4me3, and a 4 fold enrichment with H3Ac. It is adjacent to the 3'end of the MATN-4 gene, which did not give any polII enrichment on its start site and is not expressed in NTERA-D1 according to Affymetrix Expression Arrays. Within this enriched spot, there is a FirstExon promoter prediction (on reverse strand) and a CpG island (692 bp) immediate upstream of the promoter prediction. Additionally, it also contains the candidate first exon of a human spliced EST (DB444499, UCSC Genome Browser) found in testis.

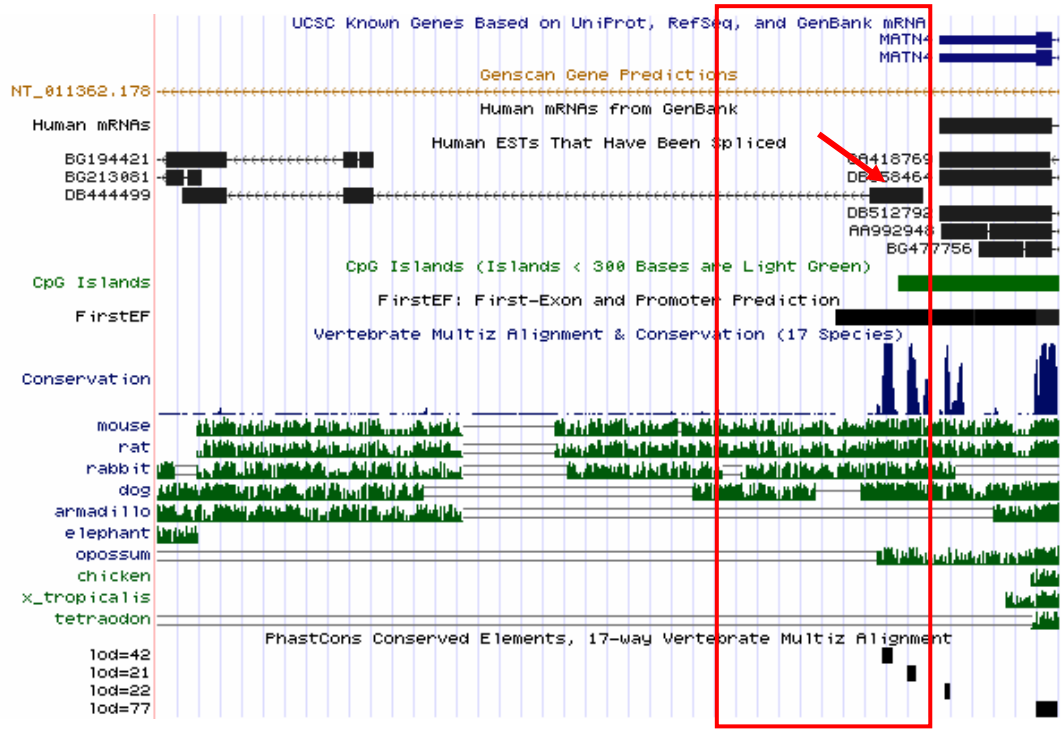


Figure 5.18 The annotation of the an polII enriched region (43,354,134..43,355,325 bp on chromosome 20) reproduced from UCSC Genome Browser. The red squared denotes the boundaries of the enriched spot (3.8 fold polII and 40 fold H3K4me3 enrichments). Track information is displayed on the left hand side of the annotation display.

Figure 5.18 shows feature annotations (UCSC Genome Browser) mapped onto this polII enriched spot (in red square). Combined, the presence of a promoter prediction, a CpG island and a spliced EST, strongly supports the existence of a new gene, which is tissue-specific (not present in HeLa S3).

There are four adjacent spots (located between 43,821,287 and 43,829,560 bp) which gave polII (~5 fold) as well as H3K4me3 (~11 fold) enrichment, and they do not coincide with any genic feature. There is no evidence such as gene or promoter predictions or high sequence conservation across species, yet this polII enrichment may not be an artefact. Further experimental testing is required but none of these regions contained any known micro RNA, or any TARs.

In summary, out of 35 genes expressed in NTERA-D1, 5' ends of 14 were enriched with RNA polymerase II. TSSs of two non-expressed genes (*C20orf165* and *WFDC10A*) were enriched with polII (2- and 1.9 fold respectively).

The genes, *PLTP*, *WISP2* and *ZNF334* were enriched by polII only in NTERA-D1 but not in HeLa S3 cells, and these three genes were only expressed in NTERA-D1 cells. Therefore, it was possible to detect expression profile differences of the genes in ChIP experiments in contrast to the gene reporter assays where there was less correlation between the promoter activity of the gene and its expression status (section 4.5.3.2). This result is expected as ChIP experiments can successfully capture the genomic environment of the genes while gene reporter assays only mimic the trans regulatory content of the gene.

SERINC3 and *CD40* are expressed in both cell lines, but *SERINC3* gave polII enrichment only in HeLa S3 cells, and *CD40* only in NTERA-D1 cells. This may indicate a different mode of gene regulation depending on the cell type. Where there is no ChIP signal, polII complexes might be hindered by other proteins and its epitope not accessible for the antibody in one cell line, although in the other cell line, there might be another set of proteins used to activate the gene that might not have the same hindrance effect. However, the fact that in each cell line circa 50% of the expressed genes did not give a polII enrichment favours the explanation of sensitivity.

It is well established that actively transcribed genes are decorated by a set of specifically modified histones (Allfrey et al., 1964) (Litt et al., 2001). As mentioned, H3Ac and H3K4me3 are the most common markers found on actively transcribed genes. Below, I will discuss the epigenetic markers involved in transcriptionally active genes and correlate the activating histone code with the observed polymerase II activity described so far.

5.6 Histone Modifications on Transcription Start Sites

ChIP experiments were performed with antibodies recognising histone H3 tri-methylated at K4 position (H3K4me3) and histone H3 acetylated on lysine 9 and 14

(H3Ac) in both HeLa S3 and NTERA-D1 cells. The ChIP samples were characterized using the custom 3.5 Mb tile-path array.

5.6.1 Results in HeLa S3 cells

5.6.1.1 H3K4me3

An antibody recognizing H3K4me3 was used to locate transcriptionally active promoter regions. The H3K4me3 polyclonal antibody (Abcam, #ab8580) was raised in rabbit and it recognizes tri-methylated K4 residues on histone H3 proteins. This antibody has a weak reactivity to di-methylated K4 of histone H3 but it has no cross reactivity with any modified form of K9 of histone H3.

A lower threshold value (1.50) was used to decide whether a spot enriched or not by H3K4me3 since the relative distribution of the background and actual signal with this antibody differs from that obtained with the polII antibody. On the array 50 spots gave enrichment with H3K4me3 antibody in HeLa S3 cells. Of those, 42 were within 2 kb of annotated TSSs of alternative transcripts representing 22 genes. Out of eight enriched spots not associated with any TSS, two were not considered as true positives due to possible cross-hybridization, since they had over 90% sequence identity with ubiquitously expressed genes elsewhere in the genome. One of the spots (located between 44,159,993 and 44,162,444 bp) contains a processed pseudogene (*RPL13P2*), which has a 90% sequence identity with ubiquitously expressed *RPL13* ribosomal protein gene. The other spot (located between 45,817,478 and 45,819,561 bp) does not contain any genic feature but it has 85% sequence identity with *GTFIIIC* (general transcription factor IIIC polypeptide I), a ubiquitously expressed gene required for RNA polymerase III transcription. The remaining six enriched spot will be discussed in further sections since they gave higher signals with other antibodies used in this study.

The heat map of H3K4me3 signals within ~6 kb upstream and downstream of the annotated start sites of the 23 positive transcripts (representing 22 genes) is given in Figure 5.19.

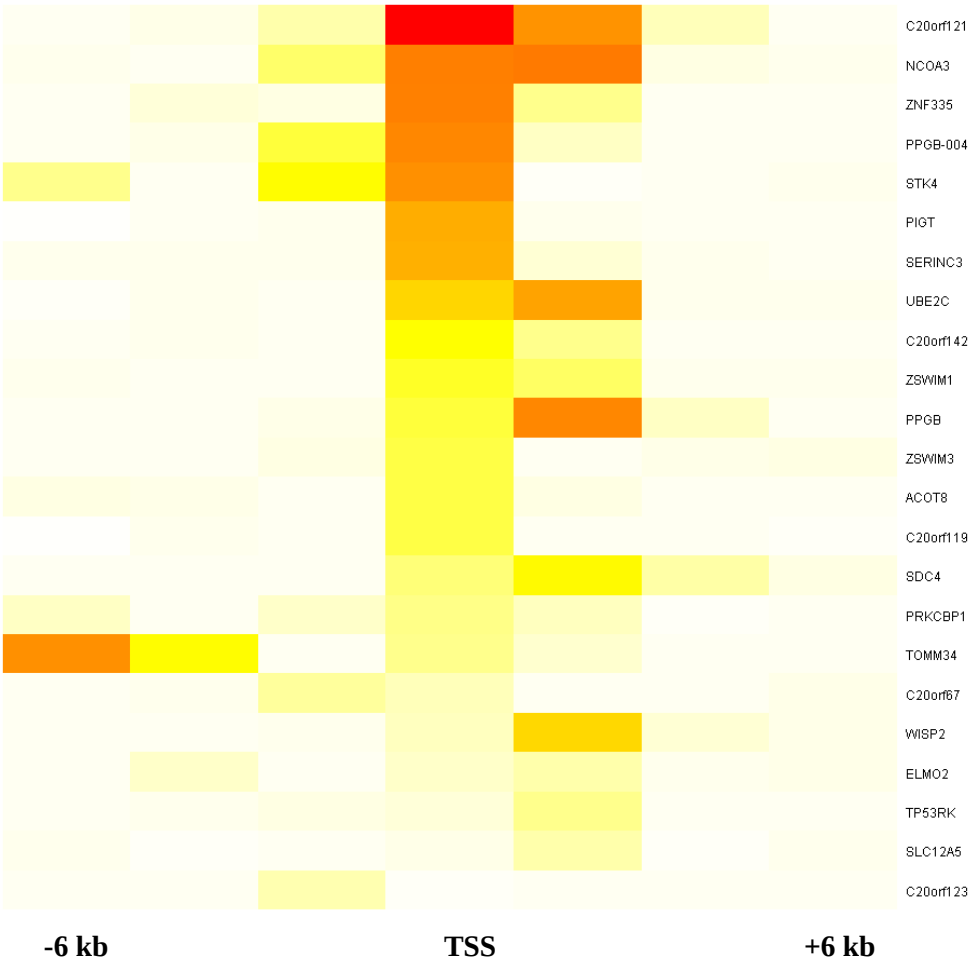


Figure 5.19 The heat map displaying ~6 kb upstream and downstream of the annotated start sites of 22 genes that showed an H3K4me3 enrichment around their TSS.

Also, Table 5.4 shows the enrichment levels of H3K4me3 enriched start sites of the 23 transcripts (representing 22 genes) by other 8 antibodies used in this study. Out of 28 genes expressed in HeLa S3 according to Affymetrix Expression Arrays, 20 were enriched with H3K4me3 around their TSS, and out of these 20 H3K4me3 enriched genes, 13 were enriched with RNA polymerase II as well.

C20orf62 gave an enrichment on the spot carrying its TSS, which also contains a pseudogene (*RPL37AP1*) that has 96% sequence identity (across 700 bp) to a

ubiquitously expressed ribosomal protein (*RPL37A*). Since *C20orf62* gene is not expressed in HeLa S3 cells, this enrichment was attributed to a cross hybridization, therefore it is not included within the active gene set.

The start site of *SLC12A5*, which is not expressed in HeLa S3 cells was enriched with H3K4me3. Interestingly, the H3K4me3 enrichment was observed downstream rather than on the actual TSS. The downstream spot is also enriched with H3Ac (Table 5.4). *SLC12A5* is not enriched anywhere across its sequence with polII. Also, its core promoter region did not show any activity in the gene reporter assays (section 4.5.1). In NTERA-D1 cells, its start site as well as its downstream region are enriched with H3K4me3 and H3Ac, but also with H3K27me3, an epigenetic marker associated with silenced genes (Cao et al., 2002). Hence, the enrichment profile across this gene will be discussed in detail in section 5.10.

There are five genes expressed in HeLa S3 (*TOMM34*, *ACOT8*, *C20orf142*, *ELMO2* and *TP53RK*) that gave an H3K4me3 enrichment but no polII enrichment on their TSS. Reporter assays for *ACOT8*, *ELMO2* and *TP53RK* genes gave strong core promoter activity (section 4.5.1). This indicates that the lack of enrichment by polII on the start site of these genes is most likely due to experimental limitations as explained in section 5.5.1, since detecting a promoter activity by reporter assay confirms the lack of any dominant trans-acting silencing on the promoter.

Histone H3 K9 di-methylation (H3K9me2) and K27 tri-methylation (H3K27me3) are associated with heterochromatic regions and silencing (Shilatifard, 2006) and will be discussed in section 5.8. However, it is worth mentioning here that it is not surprising to see no H3K4me3 enriched region being enriched with either of these two epigenetic markers in HeLa S3 cells.

Index	HUGO Gene ID	polII	H3K4me	H3K4me2	H3K4me3	H3Ac	H4Ac	CTCF	Expression
1	ACOT8		2.0d	3.5	6.3	2.2			P
2	C20orf119			4.6	6.2	2.9			P
3	C20orf121	7.4	3.3d	18	36.7	14.4	6.5		P
4	C20orf123				2.4u				No probe
5	C20orf142			4.6	9	3.3			P
6	C20orf67	2.2u		2.2u	2	1.6			P
7	ELMO2		2.9d	2.1d	2.6d	1.6d	1.6u		P
8	NCOA3	4.4	0	6.7	22.8	11.4	3.4		P
9	PIGT	3.8	0	7.3	17.9	7.9	3.7		P
10	PPGB (RT)	1.39*	2.70	4.59	6.62	2.84	2.55		P
11	PPGB (AT)	3.2	2.9	14.3	21.8	9.7	2.5		P
12	PRKCBP1	2.5	3.1	7.2	3.9	2.6	3.3		P
13	SDC4	5.3d	2.5d	2.9	4.5	2.1			P
14	SERINC3	2		7.3	17.5	5.7	1.9		P
15	SLC12A5				2.6d	2.2d			A
16	STK4	3.7	1.7	9.2	21	12.1	2.5		P
17	TOMM34		2.7d	4.2d	3.7	2.1			P
18	TP53RK		2.1d	4.4d	3.6d	3.7d			P
19	UBE2C	3	2.0d	4.2	13.4	2.9			P
20	WISP2	2.9d	4.5	2.7	1.9	2.5	2		P
21	ZNF335	3.2	3.4d	9.1	22.7	12.9	6.7	4.7d	A
22	ZSWIM1	2.4	1.9d	3	7.4	3.3	1.7		P
23	ZSWIM3		2.0u	3.5	6.3	2.2			P

Table 5.4. Enrichment levels of seven antibodies used in this study for 23 transcripts representing 22 genes in 3.5 Mb region in HeLa S3 cells. “u” marks a signal coming from ~2kb upstream of the annotated start site of the region and “d” marks a signal coming from ~2 kb downstream of the annotated start site of the region. The “expression” column displays the expression status of the corresponding gene; “A“ (Absent) stands for no expression while “P” (Present) means the gene is expressed. * The polII enrichment on this gene is reported although it is below the selected threshold. H3K27me3 and H3K9me2 antibody columns are omitted since none of the spots showed any enrichment with these antibodies.

Note that none of the polII enriched spots, which have no association to any gene annotations, showed enrichment with H3K4me3 (see section 5.5.1). However, they were enriched with other modified histones, H3K4me and H3K4me2 and will be discussed in section 5.7.1.

As described earlier, the TSS of *KCNK15* is absent but the neighbouring spot spanning 2.2 kb of the first intron (42,809,326..42,811,520 bp) showed 3.3 fold enrichment for polII. Likewise, there is 3.7 and 3.6 fold enrichment with H3K4me3 and H3Ac antibodies respectively, which also supports the fact that the gene is actively transcribed in HeLa cells.

There is another H3K4me3 enriched spot (~1.8 fold) carrying the second 5' exon *SULF2*. The annotated start site of this gene did not give any enrichment above threshold with any antibody. In NTERA-D1 cells, there is a strong H3K4me3 enrichment on the actual TSS as well as the second 5' exon. There is also a promoter prediction (FirstExon) around the second 5' exon. These observations combined might indicate a tissue specific alternative TSS for this gene.

5.6.1.2 H3Ac

ChIP experiments in HeLa S3 cell were performed with the antibody (Upstate #06-599), which recognizes histone H3 acetylated at lysine 9 and 14. This polyclonal antibody is raised in rabbit and has no cross-reactivity with other modified histones. There are 60 spots on the array enriched with H3Ac. Out of these, 52 were within 2 kb of a TSS of 17 transcripts (17 genes) in the region. Figure 5.20 shows the heat map displaying the signals within 6 kb upstream and downstream of these transcripts.

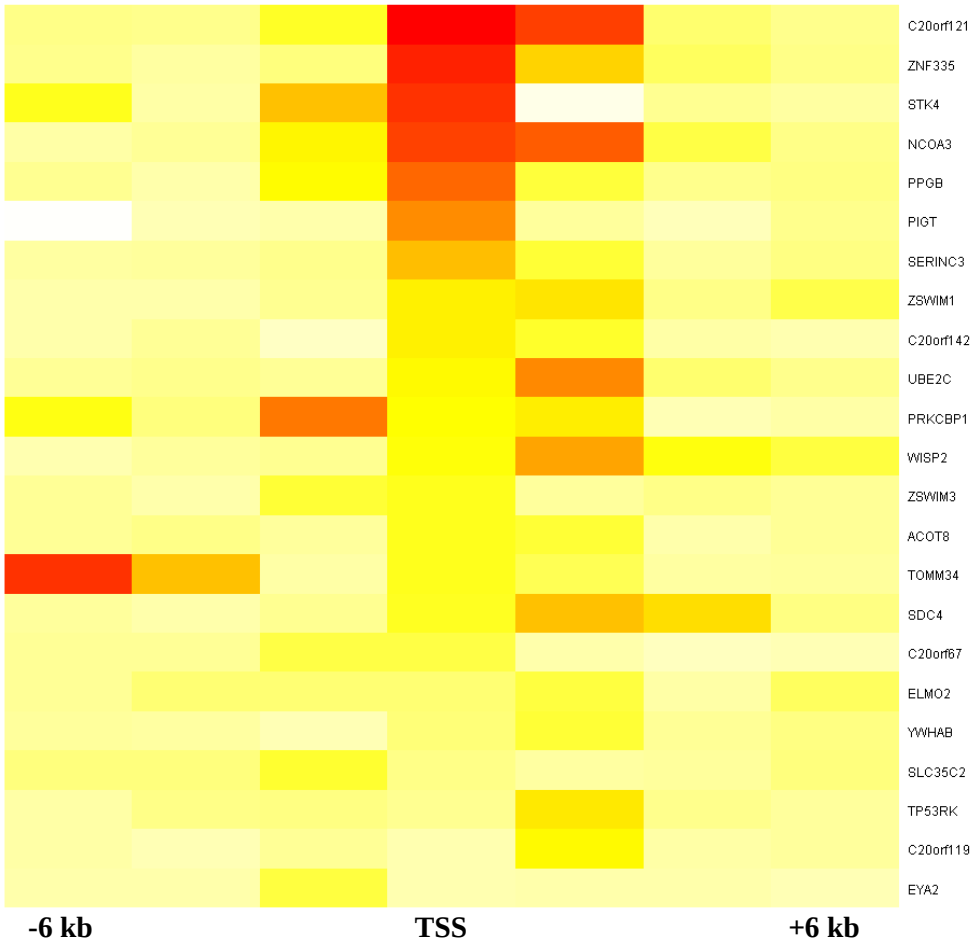


Figure 5.20 Heat map displaying the signals around 6 kb upstream and downstream of the annotated start sites of 17 H3Ac enriched transcripts in HeLa S3 cells. Higher signal is indicated with colours towards red while lower signals would be towards white.

Unlike H3K4me3 enrichment signals where the highest signal is mainly located on the TSS spot (see Figure 5.19), histone acetylation seems to be more widespread with a trend toward downstream sequences. This may mean that the factors which play a role in accurate positioning of the initiation machinery require H3K4me3, while histone 3 acetylation is needed for recruiting the initiation machinery around the start site.

Out of 24 H3K4me3 enriched start sites, only one was not enriched with acetylated histone H3. This finding is in agreement with previous studies (Kim et al., 2005) (Bernstein et al., 2005) where the acetylated histone H3 is found in 90-99% of H3K4me3 enriched genes. Also, it is shown that MLL, which is responsible for

histone H3 methylation at K4 residues is stimulated by acetylated histone H3 peptides (Milne et al., 2002) (Nakamura et al., 2002). Therefore, histone H3 acetylation might play an important role in initiation of histone H3 tri-methylation at K4.

C20orf123, which showed H3K4me3 but no H3Ac enrichment on upstream of its start site did not give any enrichment with other antibodies either.

The remaining eight sites that gave enrichments with H3Ac gave higher enrichments with other antibodies, therefore the features of these sequences will be discussed in 5.7.1.

5.6.1.3 H3K4me and H3K4me2

In 5.6.1.1, I showed that tri-methylation of histone H3 at K4 on the TSS is a feature of transcriptionally active genes. To find out whether mono and di-methylation of histone H3 at K4 shows a similar pattern, ChIP experiments were carried out as before with polyclonal antibodies Abcam #ab8895 and Abcam #ab7766, which recognize the H3K4me and H3K4me2 form respectively. Both antibodies are raised in rabbit and do not have cross reactivity to histone H3 methylated at K9. The mono- and di-methylation profiles of the 6 kb distances of the start sites of genes enriched with H3K4me3 are displayed in Figure 5.21 which shows that tri-methylation is more centred on the actual TSS whereas there is virtually no tri-methylation in the flanking sequences. As mentioned earlier, this may indicate that this modification plays a role in fine positioning of the polII transcription initiation complex on the actual start site (see section 5.6.1.2).

High H3K4me2 enrichment was observed on H3K4me3 enriched start sites (upper half of the H3K4me2 heat map in Figure 5.21). TSSs that are not strongly tri-methylated are relatively more enriched with H3K4me2. This may be due to the fact

that H3K4me2 is an intermediate state to H3K4me3. The H3K4me2 type of enrichment pattern was also observed with H3K4me.

As shown in Table 5.4, only two H3K4me3 enriched start sites (of *C20orf123* and *SLC12A5*) were not enriched with either H3K4me2 or polII. Yet, these genes are not expressed in HeLa S3 cells. This observation supports the idea that di-methylation of histone H3 at K4 on TSSs can be also seen as a mark for active genes. Again H3K4me2 is an intermediate to H3K4me3. However, as will be discussed in later sections (5.6.1.3 and 5.7.1), H3K4me2 can exist as an independent epigenetic functional marker from H3K4me3 in the genome. Histone H3 mono-methylation (H3K4me) was observed in 60% of the H3K4me3 enriched start sites. H3K4me is the first step in the methylation process of histone H3, so the presence of H3K4me on start sites is to be expected. On the other hand, HK4me has recently been found in distant regulatory sequence elements (Ren, B., unpublished data).

In summary the above results suggest that promoter regions are clearly marked with tri-methylated or di-methylated histone H3 at K4, whereas mono-methylated histone H3 at K4 does not have the same discriminatory power.

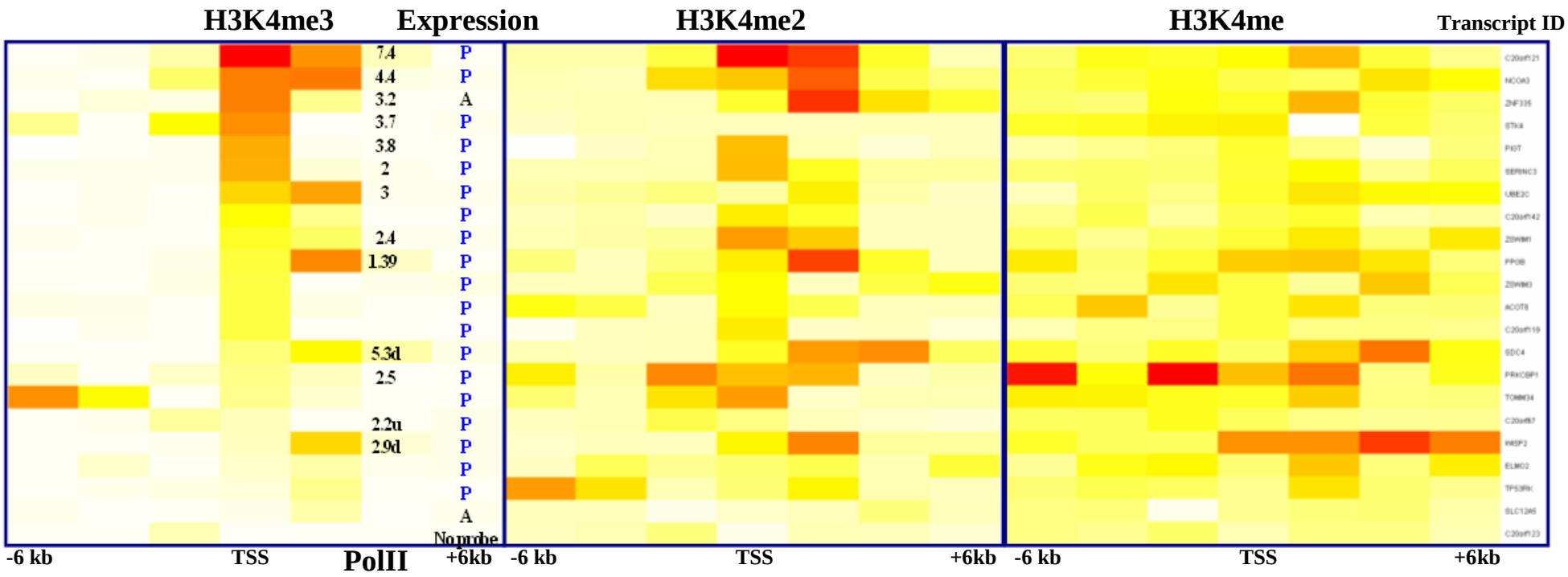


Figure 5.21 H3K4me3, H3K4me2 and H3K4me profiles of 22 genes that showed H3K4me3 enrichments on their start sites in HeLa S3 cells. Each profile is presented as a heat map which displays the signals obtained within ~6 kb distance of TSSs. The first column on the left lists the polII enrichments and the sec column lists the expression profiles ("P" stands for the gene is expressed and "A" denotes no expression) for the corresponding genes. "d" and "u" means that the signal is detected at ~2 kb downstream or upstream of TSS respectively.

5.6.1.4 H4Ac

An antibody recognizing the acetylated form of histone H4 at lysine 5,8,12 and 16 was employed to determine genomic sites enriched with this modified histone type. The antibody, Upstate #06866 is raised in rabbit and has a weak cross-reactivity with acetylated histone H3. There were 60 spots on the array that showed enrichment with this modified histone but only 19 were also enriched with H3K4me3. Only half (53%) of the H3K4me3 enriched sites were also enriched with H4Ac. This is expected as histone H4 acetylation plays a major role in chromatin structure changes and protein interactions (Shogren-Knaak et al., 2006).

A nice illustration of how the above ChIP results can be combined to improve the current annotation comes from the following example. A spot containing the TSSs of four non-coding transcripts of *DBNDD2* (*C20ORF35*) (DBNDD2-007, -009, -011 and -013) gave an 8.8 fold H3K4me3 enrichment. Interestingly, this spot also carries the start sites of three coding transcripts of *C20orf169*. However, the gene record for *C20orf169* has been removed from the Entrez Gene database

(http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene) since, except its last exon, all other exons are shared with the *DBNDD2*, and the mRNA evidence supporting *C20orf169* gene is used in the annotation of *DBNDD2*. The annotation of the region is shown in Figure 5.22.

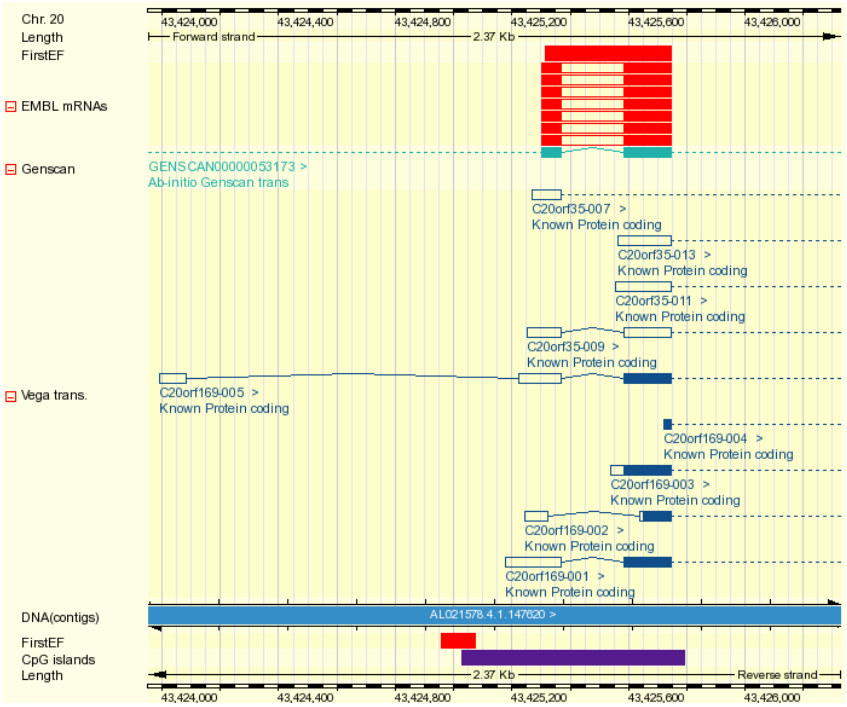


Figure 5.22. Ensembl annotation of the spot (its genomic coordinates 43,423,952..43,426,324 bp) that showed a 8.8 fold H3K4me3 and 5.3 fold H3K4me2 enrichments. It carries the start site of four non-coding transcripts of *DBNDD2* (*C20orf35*). It also carries the first exon of 5 coding transcripts of the *C20orf169* whose record was removed from the Entrez Gene Database.

The fact that this region gave a significant 8.8 fold enrichment for H3K4me3 as well as H3K4me, H3K4me2 and H4Ac enrichments, there are two possible explanations: either, the annotated non-coding transcripts of *DBNDD2* are transcriptionally active or *C20orf169* actually is a real gene. *DBNDD2* is “expressed” in HeLa S3 cells but none of the start sites of its coding transcripts showed enrichment with the antibodies used in this study. Given that the spot in question contains a promoter prediction (FirstExon) and a CpG island, I believe that it is necessary to perform further experiments to reassess annotation of the region and re-instate the *C20orf169*, which is most likely a true gene.

5.6.2 Results in NTERA-D1 cells

5.6.2.1 H3K4me3

There were 83 H3K4me3 enriched spots in the ChIP experiments carried out in NTERA-D1 cells. Of these, 73 carried sequences within 2 kb distance of a TSS. Three spots carrying pseudogenes were eliminated due to possible cross hybridization with ubiquitously expressed genes elsewhere in the genome. Other enriched spots that are not close to any TSS were already discussed in section 5.5.2.

Out of the 66 genes whose TSS was represented by at least one spot on the array, 30 showed H3K4me3 enrichment. Of the 35 genes represented on the array and expressed in NTERA-D1 cells, 25 showed enrichment. The enrichment levels within 6 kb distance of the start site of the 30 genes are shown in Figure 5.23.

Table 5.5 lists the signals from the H3K4me3 enriched start sites together with the signals obtained with the other antibodies used in this study. Of the 30 genes enriched with H3K4me3, 13 were enriched with RNA polymerase II as well (Table 5.5). The four genes (*KCNS1*, *SLC12A5*, *WFDC10A* and *C20ORF165*) that are not expressed but whose start sites are enriched with H3K4me3, showed enrichments with either CTCF or H3K27me3, which play a role in silencing genes.

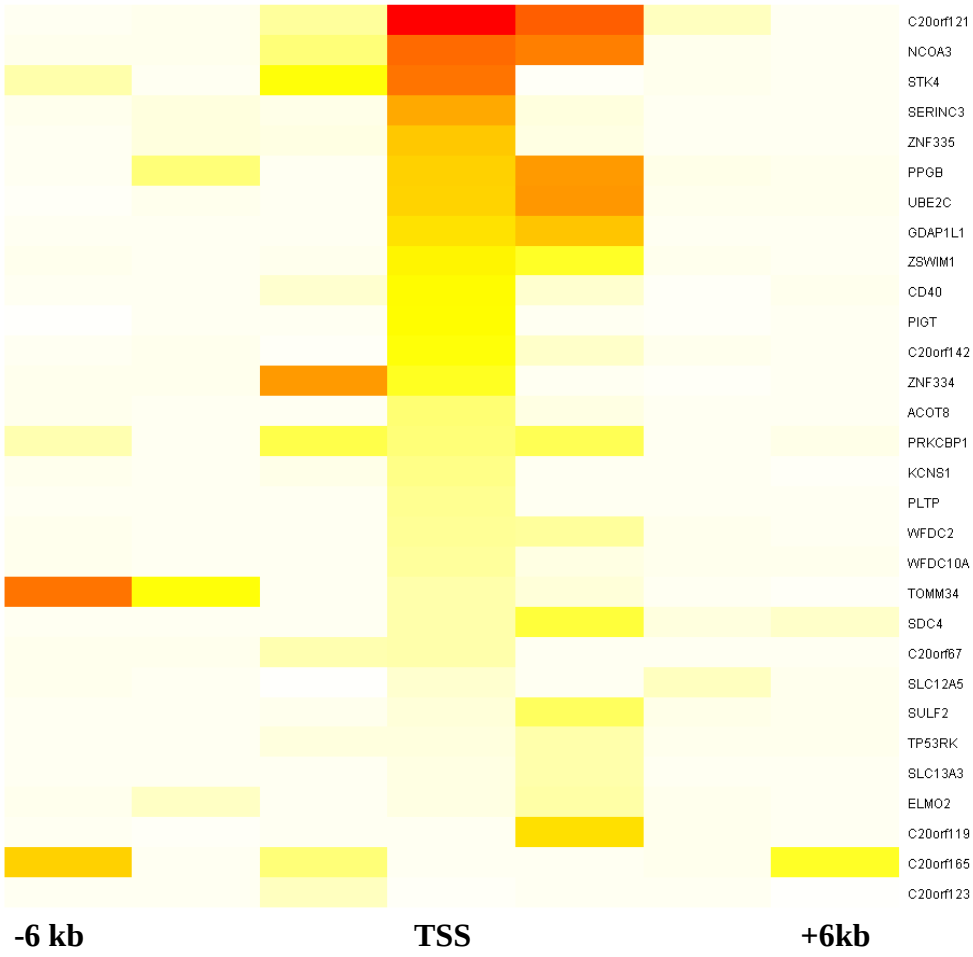


Figure 5.23 Heat map displaying the H3K4me3 enrichment levels of spots spanning 6 kb upstream and downstream sequences of 30 enriched TSSs of representative transcripts in NTERA-D1 cells.

Index	HUGO Transcript ID	polII	H3K4me	H3K4me2	H3K4me3	H3Ac	H4Ac	H3K27me3	CTCF	Expression
1	ACOT8			4.02	10.05					P
2	C20orf119			5.26	26.42	2.76				P
3	C20orf121	5.15	1.8d	25.77	79.56	7.1				P
4	C20orf123			1.6u	3.91u					No probe
5	C20orf142			7.68	18.47	2.66				P
6	C20orf165	1.9d		5.8u	9.5u				2.2u	A
7	C20orf67	1.9u		3.93	5.61	1.55				P
8	CD40	2.18		12.59	19.97	1.74				P
9	ELMO2			1.62	6.02d					P
10	GDAP1L1			14.25	26.25	3.4			3.3d	P
11	KCNS1			7.24	8.33			0	3.20u*	A
12	NCOA3	4.75		16.28	54.36	5.18				P
13	PIGT	1.98		7.57	19.6	2.2				P
14	PLTP	3.49			7.49	1.79				P
15	PPGB	3.34		21.91	42.93	4.79				P
16	PRKCBP1	1.75	2.5	7.75	9.39	2.37	3.46			P
17	SDC4	2.1d		4.21	5.65	1.84				P
18	SERINC3			12.24	39.36	3.89				P
19	SLC12A5			8.55	2.54			9.52		A
20	SLC13A3			7.6	5.475					P
21	STK4			18.35	51.9	4.65				P
22	SULF2			9.4d	1.91	1.91d				P
23	TOMM34			5.5d	5.7					P
24	TP53RK			1.67	1.51					P
25	UBE2C	3.62		6.47	29.56	1.81				P
26	WFDC10A			7.99	6.89				4.5	A
27	WFDC2			7.4	7.29	1.51				P
28	ZNF334			9.6	42.93	4.21				P
29	ZNF335	2.74		11.5	31.97	3.43			3.9d	P
30	ZSWIM1	2.6d		4.53	21.51	1.66d				A

Table 5.5. Enrichment levels of 30 H3K4me3 enriched TSSs with other antibodies used in this study in NTERA-D1 cells. In the expression column "A" stands for no expression and P denotes that the gene is expressed in NTERA-D1 cells. "u" and "d" denote that the signal is detected 2 kb upstream or downstream of the TSS respectively. * This signals is placed around 4 kb upstream of the KCNS1 TSS.

There are three other genes (*KCNS1*, *WFDC10A* and *C20orf165*) that are not expressed in NTERA-D1 cells, yet they showed enrichments with H3K4me3 (but not with H3Ac) around their start sites. Interestingly, the TSSs of *WFDC10A* and *C20orf165* also showed enrichment with CTCF protein, which can act as a silencer, as well as the region 4 kb upstream of *KCNS1* (see section 5.9). CTCF can silence promoter regions of its target genes, and in the context of these three genes, it may function as a repressor since they are not expressed. Note that there is also a CTCF enrichment around 4 kb upstream of the *KCNS1* in HeLa S3 cells, while, unlike in NTERA-D1, there is no other H3K4me3 or polII enrichment on the gene in HeLa S3 cells and the gene is not expressed in HeLa S3 cells either. This indicates that if CTCF does function as a silencer on *KCNS1*, its recruitment is probably not dependent of the H3K4me3 status of the region since its binding is detected in two different chromatin environments.

The promoter region of *WFDC10A* is also enriched with CTCF protein in both cell lines. However, there is an H3K4me3 enrichment only in NTERA-D1 cells and this gene is not expressed in either cell line. This observation represents another case where CTCF binding is not dependant on at least the H3K4me3 status of the start site.

SDC4 showed a similar H3K4me3 and H3Ac enrichment pattern in both cell lines such that the higher enrichments were located around 2 kb downstream of the TSS (see Figure 5.23 and Figure 5.25). In NTERA-D1 cells, there is a second polII (~3 fold) and H3K4me3 peak (~12 fold) approximately 5 kb downstream of the start site (Figure 5.24). There is an mRNA (G43350, located between 43,407,194 and 43,407,343 bp), located 1 kb upstream of the second peak, which may represent an as yet un-annotated first exon. Except the two multi-species conserved blocks located within the spot giving the second peak, there is no other mRNA or EST evidence for

an alternative exon. Based on these observations, the 5' annotation of this gene needs to be inspected for a mis-annotated first exon or a tissue-specific second exon.

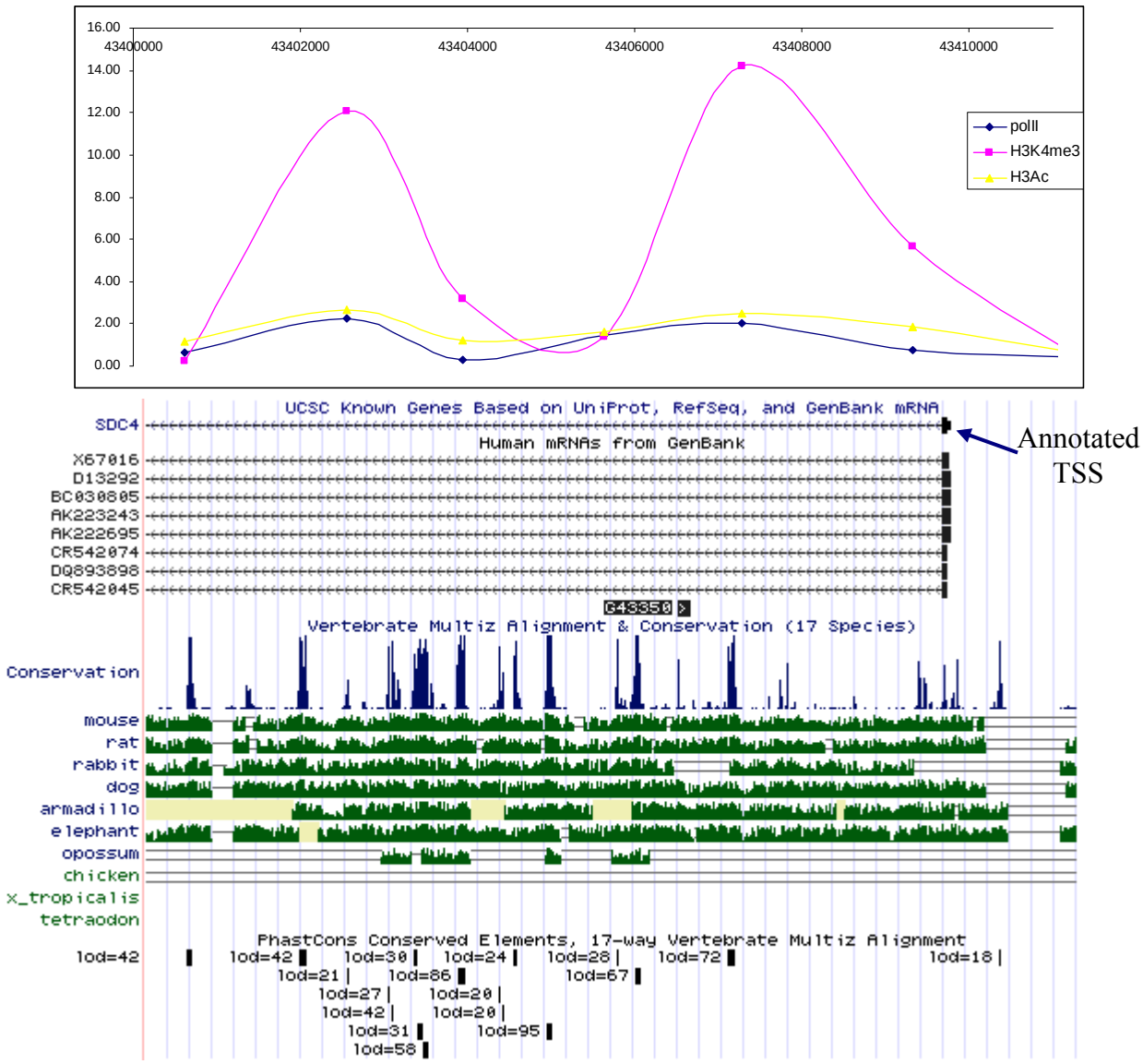


Figure 5.24 H3K4me3, polII and H3Ac peaks within the first 10 kb of SDC4 gene. The annotation is reproduced from UCSD Genome Browser.

As expected, except one case (*SLC12A5*), none of the H3K4me3 enriched sites gave enrichments with H3K27me3 or H3K9me2.

5.6.2.2 H3Ac

There are 50 spots that were enriched with acetylated histone H3 and 40 of them carry sequences that are within 2 kb distance of a TSS. Of the remaining 10, the profiles of

seven were discussed in the previous section since they coincided with H3K4me3 enrichments. Finally, the remaining three spots will be discussed in 5.7.2.

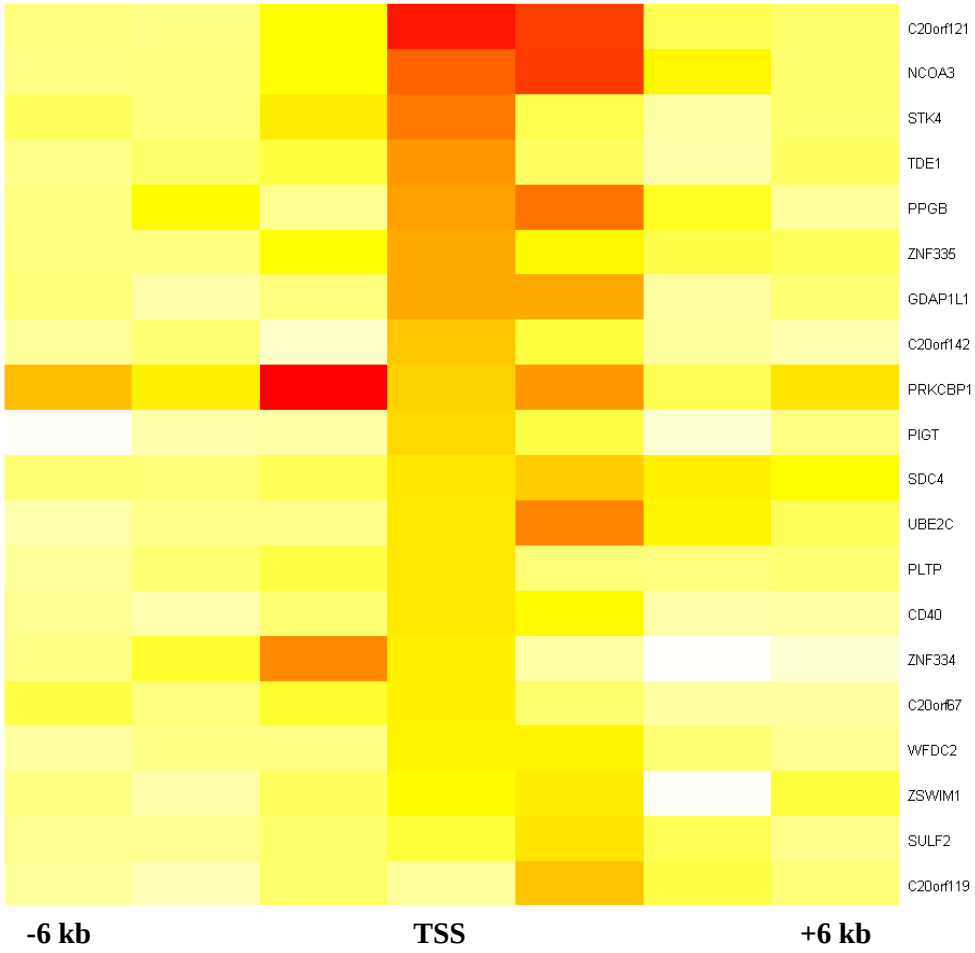


Figure 5.25 Heat map displaying the signals around 6 kb upstream and downstream of the annotated start sites of 21 H3Ac enriched transcripts in NTERA-D1 cells. Higher signal is indicated with colours towards red while lower signals would be towards white.

Start sites of 21 transcripts that showed H3Ac enrichment (40 spots) and the signals within 6 kb distance of these start sites are shown in Figure 5.25. Like in HeLa S3 cells, histone acetylation appears to be more widespread around the TSS whereas tri-methylation of histone H3 at K4 is mostly centred on the actual TSS (see section 5.6.1.2 and 5.6.2.1).

5.6.2.3 H3K4me and H3K4me2

The H3K4me and H3K4me2 enrichment profiles of the TSSs that are enriched with H3K4me3 in NTERA-D1 cells are shown in Figure 5.26. As observed in HeLa S3

cells (see section 5.6.1.3), H3K4me3 is very specific to the TSS, H3K4me2 is observed both on and immediately downstream of the start site. On the other hand, H3K4me does not appear to have such a location preference; the enrichment levels do not change greatly around the TSS. However, genes exhibit different H3K4me patterns such as the *PRKCB1*, which is enriched up to 6 kb either way of the start site, and *NCOA3*, *ZNF335* and *SDC4*, which are enriched with H3K4me across the downstream sequences (Figure 5.26).

5.6.2.4 H4Ac

Interestingly, in NTERA-D1 cells, the antibody recognizing acetylated histone 4 at lysines 5,8,12 and 16 (H4Ac) showed enrichments in only three spots. All enriched spots coincide with the start site of *PRKCBP1*. The poor performance of this antibody is intriguing since it worked well in HeLa S3 cells. Also, all antibodies except H4Ac gave higher signals in NTERA-D1 than HeLa S3 cells.

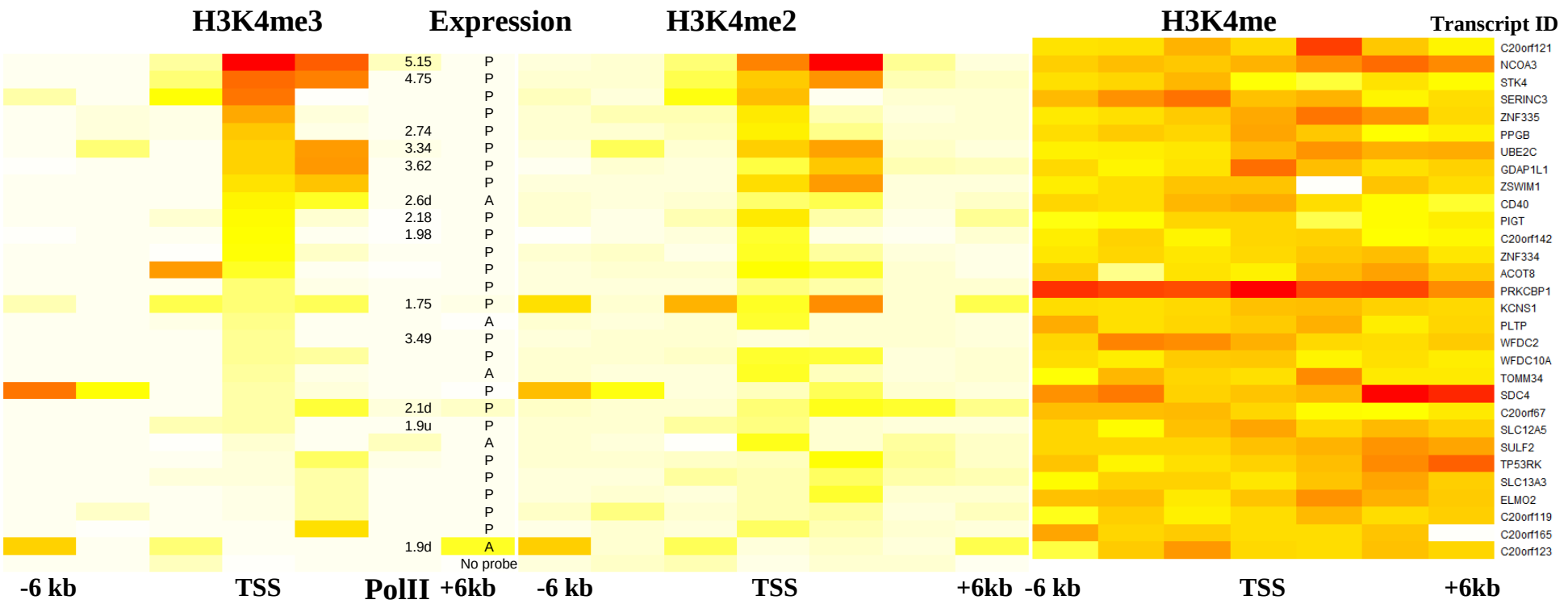


Figure 5.26 H3K4me3, H3K4me2 and H3K4me profiles of 23 genes that showed H3K4me3 enrichments on their start sites in NTERA-D1 cells. Each profile is presented as a heat map which displays the signals obtained within ~6 kb distance of TSSs. The first column on the left lists the polII enrichments and the sec column lists the expression profiles ("P" stands for the gene is expressed and "A" denotes no expression) for the corresponding genes. "d" means that the signal is detected at ~2 kb downstream of TSS.

5.7 Histone Modifications marking possible regulatory elements

In this study, antibodies recognizing seven different modified histones were employed to derive a partial histone map of a 3.5 Mb region at 20q12-13.2. So far, I have described the results obtained with five antibodies namely, mono-, di- and tri-methylated histone H3 at K4 (H3K4me, H3K4me2 and H3K4me3), acetylated histone H3 at lysine 9 and 14 (H3Ac) and acetylated histone H4 at lysines 5,8,12 and 16 (H4Ac), all of which are commonly found in euchromatic regions. In order to look at the number of occurrences of observing each possible combination of modified histones at a given site, all sites carrying different combinations of modified histones were listed (Table 5.6) and then plotted as shown in Figure 5.27. However, since there are only three sites enriched with H4Ac in NTERA-D1 cells, as opposed to 64 sites in HeLa S3 cells, H4Ac combinations were excluded from the plots (Figure 5.27).

In both cells, there are no spots that are enriched only with H3K4me + H3K4me3 or H3K4me + H3K4me3 + H3Ac as expected, since methylation of histone H3 at K4 is performed in a stepwise manner by the same enzyme (MLL) and H3K4me2 is the intermediate molecule between H3K4me and H3K4me3.

In HeLa S3 cells, there are three times the number of sites enriched with only H3K4me than in NTERA-D1 cells. Among the sites that are only enriched with H3K4me, 80% and 85% are within inter- or intra-genic regions in HeLa S3 and NTERA-D1 cells respectively, meaning that they do not possess the potential for further methylation. This observation again supports the opinion that H3K4me is not associated with TSSs and probably exists as a marker for other regulatory regions, most likely in combination with other modified histones.

H3K4me	H3K4me2	H3K4me3	H3Ac	H4Ac	# of occurrences in HeLa S3	# of occurrences in NTERA-D1
H3K4me	H3K4me2	H3K4me3	H3Ac	H4Ac	14	2
H3K4me	H3K4me2	H3K4me3	H3Ac		9	2
H3K4me	H3K4me2	H3K4me3			2	3
H3K4me	H3K4me2		H3Ac	H4Ac	6	
H3K4me	H3K4me2		H3Ac		11	3
H3K4me	H3K4me2			H4Ac	11	
H3K4me	H3K4me2				17	10
H3K4me		H3K4me3	H3Ac	H4Ac		
H3K4me		H3K4me3				
H3K4me			H3Ac	H4Ac	15	
H3K4me			H3Ac			1
H3K4me				H4Ac	9	
H3K4me					71	25
	H3K4me2	H3K4me3	H3Ac	H4Ac	5	
	H3K4me2	H3K4me3	H3Ac		12	35
	H3K4me2	H3K4me3			3	38
	H3K4me2		H3Ac	H4Ac		
	H3K4me2		H3Ac		1	3
	H3K4me2			H4Ac		
	H3K4me2				10	52
		H3K4me3	H3Ac	H4Ac		
		H3K4me3	H3Ac		3	1
		H3K4me3			4	
			H3Ac	H4Ac		
			H3Ac		2	2
				H4Ac	3	1

Table 5.6 Number of occurrences of different histone combinations at one site in HeLa S3 and NTERA-D1 cells.

Next, 10 out of 11 (90%) and 40 out of 54 (73%) sites which were only enriched with H3K4me2, were within inter- or intra-genic regions in HeLa S3 and NTERA-D1 cells respectively. Sites that were enriched only with H3K4me and H3K4me2 were not also close to annotated start sites. However, nearly 90% of the sites that were enriched with H3K4me + H3K4me2 + H3Ac were close to TSSs. This observation can lead to two conclusions; (i) H3 acetylation may indeed have a stimulatory effect for a site to be enriched in H3K4me3 (Milne et al., 2002; Nakamura et al., 2002) and (ii) H3Ac may not be a common epigenetic marker in regulatory elements other than promoters.

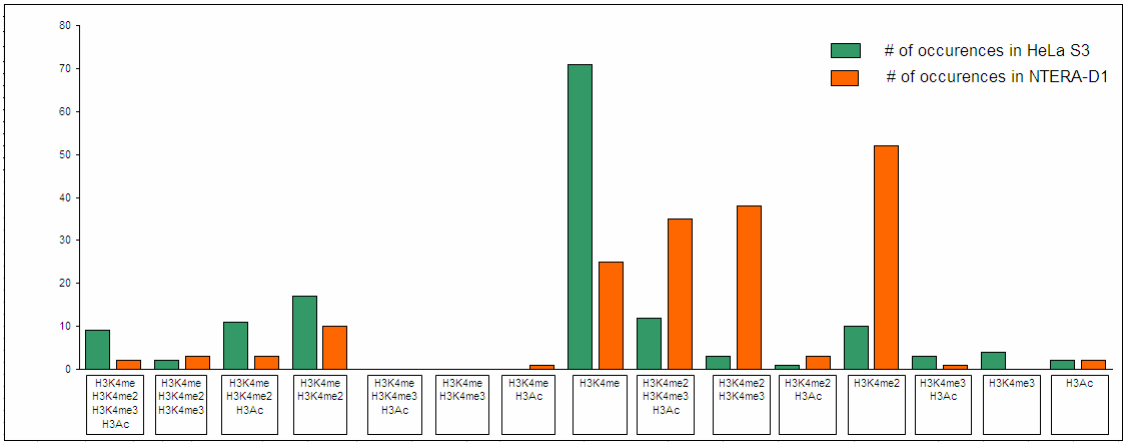


Figure 5.27 Plots of number of occurrences of all possible modified histone combinations in HeLa S3 and NTERA-D1 cells

5.7.1 HeLa S3 cells

It is interesting to look at the spots that gave enrichment with a number of modified histones but are not located in close proximity of any TSS. In HeLa S3 cells, there are 99 sites that showed enrichments with H3K4me2; 45 of which showed enrichment also with H3K4me3 and are associated with TSSs. Out of the remaining 54 H3K4me2 enriched sites, 80% showed enrichment also with H3K4me. The spots that showed low enrichment with only one antibody will not be discussed here due to insufficient experimental evidence for their potential functions. Table 5.7 lists 28 selected spots that gave H3K4me2 enrichment and their short feature descriptions.

Index	Start Coordinates	End Coordinates	Spot Information	polII	H3K4me	H3K4me2	H3Ac	H4Ac	CTCF
SP1	42469627	42471927	within HNF4A gene; contains one highly conserved region	6.71		3.23			
SP2	42670305	42672476	low conservation		1.74	1.84		2.3	6.59
SP3	42672013	42674371	within PKIG gene; no features		2.71	2.15		2.18	7.55
SP4	42704657	42706871	no conservation		2.63	1.77			
SP5	42754161	42755685	contains highly conserved regions	5.98	3.47	2.92		2.18	
SP6	42755951	42756933	low conservation	1.58	9.61	2.21		3.56	
SP7	42756907	42757833	low conservation	1.25	5.91	5.71	2.15	2.84	
SP8	42757760	42758740	no conservation	1.5	5.29	4.76	1.6	2.99	
SP9	42760403	42761222	contains one ultra conserved region		9.73	2.24	2.04	4.98	
SP10	43235675	43237871	contains 5' end of inactive PI3 gene		3.23	2.27			
SP11	43381309	43383563	Contains one highly conserved region		2.25	1.6			2.78
SP12	43839764	43842639	contains 5' end of inactive WFDC3 gene		2.99	1.64			
SP13	43883251	43885299	contains 3' end of TNNC2 gene		2.73	1.52			
SP14	43896779	43899401	contains 5' end of non-coding transcript of SNX21 gene		3.79	1.51			
SP15	43922315	43924699	within ZSWIM3 gene; contains one highly conserved region		2.88	2.25			
SP16	44147270	44149246	within NCOA5 gene; contains one highly conserved region	1.83	2.73	2.91			
SP17	44275725	44277729	within CDH22 gene; contains one highly conserved region		5.49	2.06			
SP18	44380475	44382544	Contains one highly conserved region		2.77	3.83			
SP19	44461256	44463475	within ELMO2 gene; contains highly conserved regions		1.77	2.5			
SP20	44622588	44624763	no conservation			1.7			4.74
SP21	44624669	44627010	contains an exon of SLC13A3 gene			2.85			5.29
SP22	45395404	45396897	within PRKCBP1 gene; contains a weak conserved region		2.52	3.39	1.62		
SP23	45453630	45455876	no conservation	2.65	2.43	1.51		4.31	
SP24	45516032	45518287	contains highly conserved regions		4.95	1.83		2.63	
SP25	45553177	45556238	contains highly conserved regions		1.87	1.9			
SP26	45636634	45639142	within NCOA3 gene; contains highly conserved regions		4.8	5.26			
SP27	45638797	45641114	within NCOA3 gene; contains highly conserved regions		3.91	3.55			
SP28	45648944	45651138	within NCOA3 gene; low conservation		3.05	2.99		1.53	

Table 5.7 The enrichment profiles of 28 H3K4me2 enriched spots. Empty cells means that there was no significant enrichment with that of specific antibody. First two columns list the genomic start and end coordinates (NCBI, version 36) of the enriched spots, while the spot information column gives a short description of the sequence of the spot.

One region encompassing the SP5 to SP9 (Table 5.7) showed a particularly interesting enrichment pattern. This region does not contain any annotated coding features although it contains a novel transcript with no open reading frame (RP11-445H22.4). An Affymetrix expression probe (241759_at) designed mRNAs, AK090842 and CR597563, (13 spliced ESTs), gave an expression signal only in HeLa S3 cells. The annotation of the region is given in Figure 5.28, together with the enrichment levels of the different proteins. The region is overall enriched with H3K4me (~5-10 fold) and H4ac (~2-4 fold). Also, the spot containing the mRNA and spliced ESTs were enriched with H3K4me2 (6 fold) and H3Ac (2.2 fold), the epigenetic markers which are often found on the TSSs of active genes.

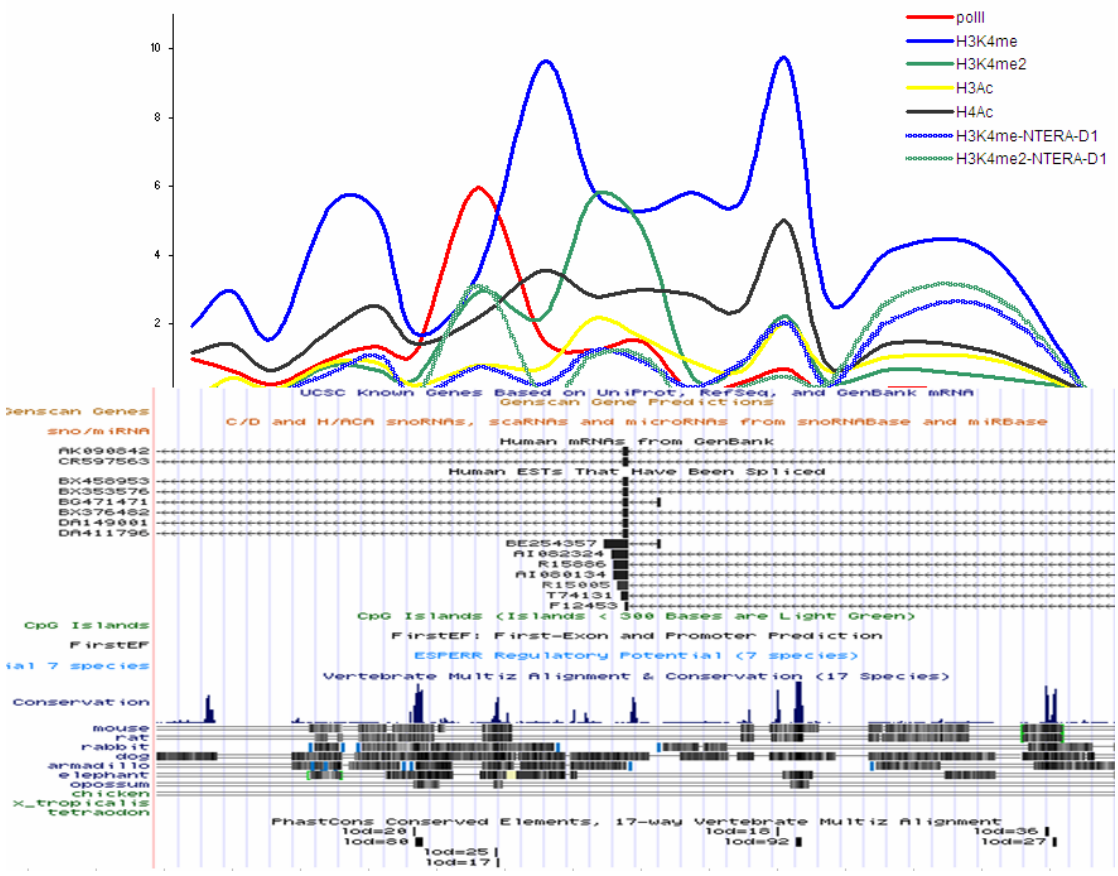
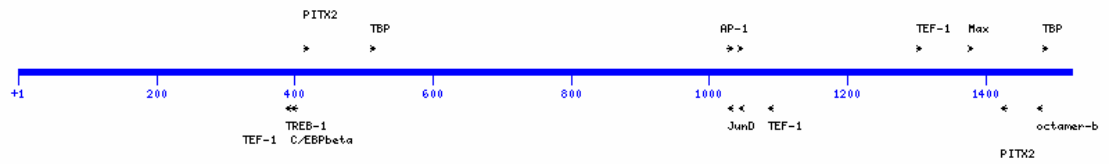


Figure 5.28 Annotation taken from UCSD Genome Browser for the region between 42,749,148 and 42,766,245 bp together with enrichment levels with proteins RNA polymerase II (polII), H3K4me, H3K4me2, H3Ac and H4Ac in HeLa S3. Also the thick blue and green lines displays enrichment levels with H3K4me and H3K4me2 in NTERA-D1 cells.

There is no H3K4me3 enrichment but SP5 shows a 5.9 fold RNA polymerase II peak. Despite the fact that the spot which showed polII enrichment is not close to mRNA or EST evidence in the region (circa 3 kb upstream), still the Affymetrix probe showed an expression for this mRNA in HeLa S3 cells. To explore whether SP5 has any promoter characteristics, I searched the sequence for putative transcription factor binding sites using the program MAPPER (Marinescu et al., 2005).



TEF-1	Transcription Enhancer Factor-1 (interacts with TBP and AP-1)
TREB-1	Tax responsive element binding protein-1 (interacts with TBP)
TBP	TATA-box binding protein
AP-1	Activator protein-1
JunD	Jun D proto oncogene (functional part of AP-1 activation complex)
Max	Myc binding protein X (interacts with TEF-1)
PITX2	paired-like homeodomain transcription factor 2 (expressed only in HeLa S3 cells
C/EBPbeta	CCAAT/enhancer binding protein beta subunit

Figure 5.29 Transcription factor binding profile of the region spanning from 42,754,161 to 42,755,685 bp that gave a 5.5 fold enrichment with RNA polymerase II. This figure is the reproduced from the graphical output of program MAPPER used to search putative binding sites.

Figure 5.29 shows a graphical display of the putative binding sites on the polII enriched region for the human transcription factors. Only factors expressed in HeLa S3 cells were included in this analysis. Interestingly, this region contains two perfect TATA boxes and binding sites for TEF-1 protein which is an activator that binds to enhancers and also interacts with TATA-box binding protein (TBP) (Gruda et al., 1993). This region also has binding sites for common transcription factors such as AP-1, JunD and Max that act as activators together on a number of human promoters. More importantly, it also contains a binding site for enhancer binding factor C/EBPβ, which interacts with RNA polymerase II initiation complex (Mo et al., 2004). However the lack of H3K4me3 and H3Ac enrichment raises questions whether this region has potential to be a real TSS. In addition, no common promoter elements such as GC-boxes (Sp1 binding sites), initiator element or CAAT-box were found.

Within SP9 (Table 5.7), there is a 201 bp long region which is highly conserved across five species. This spot showed H3K4me (~10 fold), H3K4me2 (~2.2 fold), H3Ac (~2 fold) and H4Ac (~5 fold) enrichments. The alignment of this conserved region across five species is shown in Figure 5.30.

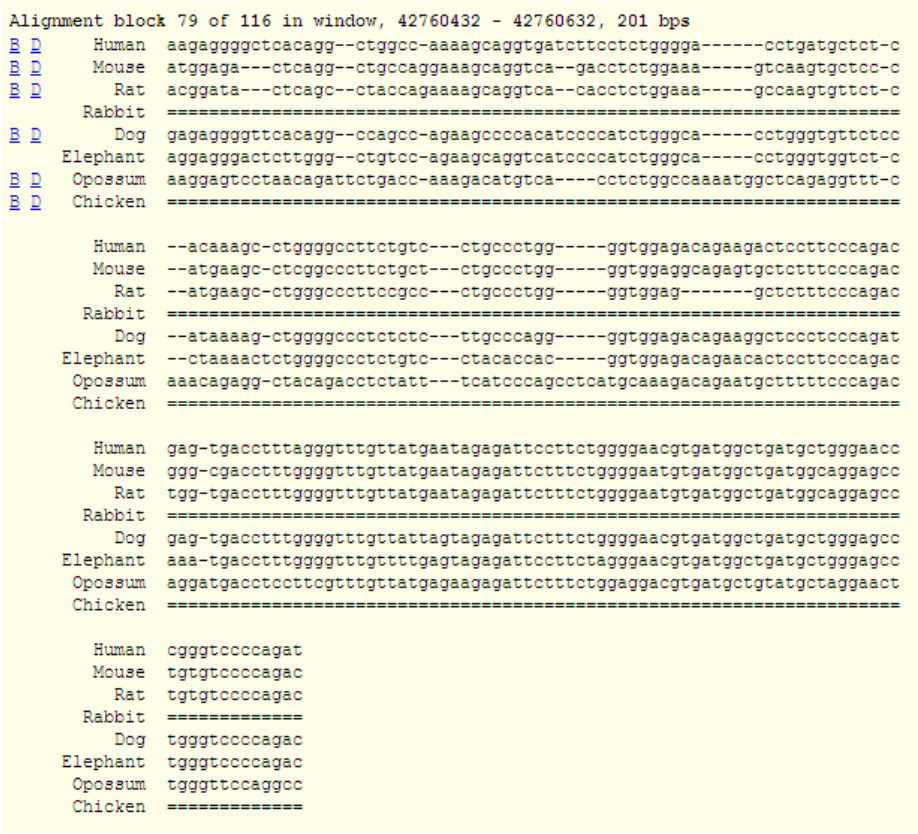


Figure 5.30 The alignment of the human DNA sequence spanning between 42,760,432 and 42,760,632 bp coordinates to mouse, rat, dog, elephant and opossum sequences. This alignment is reproduced from UCSD Genome Browser. This conserved region is within the spot that gave enrichments with H3K4me, H3K4me2, H3Ac and H4Ac modified histones.

This highly conserved region was searched for putative binding sites of transcription factors by the program MAPPER and it was found to contain binding sites for AP-2, Max-1, SPI-B, CD28RC and PU-1 transcription factors.

In light of these findings, this region (SP5 to SP9) may well be a tissue specific distal regulatory element whose function seems to be regulated by different histone modifications especially by mono-methylated histone H3 at K4 and acetylated histone H4. This region may also have strong interactions with the initiation complex on the

promoter of its target gene located elsewhere on the genome and as a result, it gets crosslinked and co-immunoprecipitated as if polII bridged these two sequences. This hypothesis can be tested by assessing the enhancer activity of this region on a promoter (or a number of promoters) via gene reporter assays. The promoter activity of the region with a polII enrichment can also be tested using gene reporter assays. Note that this region (SP5 to SP9) is enriched with H3K4me and H3K4me2 also in NTERA-D1 cells (thick blue and green curves in Figure 5.28). Interestingly, this may indicate that the above combination of histone modifications can mark distal elements irrespective of their activity status in different cell types. Additionally, SP5, which shows a polII enrichment in HeLa S3 cells, did not show any polII but CTCF (3.2 fold) enrichment in NTERA-D1 cells. This region might be repressed in NTERA-D1 cells due to CTCF binding which can function as a repressor (see section 1.3.2.1).

SP4 is located within the first intron of the *ADA* (42,704,657 to 42,706,871 bp) and showed 2.6 and 1.8 fold enrichment with H3K4me and H3K4me2. In NTERA-D1 cells, SP4 including its left and right spots (42,702,249 to 42,708,995) also showed H3K4me and H3Kme2 enrichments (see Table 5.8). An intronic enhancer of the ADA promoter has already been located between 42,706,026 and 42,709,030 region that contains three DNase hypersensitive sites (Aronow et al., 1989). This intronic enhancer is contained within the spots that showed enrichment with H3K4me and H3K4me2 in both cell lines. This observation supports the hypothesis that mono-methylated histone H3 can indeed be a marker for distal enhancer elements (Ren, B., unpublished data).

SP26 is enriched with H3K4me (3.9 fold) and H3K4me2 (3.6 fold). This region lies within the intron of NCOA3 gene and it contains two ultra conserved regions of 176 and 179 bp long; the genomic alignments of these regions are given in Figure 5.31. Analysis of the first region with the program MAPPER found a perfect-match binding

site for the six members of ETS family transcription factors. These proteins are transcriptional activators; mainly signalling pathways such as MAP kinases or Ca^{+2} specific signals activated by growth factors or cellular responses converge on the ETS family proteins, controlling their activity, interactions and specification of their downstream targets (Yordy and Muise-Helmericks, 2000). In addition, there is a binding site for a nuclear receptor factor (NR5A2), an enhancer binding factor (Li et al., 1998), which interestingly is known to interact with NCOA3 protein (Ortlund et al., 2005). To estimate the probability of finding a putative binding site for NR5A2 by chance, a 30 kb intergenic sequence was searched for its binding site, one NR5A2 binding site was found per 2.63 kb. I also searched a 600 kb sequence containing several genes, and the program found one NR5A2 binding site approximately per 2.94 kb. Thus, there is a high probability that the putative NR5A2 binding site found within this 176 bp highly conserved site is real.

Figure 5.31 The alignments across multi-species of the regions spanning from 45,637,267 to 45,637,440 bp on the left and 45,638,816 to 45,638,995 bp on the right. The alignments were taken from UCSC Genome Browser.

NCOA3 directly controls the expression of genes important for initiation of DNA replication and it has been linked to multiple types of human cancer due to its frequent

over-expression (Louie et al., 2006). Moreover, NCOA3 has been shown to bind to its own promoter to direct its auto-regulation in a positive manner (Louie et al., 2006). Therefore, SP26 may actually play a role in auto-regulation of the *NCOA3*. This possibility surely requires further experimental verification, such as assessing the activity of SP26 on the *NCOA3* promoter by using for example, gene reporter assays. A ChIP experiment using an antibody recognizing the NCOA3 protein could also offer important insights on the regulatory elements of this gene. The second conserved site had putative binding sites for interferon regulatory factors, but no other relevant binding site could be found.

SP28 is also located in an intron of the *NCOA3*. It was enriched in H3K4me (3.1 fold), H3K4me2 (3 fold) and H4Ac (1.5 fold). This region contains a single short highly conserved site and it has moderate conservation with dog and armadillo (UCSC Genome Browser). Again, SP28 was searched by MAPPER and several putative binding sites for three relevant factors were found: five binding for the enhancer binding factor TEF-1, POU2F1 (Oct-1) and NR5A2 protein. The graphical display of the binding sites is shown in Figure 5.32. All these factors are known enhancer binding factors, and the arrangement of binding sites of TEF-1 and POU2F1 is quite interesting due to its similarity to that in the SV40 enhancer, on which they bind in symmetry to exert their activation functions (see Figure 1.4). For these reasons, SP28 is also a good candidate to be tested for enhancer activity.

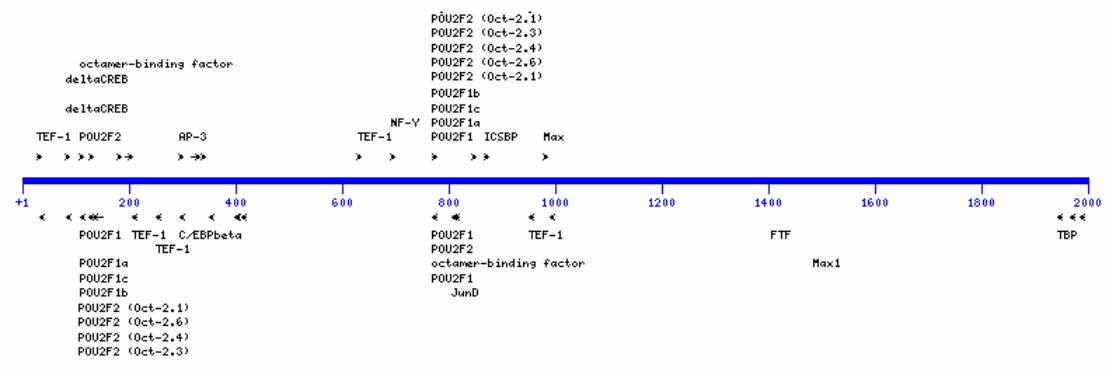


Figure 5.32 The putative binding sites on the sequence spanning from 45,648,944 to 46,651,138 bp. SP28 was enriched with H3K4me, H3K4me2 and H3Ac proteins.

SP24 showed enrichment with H3K4me (5 fold), H3K4me2 (1.8 fold) and H4Ac (2.6 fold) and contains multi-species conserved regions. No relevant binding sites were found by MAPPER on the conserved sites. SP24 contains six binding sites for Topors (topoisomerase I binding, arginine/serine rich) protein. It is unusual to find six Topors binding sites within 2255 bp since its consensus binding site sequence is 22 bp long. Topors have been shown to regulate the activity of p53 by ubiquitination and sumoylation, acting as tumour suppressors (Rajendra et al., 2004) (Weger et al., 2005). It also has trans-activating activities by binding to promoters and other regulatory elements of its target genes (Chu et al., 2001). Taking these observations together, this region seems to have a potential to be a distal regulatory element that requires further testing.

Within the *NCOA5*, SP16 is enriched in polII (1.8 fold), H3K4me (2.7 fold) and H3K4me2 (2.9 fold) and it has one highly conserved block of 226 bp. This region contains multiple putative binding sites for TEF-1, AP-1, RUNX1 (an enhancer binding protein) and SOX9 (enhancer binding protein). It is known that TEF-1 and AP-1 act in synergy on SV40 enhancer and exert their activation functions. In gene reporter assays, *NCOA5* showed one of the highest responses to SV40 enhancer which also carries binding sites for AP-1 and TEF-1 factors. Therefore this possible

regulatory element can also have some activatory potential on the promoter of *NCOA5* in which this element resides.

It is important to note that for *NCOA5*, the TSS was not enriched in any protein studied, although *NCOA5* was highly expressed in both cell lines. The core promoter of this gene also showed very high activity (~900 fold increase compared to background constructs in HeLa cells) in reporter assays in both cell lines. No apparent problem was detected with the spot carrying the TSS. If the lack of any sort of enrichment on the TSS is not an experimental error, this gene may have a unique regulatory pattern where its promoter is not marked with common markers as H3K4me3 and H3Ac. Just to add here that there are 10 genes in HeLa S3 and 7 genes in NTERA-D1 cells that are expressed but their start sites are not enriched with polII or H3K4me3. Also, in a genome-wide promoter study by Ren et al, it was found that the start sites of 15% of expressed genes were not enriched with either polII, H3K4me3 or H3Ac (Kim et al., 2005).

SP1 lies within the *HNF4A* and showed polII (6.7 fold) and H3K4me2 (3.2 fold) enrichments. It contains a highly conserved region of 291 bp in which MAPPER found putative binding sites for interacting factors SMACA3 and Sp3. SMACA3 is a member of SWI/SNF chromatin modelling factors and Sp3 is a transcription factor acting as an activator or a repressor on several human promoters. No other binding sites for relevant proteins were found in SP1. The enrichment profile of SP1 its flanking region will be discussed in detail in section 5.10.

5.7.2 NTERA-D1 cells

This cell line exhibited a different histone enrichment profile than HeLa S3. In NTERA-D1, only 25 spots were enriched with H3K4me whereas this number was 71 in HeLa S3 cells. Also, in NTERA-D1 cells, 70% of the H3K4me enriched spots were

in close proximity of annotated start sites while only 45% of those spots were in close proximity of annotated start sites in HeLa S3 cells. There were 52 spots enriched only with H3K4me2 in NTERA-D1 but only 10 spots in HeLa S3 cells. In NTERA-D1, 35% of these H3K4me2 enriched spots were in close proximity of annotated start sites.

Also, the H3Ac enrichment profiles were different between these two cell lines; the percentage occurrences of H3Ac enriched spots in combination with other modified histones is shown in Figure 5.33. In NTERA-D1, H3Ac enrichment was closely associated with H3K4me3 enrichment, whereas in HeLa S3, H3Ac was equally present together with H3K4me and H3K4me3 enrichments.

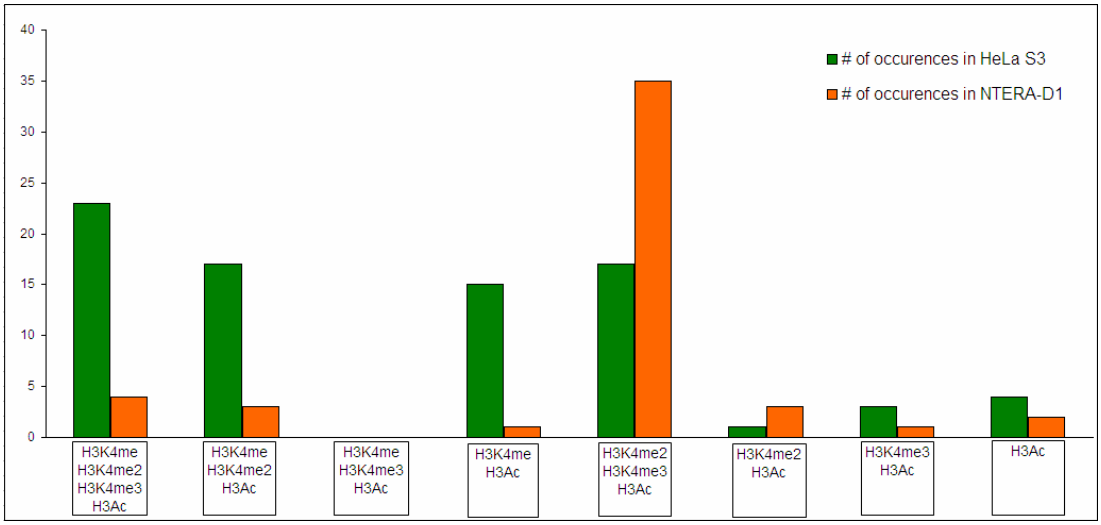


Figure 5.33 Percentages of the spots enriched with possible combinations of modified histones that includes H3Ac.

The enrichment profile of NTERA-D1 with H4Ac was very poor, only three spots showed enrichment higher than the selected threshold and all of them were in close proximity of the TSS of *PRKCBP1*. There were several enriched spots with H4Ac but they did not exceed the threshold level. Maybe this antibody did not work optimally in this cell type, although this is highly improbable since it worked very well in HeLa S3 and NTERA-D1 cells showed overall higher enrichment levels with most antibodies (except H4Ac and H3K9me2) compared to HeLa S3 cells.

It cannot be excluded that the observed differences are due to the use of tissue-specific combinations of modified histones in order to recruit different sets of activators on distal regulatory sites. A set of ChIP experiments using antibodies recognizing different sets of modified histones can be performed in NTERA-D1 cells to see their potential for marking distal regulatory elements.

These differences can also be due to harvesting the two cell lines at different time points in the cell cycle. This possibility can be tested by performing ChIP experiments with synchronized cell lines.

Index	Start Coord.	End Coord.	Spot Information	H3K4me	H3K4me2	H3Ac	H3K27me3	H3K9me2	CTCF
SP1	42385488	42387258	contains moderately conserved regions		2.038				
SP2	42586728	42588766	contains two short highly conserved region		1.718				
SP3	42670305	42672476	within PKIG gene; low conservation		1.583				4.1
SP4	42702249	42704639	within ADA gene; no conservation		3.845				
SP5	42704657	42706871	within ADA gene; no conservation	2.91	5.957	1.84			
SP6	42706827	42708995	within ADA gene; no conservation	2.289	2.717				
SP7	42754161	42755685	no conservation		3.17				
SP8	42754237	42755121	no conservation		2.933				
SP9	42755040	42757910	Low conservation		1.692				
SP10	42760403	42761222	contains one highly conserved region	2.046					
SP11	42761596	42763328	no conservation	2.167	2.758				
SP12	42762937	42764899	no conservation	2.541	2.917				
SP13	42789250	42791223	contains 3' end of WISP2 gene	1.952	1.78				
SP14	42790775	42793081	no conservation		2.987				
SP15	42825732	42827861	no conservation	1.69					
SP16	42845959	42847575	within RIMS4 gene; no conservation		1.73				
SP17	42866586	42869054	within RIMS4 gene; contains two highly conserved region	1.67					
SP18	43146874	43149361	no conservation		2.402				
SP19	43270902	43273021	contains 3' end of SEMG1 gene		2.948				
SP20	43319244	43320783	no conservation		2.095				
SP21	43398547	43400836	within SDC4 gene; contains one highly conserved region	2.346	2.43				
SP22	43400608	43402868	within SDC4 gene; contains two highly conserved regions	2.079	2.028				
SP23	43827441	43829560	no conservation		4.257				
SP24	43884167	43885901	contains 3' end of TNNC2 gene	2.03					
SP25	44063015	44065393	contains moderately conserved regions with mouse and rat	1.739					
SP26	44083316	44085609	contains 5' end of RP11-465L10.10 processed transcript; contains one promoter prediction		2.132				
SP27	44117171	44119404	within SLC12A5 gene; contains highly conserved regions		1.897		7.787	1.267	
SP28	44198835	44200465	no conservation		5.552				
SP29	44256074	44258486	within CDH22 gene; contains one ultra conserved region and a spliced EST	1.681					
SP30	44263883	44266386	no conservation	2.154					
SP31	44265868	44267498	no conservation	1.928	1.967				
SP32	44277449	44279433	within CDH22 gene; contains a highly conserved region		1.57				

SP33	44298319	44300488	within CDH22 gene; contains a highly conserved region		2.667				
SP34	44388316	44389991	contains one ultra conserved region	2.136					
SP35	44404114	44405724	no significant conservation			1.506			
SP36	44410407	44412877	contains the 3' end of SLC35C2 gene	2.47	3.077	1.546			3.9
SP37	44682236	44684551	within SLC123A3 gene; no conservation		1.807				
SP38	44736072	44738637	contains moderately conserved regions		2.018				
SP39	44746060	44748961	contains 3' end of TP53RK gene		1.937				
SP40	44764688	44766247	contains moderately conserved regions		1.88				
SP41	44775942	44778380	no conservation		1.947				
SP42	44939262	44941435	no conservation		5.843				
SP43	45007120	45009126	within EYA-2 gene; contains one ultra conserved region		3.875				
SP44	45053697	45056047	within EYA2 gene; no conservation		1.785				
SP45	45085601	45088163	within EYA2 gene; no conservation		1.715				
SP46	45179954	45182499	within EYA2 gene; contains two highly conserved region		2.675				
SP47	45433592	45436002	contains one moderately conserved region		2.185				
SP48	45581712	45582950	within NCOA3 gene; no conservation		4.692				
SP49	45617423	45620210	within NCOA3 gene; no conservation		4.438				
SP50	45620195	45622416	no conservation		2.775				
SP51	45636634	45639142	within NCOA3 gene; contains ultra conserved regions		1.828				
SP52	45648944	45651138	within NCOA3 gene; moderately conserved		2.557				
SP53	45748303	45750946	within SULF2 gene; low conservation		2.962				
SP54	45819538	45820937	within SULF2 gene;contains one highly conserved region		2.137				
SP55	45824411	45826796	one short highly conserved stretch			1.708			
SP56	45828163	45830521	within SULF2 gene;no conservation		4.222				

Table 5.8 The enrichment profiles of 56 enriched spots not in close proximity of any annotated start sites in NTERA-D1 cells. Empty cells means that there was no significant enrichment with that specific antibody. The first two columns list the genomic start and end coordinates (NCBI, version 36) of the enriched spots, while the spot information column gives a short description of the sequence of the spot.

Table 5.8 lists all enriched spots and a short sequence feature description in NTERA-D1 that are not in close proximity (more than 2 kb apart from) with a TSS. Several of these sites contain highly conserved sequences, while some are more than 95% conserved over 100 bp long stretches. The intronic sites were searched for putative binding sites for enhancer-binding proteins such as SMAD-3 and SMAD-4, Octamer binding proteins, TEF-1, AP-1 and AP-3. Since most of these spots were only enriched with one histone modifications, more evidence is needed for them to be treated as potential distal regulatory sites.

5.8 Histone Modification involved in transcriptionally inactive regions

ChIP experiments were performed using antibodies recognizing tri-methylated histone H3 at K27 (H3K27me3) and di-methylated histone H3 at K9 (H3K9me2) in both HeLa S3 and NTERA-D1 cells. These modified histones are often associated with X-chromosome inactivation (Plath et al., 2003), gene silencing (Plath et al., 2003) and heterochromatin formation and maintenance (Richards and Elgin, 2002) (Grewal and Moazed, 2003).

5.8.1 H3K27me3

5.8.1.1 HeLa S3

None of the spots containing TSS were found to be enriched with H3K427 higher than the threshold. A heat map was generated that displays the H327me3 signals within 6 kb distance of all the genes in comparison with H3K4me3 signals (Figure 5.34).

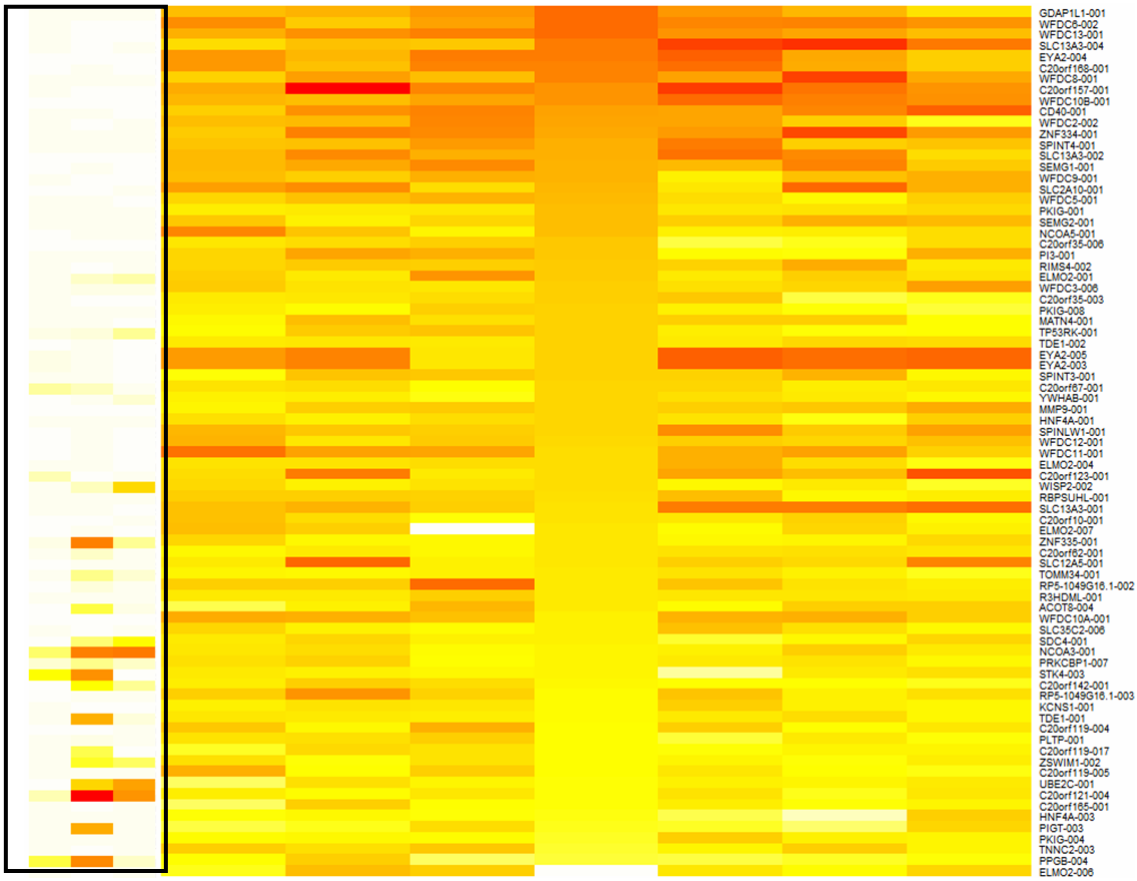


Figure 5.34 Heat map displaying H3K27me3 signals of 6 kb distance of 79 TSSs in HeLa S3 cells. The small heat map on the left displays the H3K4me3 signals of 2 kb distances of the corresponding start sites.

Figure 5.34 shows that H3K27me3 signals are negatively correlated with H3K4me3 signals at the start sites of genes in the region. This confirms that genes indeed carry different histone codes depending on their transcriptional activity status. Table 5.9 lists the 15 spots that gave signals above the selected threshold (1.50) with H3K27me3 antibody together with a short feature description of their sequences. One interesting point is that all the spots that are H3K27me3 enriched are located after a gene rich segment of the region. This may be a sign of a start of a different chromatin domain within the region.

Index	Start Coord.	End Coord.	Spot Information	H3K27me3	CTCF
SP1	44163647	44166500	no conservation	1.547	
SP2	44190355	44192401	contains 3' end of CD40 gene	2.502	4.844
SP3	44198835	44200465	no conservation	1.537	
SP4	44202007	44204336	contains short conserved regions	2.188	
SP5	44359438	44361613	5' end of a GENSCAN gene prediction; no conservation	1.62	
SP6	44360913	44362559	5' end of a GENSCAN gene prediction; contains one short highly conserved region	1.567	
SP7	44382683	44384608	no conservation	1.885	
SP8	44399167	44401245	contains one highly conserved region	1.765	
SP9	44401188	44402830	contains one short highly conserved region	1.505	
SP10	44540333	44542416	no conservation	1.771	
SP11	44637193	44639415	within SLC13A3 gene; contains one short moderately conserved region	3.175	
SP12	44705037	44707354	within SLC13A3 gene; no conservation	1.823	
SP13	45179954	45182499	within EYA2 gene; contains two highly conserved regions	1.663	
SP14	45807042	45808720	within SULF2 gene; no conservation	2.093	
SP15	45808066	45810484	within SULF2 gene; no conservation	2.311	

Table 5.9 The enrichment profiles of 15 H3K27me3 enriched spots in HeLa S3. Empty cells means that there was no significant enrichment with that of specific antibody. First two columns list the genomic start and end coordinates (NCBI, version 36) of the enriched spots, while the spot information column gives a short description of the sequence of the spot.

Since there are several sites that are H3K27me3 enriched, this modified histone can be a functional marker of distal regulatory elements as well. It is known that histone H3 K27 tri-methylation facilitates binding of polycomb protein (a member of poycomb complex) to histone H3 and regulates the silencing of polycomb group genes (Cao et al., 2002). This constitutes an example of how different histone codes on a genomic site direct the recruitment of different sets of factors on a site. Therefore, I postulate that such regions have the potential to be functional regulatory elements and needs to be pursued further by ChIP experiments performed with antibodies recognizing a variety of transcription factors.

5.8.1.2 NTERA-D1

In NTERA-D1 cells, 49 spots were enriched with H3K27me3 and the histone enrichment profile is different than HeLa S3 cells. A heat map was generated that shows the enrichment levels within 6 kb of the TSSs in the region (Figure 5.35).

Nine genes showed H3K27me3 enrichment within 2 kb distance of their TSSs, and four of them also showed enrichment with H3K4me3. This is quite different from HeLa S3 cells in which no H3K4me3 enriched genes showed enrichment around the start site. On the other hand, none of the H3K27me3 enriched spots showed enrichment with polII or acetylated histone H3 or H4 in both cell lines. Table 5.10 lists the spots that are enriched with H3K27me3 along with the enrichment levels of the spots with other proteins used in this study.

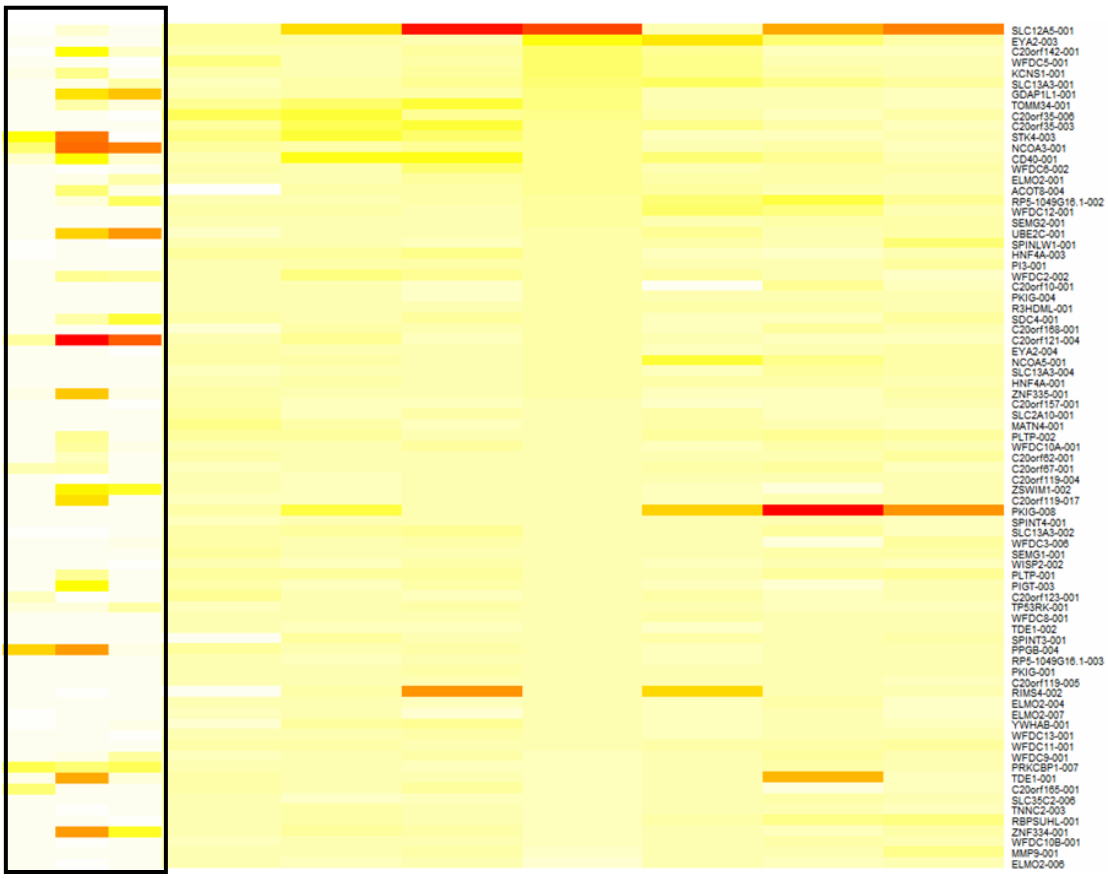


Figure 5.35 Heat map displaying H3K27me3 signals of 6 kb distance of 79 TSSs in NTERA-D1 cells. The small heat map on the left displays the H3K4me3 signals of 2 kb distances of the corresponding start sites.

Index	Start Coordinates	End Coordinates	Spot Information	H3K4me	H3K4me2	H3K4me3	H3K27me3	H3K9me2	CTCF
SP1	42387722	42389987	contains a promoter prediction				3.26		
SP2	42451724	42453980	within HNF4A gene; low conservation				1.58		
SP3	42580025	42581664	within SERINC3 gene; no conservation				5.1		
SP4	42677109	42679491	close to 5' end of PKIG gene	1.984			4.07		
SP5	42681139	42682610	contains 3' end of ADA gene				12.23		
SP6	42683191	42685461	within ADA				6.48		
SP7	42685423	42687531	contains 3' end of PKIG gene				1.69		
SP8	42689279	42691036	close to 5' end of non-coding transcript of ADA gene	1.657	10.82	3.273	21		
SP9	42868663	42871244	close to 5' end of RIMS4 gene				3.86		
SP10	42873659	42875941	close to 5' end of RIMS4 gene				6.44		
SP11	43024757	43026900	no conservation				1.61		
SP12	43465098	43467459	close to 5' end of DBNDD2 gene				1.6		
SP13	44086840	44088464	contains two highly conserved regions				3.72		
SP14	44088445	44090983	close to 5' end of SLC12A5 gene				11.46		
SP15	44090749	44092135	Contains 5' end of SLC12A5 gene		8.547	2.54	9.52		
SP16	44092559	44094553	close to 5' end of SLC12A5 gene		2.453	3.863	5.68		
SP17	44094336	44095961	within SLC12A5 gene				7.2		
SP18	44095668	44097726	within SLC12A5 gene				6.13		
SP19	44097020	44099292	within SLC12A5 gene				9.07		
SP20	44099144	44100963	within SLC12A5 gene				2.18		
SP21	44101265	44103436	contains 3' end of SLC12A5				1.72		
SP22	44103491	44106129	within SLC12A5 gene				3.87	1.082	
SP23	44104967	44107170	within SLC12A5 gene				2.16		
SP24	44106646	44108979	within SLC12A5 gene				1.51		
SP25	44113102	44115230	within SLC12A5 gene				1.59		
SP26	44114762	44117270	within SLC12A5 gene				8.4		
SP27	44117171	44119404	within SLC12A5 gene		1.897		7.79	1.267	
SP28	44120237	44121938	close to end of SLC12A5 gene (see section 5.5.2.1)		13.56	4.843	1.78		
SP29	44149238	44150590	close to 5' end of NCOA5 gene				1.55		
SP30	44176229	44177856	contains one ultra conserved region				1.93		
SP31	44177841	44179824	close to 5' end of CD40 gene		1.632	2.821	1.93		
SP32	44231425	44234066	contains short moderately conserved regions				2.83		
SP33	44233981	44235883	contains short moderately conserved regions				3.05		
SP34	44307554	44309441	within CDH22 gene; contains 3 ultra conserved regions				2.66		

SP35	44310685	44312911	close to 5' end of CDH22 gene		3.587	2.568	3.69		
SP36	44311936	44314006	contains 5' end of CD22 gene				1.69		
SP37	44313930	44316183	close to 5' end of CDH22 gene				2.54		
SP38	44364653	44367357	contains one moderately conserved region				2.08		
SP39	44366974	44369194	contains moderately conserved regions		2.217	1.873	2.69		2.884
SP40	44371086	44372681	contains 5' end of a gene prediction				3.45		
SP41	44375439	44377976	contains one highly conserved region				1.93		
SP42	44956175	44958701	contains 5' end of EYA2 gene				2.04		
SP43	44958627	44959383	close to 5' end of EYA2 gene				3.27		
SP44	45093985	45096221	within EYA2 gene; no conservation				1.51		
SP45	45096065	45098568	within EYA2 gene; contains short highly conserved regions				2.22		2.803
SP46	45533578	45536011	no conservation				14.04		
SP47	45535724	45537859	no conservation				22.34		
SP48	45588657	45591804	within NCOA3 gene; low conservation				2.02		
SP49	45591796	45593160	within NCOA3 gene; contains one highly conserved region				2.45		

Table 5.10 The enrichment profiles of 49 H3K27me3 enriched spots in NTERA-D1 cells. Empty cells means that there was no significant enrichment with that specific antibody. The first two columns list the genomic start and end coordinates (NCBI, version 36) of the enriched spots, while the spot information column gives a short description of the sequence of the spot.

There is one particularly interesting region located between 45,533,578 to 45,537,859 bp which showed the highest enrichment (~20 fold) with H3K27me3 in NTERA-D1 cells. This spot is not within or close to any gene and it has no significant sequence conservation either. A TF binding site search of the region with MAPPER gave 13 different binding sites for PITX2 protein, a homeodomain containing transcription factor, and five sites for AREB6, another homeodomain zinc-finger protein. As mentioned, tri-methylation of histone H3 at K27 serves as a signal for recruiting polycomb gene silencing complex (the proteins necessary to maintain the silence state of *HOX* cluster) (Cao et al., 2002) on promoter regions of target genes for silencing. Since this region does not lie in close proximity of any promoter region, we cannot speculate a similar mechanism. However, it is known that polycomb proteins can exert their silencing effects via long range interactions where their distal DNA binding sites is looped onto the target promoters (Min et al., 2003) (Paro and Hogness, 1991) (Simon and Tamkun, 2002). Based on these putative binding sites for PITX2 and high levels of H3K27me3, these regions may harbour a cis-acting regulatory element of an as yet unknown target promoter. Sequences of some other H3K27me3 enriched spots were searched for putative binding regions, where no such significant sites can be found.

5.8.2 H3K9me2

ChIP experiments performed with this antibody did not show enrichment above the selected threshold of 1.5 in both cell lines. However, in HeLa S3 cells, the TSSs of three genes, *GDAP1L1*, *RIMS4* and *SLC13A3* showed ~1.5 fold enrichment. None of these genes showed any enrichment with polII or other modified histones and they were not expressed in HeLa S3 cells. Although the enrichment levels are relatively low, these results are in agreement with the findings that histone H3 methylation at K9 is associated with inactive chromatin regions (Shilatifard, 2006).

5.9 CTCF

An antibody recognizing the CCCTC-binding protein, CTCF, was employed in ChIP experiments in both HeLa S3 and NTERA-D1 cells in order to locate binding sites within the region of interest. CTCF is a versatile protein functioning as an activator or repressor on promoters or silencer sequences, or a chromatin insulator protein (Ohlsson et al., 2001). It is also located as part of multi-protein complexes regulating histone acetylation and deacetylation (Ohlsson et al., 2001). Here, I am reporting several potential CTCF binding sites and elaborate on their functional roles.

In HeLa S3 cells, there were 55 spots enriched with CTCF above the selected threshold level (1.75). Of these, 43 were also enriched with CTCF in NTERA-D1 cells. Similarly, out of 48 spots that showed enrichment with CTCF in NTERA-D1 cells, 45 were also enriched in HeLa S3 cells. Among CTCF enriched spots, 11 were within 2 kb distance of an annotated TSS, while 21 of them resided within the introns of annotated genes. Also, 12 intergenic regions were enriched with CTCF. A study was performed to locate CTCF binding sites on human chromosome 22 using ChIP assays, where they found around 200 binding sites within genes, promoters or intergenic regions (Mukhopadhyay et al., 2004). This is in agreement with the distribution of CTCF binding sites observed in this study. The summary of CTCF enriched spots that are close to or contain an annotated start site is given in Table 5.11.

Index	Start Coordinates	End Coordinates	Spot Information	CTCF peak HeLa S3 cells	Expression in HeLa S3 cells	CTCF peak NTERA-D1	Expression in NTERA-D1 cells
SP1	43950364	43952334	~1 kb upstream of C20orf165 gene		A	2.21	A
SP2	42309966	42311607	~1 kb upstream of GDAP1L1 gene	1.92	A	3.3	P
SP3	43173696	43176384	~2 kb downstream of WFDC5 gene; no conservation	2.26	A		A
SP4	45266520	45269030	~2 kb upstream of PKRCBP1 gene		P	1.78	P
SP5	42809326	42811520	2 kb downstream of 5' end of KCNK15 gene	6.62	P	4.61	A
SP6	43366633	43368867	contains 5' end of MATN4 gene	6.49	A	4.11	P
SP7	43368246	43370427	contains 5' end of RBPSUHL gene	8.67	A	5.46	A
SP8	43423952	43426324	contains 5' end of rejected C20orf169 gene	2.3	No probe		No probe
SP9	43691558	43692227	contains 5' end of WFDC10A gene	4.57	A	4.5	A
SP10	43692227	43694607	contains 5' end of WFDC9 gene	2.14	A		A
SP11	44031431	44032950	~ 2 kb upstream of ZNF335 gene	4.70	A	3.86	P
SP12	45724594	45727077	Contains 5’ end of SULF2 gene	3.79	A	5.15	P

Table 5.11 The summary of CTCF enriched spots that are close to or contain an annotated start site in both cell lines. H3K4me3 enrichments of the spots are also given in both cell lines.

Out of eight genes whose start sites were enriched with CTCF, six of them were not expressed in HeLa S3 cells. The spot (SP8 in Table 5.11) containing the TSS of rejected *C20orf169* also showed enrichment with H3K4me3. Since C20orf169 has an ambiguous annotation (see Figure 5.22), the role of CTCF on the putative promoter of this gene is not clear. On the other hand, it is possible that CTCF acts as a repressor on the promoter of genes that are not expressed. It is known that CTCF can repress promoters by assisting histone deacetylase complexes assemble on promoter regions (Lutz et al., 2000). The fact that the putative *C20orf169* promoter was not enriched with acetylated histone H3 may be indicative of such a repression mechanism.

The *ZNF335* showed a very interesting enrichment pattern. It is enriched with CTCF, polII as well as modified histones marking for active genes (H3K4me3 and H3Ac) although no significant mRNA expression was detected by Affymetrix Expression Arrays. This gene encodes for a zinc finger protein which indirectly activates ligand-bound nuclear hormone receptors (Mahajan et al., 2002), and it also has a potential phosphorylation site (Beausoleil et al., 2004). The core promoter region of this gene does not contain any Sp1 binding site but a putative AP1 binding site on 14 bp downstream of its annotated start site, and this core promoter did not show any activity but its activity was restored when it was cloned with SV40 enhancer (see section 4.6.1). Since no mRNA expression can be detected despite the fact that the promoter region seems to be actively transcribed, there should be a mechanism which can either halt the initiation complex or decrease the mRNA stability. It has been shown that CTCT can act as a silencer by binding on promoter sites of the genes to be silenced (Klochkov et al., 2006) (Rakha et al., 2004). In the case of *ZNF335*, CTCF can be acting as a repressor by having a negative effect on the elongation process of the transcription.

In NTERA-D1 cells, there are four non-expressed genes, which were enriched with CTCF around their start sites. CTCF binding may be the cause of no transcription. On the other hand, the promoter region of *GDAP1L1* is enriched with CTCF and this gene is actively transcribed in NTERA-D1 cells. As discussed CTCF is pluripotent and in this instance may be part of an activation complex. Furthermore, a study by Norton et al showed that CTCF interacts with the large subunit of RNA polymerase II *in vitro* (Klenova et al., 2002), which can lend support to this hypothesis.

I attempted to localize CTCF binding sites within the promoter regions listed in Table 5.11 by MAPPER. Unfortunately, the program could not locate CTCF binding sites on most sequences within the threshold set for the search (90% confidence level). This was somewhat expected since CTCF uses its 11 zinc-finger motifs to bind sequences up to 50 bp long (section 1.3.2.1) which leads to many possible binding sequences that cannot be detected using one single consensus sequence.

There are 12 intergenic regions that are enriched with CTCF protein which are given in Table 5.12 and Figure 5.36 displays their genomic positions relative to gene features and other potential regulatory elements.

Since CTCF is mainly found acting on insulator elements, firstly I explored the possibility of these 12 regions being insulators. So many insulator elements within an area of less than 2.1 Mb may seem unrealistic, although it is known that insulator elements are mainly found in regions with high density of coding or regulatory sequences, which is the case for this specific region (Fourel et al., 2004). Also, detecting potential insulators only concentrated within the region of highest gene density supports these findings (see Figure 5.36).

Regions	Start Coordinates	End Coordinates	Spot Information	CTCF peak HeLa S3 cells	CTCF peak NTERA-D1 cells
R1	42744108	42746320	intergenic; no conservation	4	2.78
	42745943	42748103	intergenic; no conservation	4.35	3.29
R2	43199283	43201824	intergenic; no conservation	5.02	6.23
	43201123	43203622	intergenic; no conservation	4.87	4.18
R3	43288822	43291796	intergenic; no conservation	4.22	2.88
	43291402	43293747	intergenic; no conservation	2.28	
R4	43297185	43299177	intergenic; no conservation	3.21	3.65
R5	43352110	43353759	intergenic; no conservation	2.86	2
R6	43381309	43383563	intergenic; no conservation	2.78	2
R7	43507579	43509757	intergenic; no conservation	4.16	
R8	43518653	43521025	intergenic; no conservation		2.46
R9	43832093	43834748	intergenic; low conservation	2.07	
R10	44366974	44369194	intergenic; contains two short highly conserved regions	5.26	2.88
R11	44408689	44411452	intergenic; contains short highly conserved regions	3.11	3.76
R12	44820067	44822119	Intergenic; contains short moderately conserved regions	6.63	5.39

Table 5.12.CTCF enriched spots that lie within intergenic regions. The adjacent CTCF-enriched spots are treated as one region.

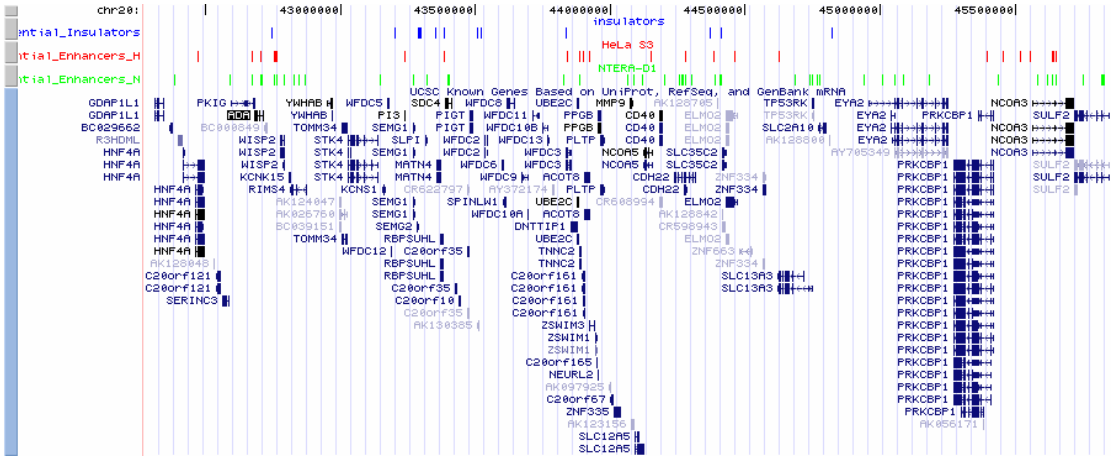


Figure 5.36 The genomic positions of 12 CTCF enriched regions (blue track) listed in Table 5.12 together with the H3K4me and H3K4me2 enriched regions in HeLa S3 (red track) and NTERA-D1 (green track) cells which are seen as potential distal regulatory elements. The region shown here covers the ~3.5 Mb region spanning from 42,274,163 to 45,850,636 bp. The gene annotations are taken from UCSC Genome Browser.

Region 1 (R1;Table 5.12) is enriched with CTCF in both cell lines and it is a potential insulator since it lies between the intronic enhancer of the house-keeping *ADA* (section 5.7.1) and the promoter of the tissue-specific *WISP2*. The schematic representation of the region is shown in Figure 5.37.

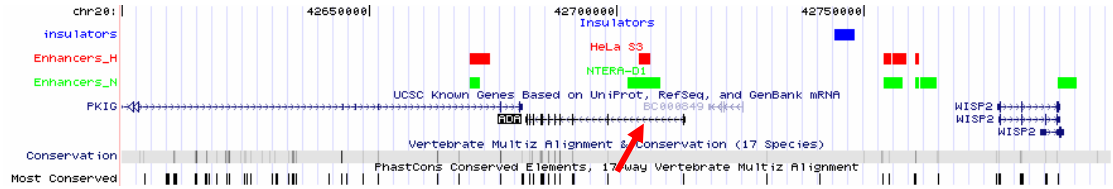


Figure 5.37 A region that contains the intronic enhancer of house-keeping ADA gene (shown with the red arrow) and a possible insulator element (blue box in "Insulators" track) that can block the activity of this intronic enhancer on the tissue-specific promoter of WISP2 gene. "Enhancers_H" and "Enhancers_N" track displays the regions that are enriched with H3K4me and H3K4me2 in HeLa S3 and NTERA-D1 cells as possible enhancer elements.

Figure 5.37 also displays the regions that are enriched with H3K4me and H3K4me2 in HeLa and NTERA-D1 cells. The intronic *ADA* enhancer falls onto one of those enriched regions in both cell lines. This enhancer is a very strong regulatory element that can increase the *ADA* activity in a tissue-independent manner in *in vivo* studies (Aronow et al., 1989). Enhancer elements exert their effects over long distances in an

orientation-independent manner. So this enhancer can be a problem if it was to act on a promoter such as that of the *WISP2* which has tissue specific expression and is probably functioning in bone turnover. An insulator element could solve this problem by blocking the communication between the enhancer and the promoter located on the other side of the insulator. The R1 CTCF-enriched site can be such an element.

There are five candidate insulator elements between 43,100,000 and 43,500,000 bp (Table 5.12). Figure 5.39 displays these putative insulator elements together with the regions that are seen as the possible cis-acting elements in section 5.7.1 and 5.7.2 due to their enrichments with H3K4me and H3K4me2. Each putative insulator lies between two candidate enhancers thereby blocking the communication between these elements to ensure proper regulation of the neighbouring genes. It is important to note that all the genes in this window exhibit a tissue-specific expression pattern.

A novel experimental design was developed to test potential insulators (Mukhopadhyay et al., 2004) and can be used to verify the putative insulator elements described here. The schematic diagram of the reporter construct used in this method is given in Figure 5.29.

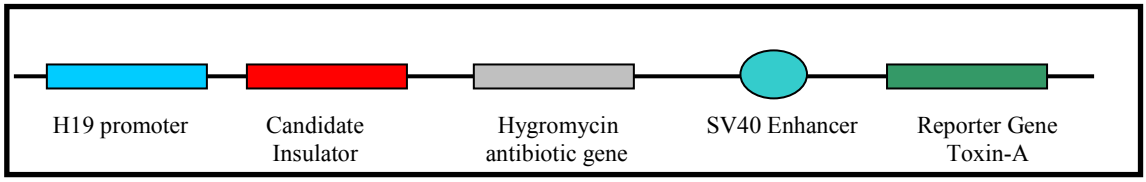


Figure 5.38 An insulator trap reporter vector construct where the candidate insulator is placed between H19 promoter and SV40 enhancer and promoter-enhancer activity is monitored by toxin-A reporter gene. In order to discriminate a possible silencing activity of the candidate fragment, hygromycin gene is placed its downstream and the cells transfected with this construct is screened with hygromycin antibiotic.

In this insulator trapping technique, a potential insulator is placed between a promoter (H19 in this example) and SV40 enhancer and the promoter-enhancer activity is monitored by toxin-A reporter gene. If the fragment is indeed an insulator, it will block the promoter-enhancer communication and toxin-A gene will not be expressed,

hence the cells will survive. The number of cells survived will be determined by a cell-counting assay. An antibiotic hygromycin gene is also placed in downstream of the candidate insulator to discriminate its possible silencing activity since cells were also treated by hygromycin. So, if the fragment is a silencer, then it will silence the hygromycin gene where cells will die.

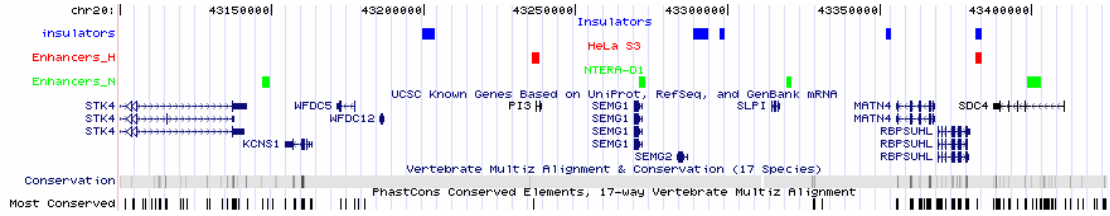


Figure 5.39 The region spanning from 43,100,000 to 43,425,000 bp where there are five candidate insulators shown as blue boxes on the insulator track. There are two more tracks, displaying H3K4me and H3K4me2 enriched regions in HeLa S3 (Enhancers_H track) and NTERA-D1 (Enhancers_N track) as possible cis-acting regulatory elements.

There are 17 more CTCF enriched regions falling within the introns of genes in the regions. These regions are listed in Table 5.13 and schematically represented on the genome in Figure 5.40.

Regions	Start Coordinates	End Coordinates	Spot Information	CTCF peak in HeLa S3	CTCF peak in NTERA-D1 cells
R1	42449867	42451865	within HNF4A gene; one moderately conserved short region	3.90	2.85
R2	42509500	42511603	within HNF4A gene; one highly conserved short region	2.38	2.62
R3	42670305	42672476	within PKIG gene; no conservation	6.59	4.10
	42672013	42674371	within PKIG gene; no conservation	7.55	6.01
R4	42861319	42863581	within RIMS4 gene; contains one highly conserved short region	2.00	2.09
R5	43071862	43073714	within STK4 gene; contains moderately conserved short regions	5.79	3.95
R6	43105913	43106599	within STK4 gene; no conservation	7.31	3.47
R7	43158018	43159290	within KCNS1 gene; no conservation	3.20	3.20
R8	43860206	43861861	within DNTTIP1 gene; low conservation	2.27	
R9	44029429	44031576	within ZNF335 gene; no conservation	2.07	
	44031431	44032950	within ZNF335 gene; no conservation	4.70	3.86
R10	44073900	44075605	within MMP9 gene; low conservation	3.97	2.28
R11	44622588	44624763	within SLC13A3 gene; no conservation	4.74	2.77
	44624669	44627010	within SLC13A3 gene; no conservation	5.29	3.78
R12	44742326	44744698	within SLC13A3 gene; no conservation	2.88	
R13	45034995	45037048	within EYA2 gene; contains short highly conserved regions	3.28	3.50
R14	45096065	45098568	within EYA2 gene; contains short highly conserved regions	4.20	2.80
R15	45212776	45215255	within EYA2 gene; no conservation	3.15	2.68
R16	45724594	45727077	within SULF2 gene; no conservation	5.15	3.79
R17	45738270	45740862	within SULF2 gene; no conservation	8.67	5.69
	45739691	45741084	within SULF2 gene; no conservation	7.49	5.30

Table 5.13 CTCF enriched regions which fall within introns.

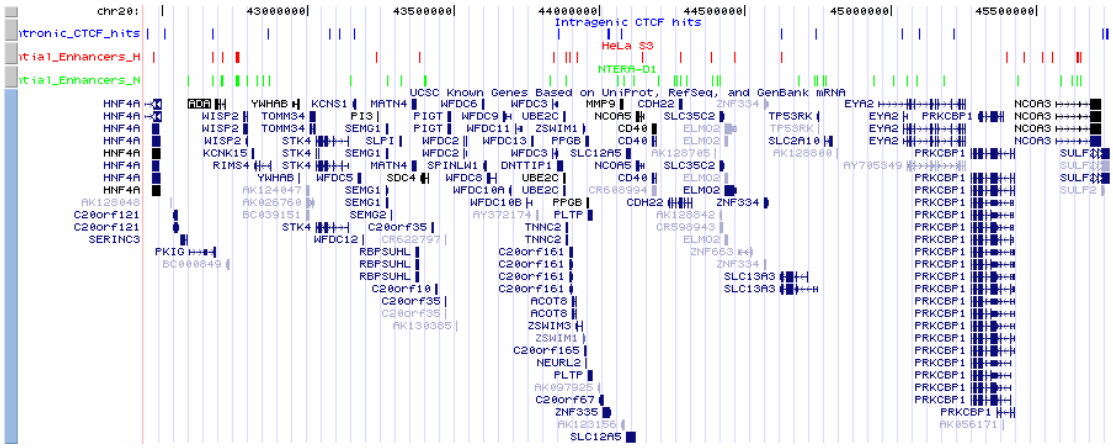


Figure 5.40 CTCF regions that falls within intronic regions shown in the blue track. Red and green tracks display the H3K4me and H3K4me2 enriched regions in HeLa S3 and NTERA-D1 cells respectively.

Interestingly, except *STK*-4, none of the remaining genes containing an intronic CTCF-enriched region showed strong H3K4me3 or polII enrichments on their start sites. While *HNF4A*, *RIMS4*, *KCNS1* and *EYA2* are not expressed in any, *PKIG*, *STK4* and *DNTTIP* genes are expressed in both cell lines according to Affymetrix expression arrays. Surely these regions may also function as insulators, however it would be unfeasible for the cell to use an intra-genic site as a silencer, since insulators need certain dynamic structural requirements such as looping over long distances, and this can be a problem when the gene carrying the insulator needs to be transcribed. On the other hand, some of these regions can be silencers and it is well established that CTCF plays a major role as a repressor in many silencer complexes (Lutz et al., 2000; Ohlsson et al., 2001; Klochkov et al., 2006). Long-range acting silencers are usually placed within introns or 3' ends of their target genes and they possess binding sites for repressor and co-repressor proteins. These regions need to be further investigated for their candidate silencing functions experimentally. This can be tested by replacing these regions together with a constitutively active strong promoter such as CMV or SV40 promoter and investigate their possible silencing effect using gene reporter assays.

5.10 Summary

Table 5.15 and Table 5.16 lists the signals obtained from eight of the antibodies used in this study in HeLa S3 and NTERA-D1 cells respectively. In both cell lines, more than half of the expressed genes (57%) showed binding of either polII, H3K4me3 or H3Ac. The remaining expressed genes did not give an enrichment with any of the above proteins. This may simply reflect suboptimal experimental conditions. However, ChIP experiments performed with H3K4me3 antibody showed very high enrichments especially in NTERA-D1 cells (~70-100 fold). It may also be that the remaining expressed genes are associated with a different set of modified histones, but this possibility is not favoured since in similar studies, most of the expressed genes are associated with H3K4me3 or H3Ac (Kim et al., 2005). Note that, the resolution of the custom-made array (~2 kb) is much lower than arrays used in similar studies (up to ~50 bp). This may also be a reason for not detecting all expressed genes via ChIP on chip. As the DNA size on the spot increases, the signal is diluted. Hence, it will be difficult to detect genes that are not highly enriched with these modifications. This will also be the case for genes that are activated in a transient fashion. Note that no correlation was observed between the expression signal of the gene obtained from Affymetrix Expression Arrays and ChIP enrichment on the start site.

Another issue is the reliability of the gene expression data. All expression arrays use probes which will detect either sense or antisense transcript of genes. Since antisense transcripts do not correspond to an actively expressed gene, a probe detecting such transcripts will lead to a false positive signal. It is not currently possible to determine the fraction of genes that produce an antisense transcript, although a new microarray platform is being developed to detect sense and antisense transcripts of a gene separately (Clevebring et al., 2006). Such a microarray platform will certainly improve the correlation between gene expression platforms and activity information

obtained by ChIP experiments. They can also assist to resolve ChIP enrichment like those obtained from genes such as *MATN4* and *SLC12A5*, in this study where their 3'end was also enriched with H3K4me3 (see sections 5.4.1 and 5.5.2).

There are certain cases where annotated start sites produced lower enrichments than the flanking regions: for example, *PRKCBP1* and *SDC4* (see sections 5.4.1 and 5.4.2). The TSSs of these genes need further experimental examination although these variations may simply be due to a mis-assembled tilepath or a mixed well during PCR to construct the array.

A small subset of genes that are not expressed also showed enrichments with either polII, H3K4me3 or H3Ac (see Table 5.15 and Table 5.16). In HeLa S3, there are four such cases; *ZNF335*, *SLC12A5*, *EYA2* and *C20ORF165*. *ZNF335* also has a CTCF enrichment on its start site (see section 5.5.1). In NTERA-D1 cells, there are four such cases; *KCNS1*, *SLC12A5*, *WFDC10A* and *C20ORF165*, and all but *KCNS1* showed CTCF enrichment on the TSS. CTCF may be part of a silencing mechanism of the above genes. CTCF enrichment and the gene expression status of the CTCF enriched genes is given in Table 5.14

CTCF-enriched genes in	Expressed	Not expressed
HeLa S3	1	8
NTERA-D1	5	4

Table 5.14. Expression profile of the genes whose start sites are enriched with CTCF in HeLa S3 and NTERA-D1 cells.

CTCF obviously operates differently in the two cell lines (Table 5.14). In HeLa S3, it most likely acts as a repressor, whereas in NTERA-D1 cells, it is also seen on the TSSs of expressed genes. Three expressed genes (*MATN*4, *RBPSUHL* and *SULF2*) that showed CTCF enrichment contain no other signal on the start site in NTERA-D1, which may suggest that these genes have a transient expression pattern whose control involves CTCF. Yet, the start of these genes was enriched with CTCF also in HeLa S3

where they are not expressed. This may be an example of how the activity of a transcription factor, in this case CTCF, depends on the context of the cell it operates in.

The antibody recognizing H4Ac showed 60 enriched spots in HeLa S3 whereas only three in NTERA-D1 cells. Figure 5.41 shows the difference between the enrichment obtained by Rabbit IgG (negative control antibody) and H4Ac in NTERA-D1 cells on a small section of the array.

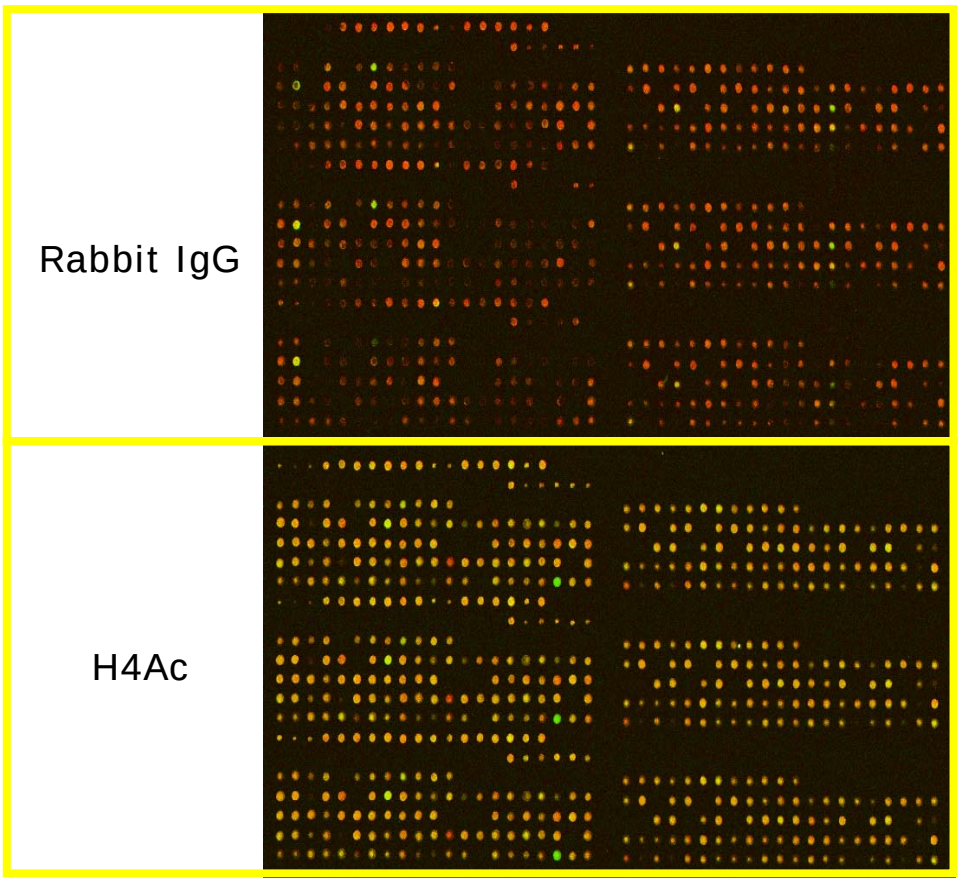


Figure 5.41 The enrichment difference between Rabbit IgG (negative control) (top) and H4Ac (bottom) antibodies on a small section of the custom-made array in NTERA-D1 cells. The green enriched spot on the bottom array carries upstream sequence of *PKRCBP1*.

The bottom array in Figure 5.41 displays a pattern as if all sites were enriched with H4Ac compared to the top array which shows Rabbit IgG enrichments. This difference may be an experimental artefact due to different working efficiencies of the

antibody. However, Figure 5.42 shows the signals on the same section of the array obtained with a non-working antibody, Sp1 (see section 5.2.1).

Sp1

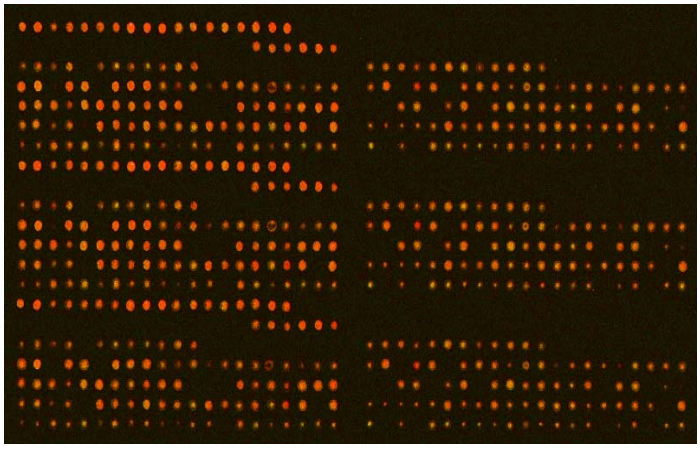


Figure 5.42. Enrichment profile of Sp1 on a subsection (the same section as in Figure 5.32) in NTERA-D1 cells.

The enrichment profile of H4Ac and Sp1 are different, that of Sp1 (non-working antibody) is similar to Rabbit IgG. This may suggest that H4Ac may be working in this cell line but the histone H4 acetylation pattern is widespread across the region. The fact that most of the other working antibodies performed relatively better in NTERA-D1 than in HeLa S3 further supports this claim. As mentioned in section 4.4.2, NTERA-D1 is established from a malignant germ cell tumour and it differentiates to neurons in response to RA (Mavilio et al., 1988; Pleasure and Lee, 1993; Segars et al., 1993). HeLa S3 cells are not from germ lines but epithelial tumour. Therefore, it is expected to see major differences in the histone code of these two cell lines.

H3K27me3 enrichment showed also very different patterns between the two cell lines. In HeLa S3, none of the gene start sites showed enrichment with H3K27me3 and the H3K27me3 enrichment was negatively correlated with that of H3K4me3 (see section 5.8.1.1). However, nine annotated start sites showed H327me3 enrichments in NTERA-D1 cells and four were also enriched with H3K4me3. Also, in HeLa S3 there were 15 spots enriched with H3K27me3 in total whereas this number was 50 in

NTERA-D1 cells. H3K27me3 is an assisting factor in polycomb-mediated gene silencing and it is shown that HOX cluster is enriched with this modification, a gene cluster which is developmentally regulated. The difference in H3K27me3 enrichment profile between the two cell lines might again be due to their different origin. The nine genes whose start sites were enriched with H3K27me3 have the same expression pattern in both cell lines, therefore H3K27me3 does not seem to play a role in the expression of these genes, but the histone enrichment profile of these genes is quite different between the two cell lines (Table 5.17).

One interesting case is *SLC12A5* where there is strong H3K4me3 on downstream of the TSS together with an H3K27me3enrichment. Also, this gene is enriched with H3K27me3 across its sequence as shown in Figure 5.43. Yet, there is a moderate H3K4me3 enrichment together with a strong H3K4me2 enrichment on the spot carrying the 3' UTR. This spot also carries a CpG island (length=1856 bp, GC%=64.3, Obs(CpG)/Exp(CpG)=0.757), a promoter prediction (FirstExon) and TSS predictions (Eponine).

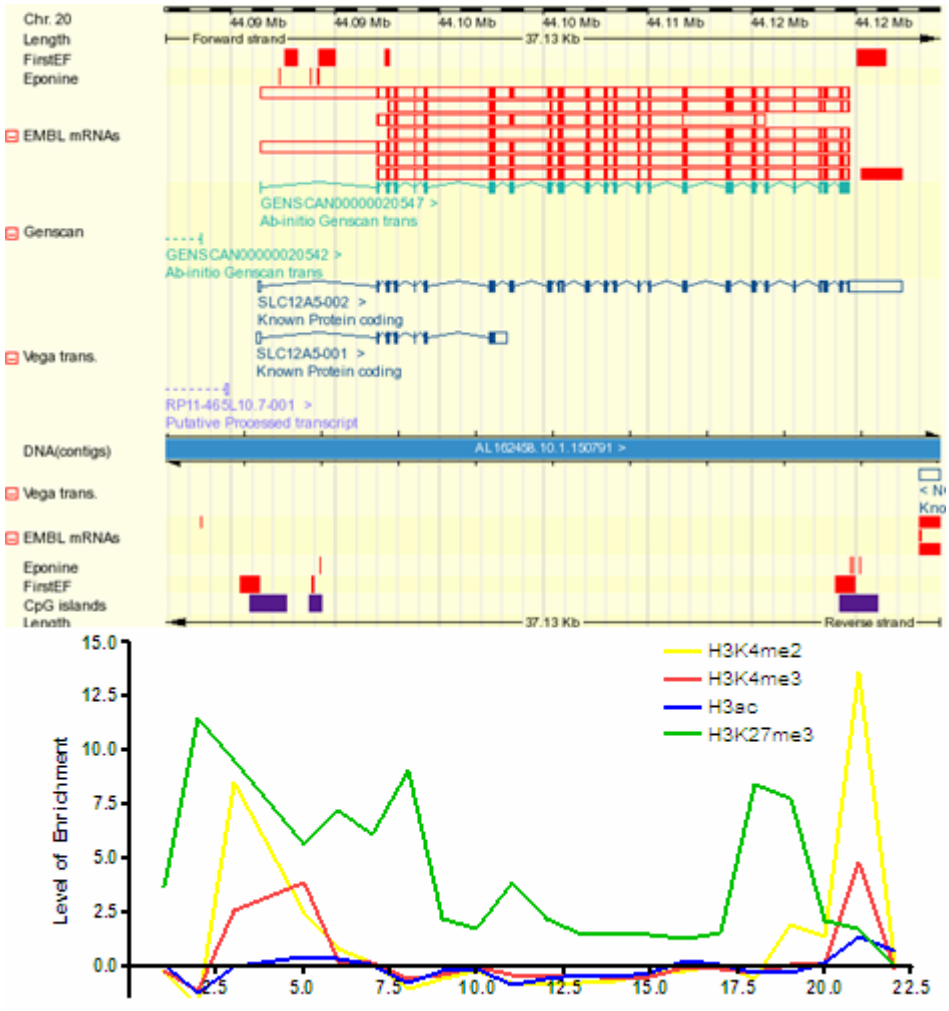


Figure 5.43 Enrichment levels on SLC12A5 gene with antibodies recognizing H3K4me2, H3K4me3, H3Ac and H3K27me3 together with the annotation of the gene taken from Ensembl Genome Browser.

SLC12A5 encodes for a membrane protein responsible for co-transporting potassium and chloride ions across the cell membrane. It has a tissue specific expression pattern, restricted to neurons in the central nervous system and retina, and it plays an important role in neuronal development (Hebert et al., 2004). There is no expression of this gene is in either HeLa S3 or NTERA-D1 cells. In HeLa S3, there is a ~2.5 fold H3K4me3 and H3Ac enrichment on the TSS, but there is no other significant enrichment with any other proteins used overall of the gene sequence. It is intriguing that NTERA-D1 requires such epigenetic marking across the entire gene, while in HeLa S3 cells such marking seems unnecessary to keep the gene silent. Such differences in epigenetic signatures in the same gene show that the histone code is

dependent on the cellular context. It might be that H3K27me3 modification interferes with an activatory complex that exists in only in NTERA-D1 cells to prevent the activation of the gene. It is still not clear why the gene has such H3K4me2 and H3K4me3 peaks on its 3' end. A possible explanation for these enrichments is the presence of a silencing mechanism by antisense RNA where the H3K4me peaks may regulate the transcription of an antisense RNA to ensure that the gene is not expressed. These results suggest that H3K27me3 has implications in regulatory mechanisms other than gene silencing.

Histone modifications such as H3K4me and H3K4me2 were found within intergenic and intra-genic regions some of which contained ultra conserved CNGs. The regulatory potential of such regions needs to be verified experimentally by either using gene reporter assays or ChIP analysis performed with antibodies recognizing common or specific enhancer binding factors such as p300 or TEF-1. The differential histone code on these elements supports the fact that different sets of histone codes are employed in different cell lines most probably depending on the activity status of the element (see section 5.6.1 and 5.6.2).

A subset of polII enriched regions in HeLa S3 cells were within inter-genic regions and the cis-acting regulatory potential was discussed in section 5.6.1. A polII enriched spot located between 42,469,627 and 42,471,927 lies within the *HNF4A* which encodes a transcription factor regulating the expression of hundreds of genes in liver and pancreas cells (Odom et al., 2004). In the beta cells of pancreas, HNF4A regulated the insulin secretion in response to glucose level. Mutations in *HNF4A* are associated with monogenic autosomal dominant non-insulin-dependent diabetes mellitus type I, also four SNPs found in the 10.7 kb region encompassing P2 promoter is shown to be associated with T2D in some populations (Bagwell et al., 2005). *HNF4A* showed an interesting enrichment pattern across the gene (Figure 5.44).

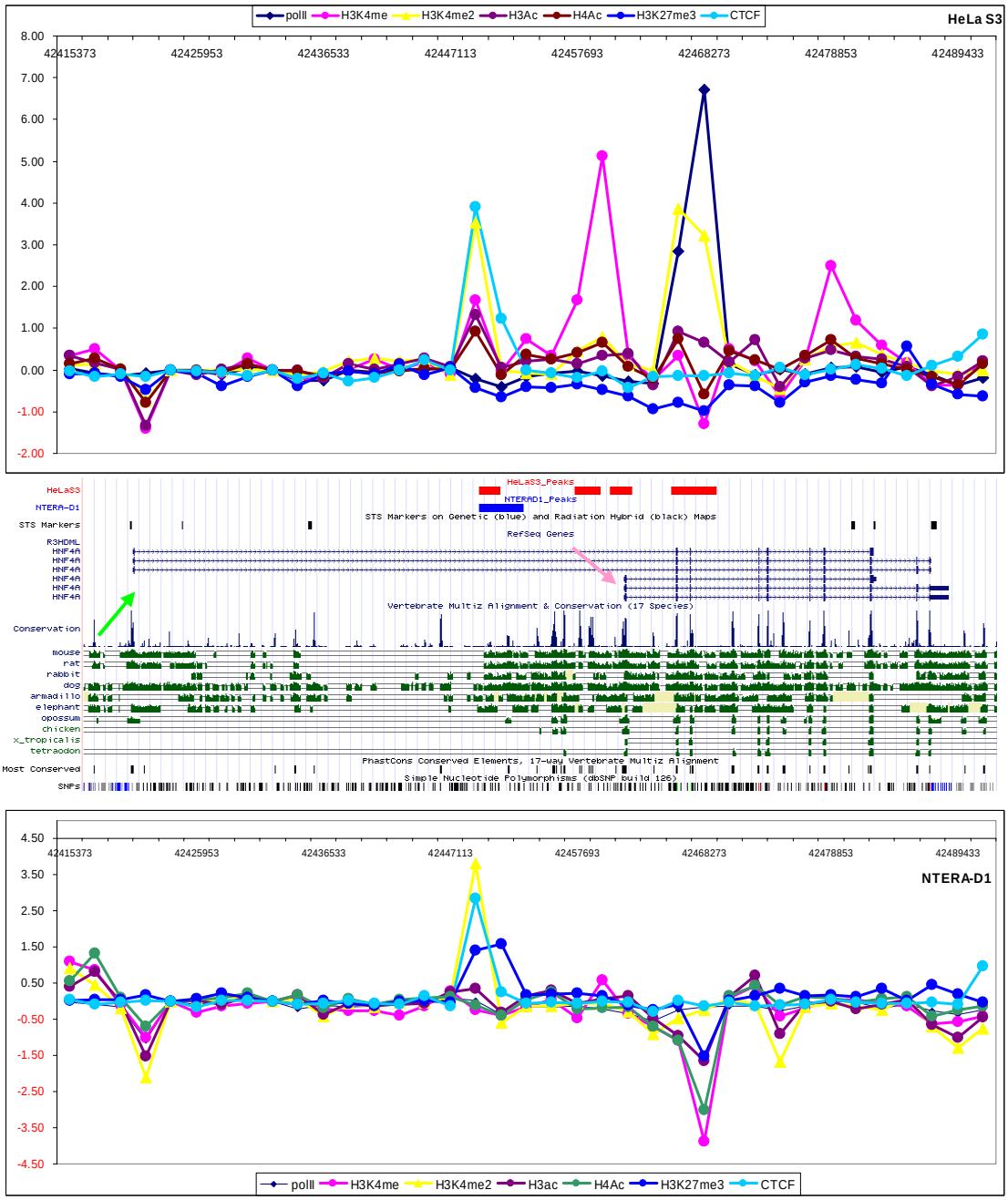


Figure 5.44. Enrichment profile across *HNF4A* in HeLa S3 and NTERA-D1 cells. The annotation is reproduced from UCSC genome browser. Green arrow denotes P2 promoter and pink arrow denotes the P1 promoter. The peaks are also displayed as custom annotation tracks (red and blue boxes).

None of the known enhancer elements of *HNF4A* showed enrichment in the two cell lines used (Mitchell et al., 2002). P2 promoter (green arrow in Figure 5.44) showed depletion for several antibodies in both cell lines. P1 promoter (pink arrow in Figure 5.44) showed H3K4me enrichment (~5.8 fold) in HeLa S3 but none in NTERA-D1 cells. The region around 32 kb upstream of the P2 promoter (first red and blue boxes in Figure 5.44) is enriched with CTCF in both cell lines. There is another region

located around 7.5 kb downstream of the P1 promoter (fourth red box in Figure 5.44) which is enriched with polII as well as H3K4me2. Interestingly, this is where most of the antibodies show a depleted signal in NTERA-D1. The polII peak in this region might again be a sign of a distal regulatory element (see section 5.6.1). Note that all these regions contain CNGs.

HNF4A is not expressed in any of the two cell lines investigated. As expected, the P2 promoter did not show any activity but it did show strong activity in synergy with the SV40 enhancer in both cell lines (see section 4.6). The cloning of the P1 promoter to pGL3-basic (enhancer-less) plasmid was unsuccessful but it was cloned to pGL3-enhancer plasmid (carrying SV40 enhancer) and P1 promoter-SV40 enhancer construct showed strong activity in both cell lines. These results suggest that *HNF4A* is under the effect of an epigenetic silencing (probably involved with intra-genic regions mentioned above) rather than a dominant trans-acting silencing mechanism. These two regions need to be investigated further in other cell lines where *HNFA* is active to see any alteration of the histone code on these elements.

In a recent study, the transcriptional map of approximately 30% of the human genome was produced at 5-nucleotide resolution using DNA microarrays (Cheng et al., 2005). Repeat-masked sequences of ten human chromosomes, including chromosome 20, were presented on high density arrays using 25-mer oligonucleotides spaced every 5 bp on average (i.e., 20 bp overlap), and the sites of transcription for poly A+ cytosolic RNA derived from eight cell lines were mapped. Approximately 9% of 74,180,611 total probe pairs detected per transcription per cell line and per chromosome, and the average number of transfrags (transcribed fragments) per cell line and per chromosome was found to be 16,864. Also, a considerable proportion of the detected transcription is cell-line specific. Strikingly, 31.8% of the detected cytosolic poly A+ sequences do not overlap with any well-characterised exon, mRNA or EST annotation

(UCSC Genome Browser, b34). Inter- and intra-genic regions that showed enrichment with several antibodies including polII and H3K4me3 were investigated whether they contain or lie within any of the transfrags obtained from Cheng et al. study. To this end, transfrags obtained from eight cell lines that are longer than 300 bp were taken (9,439) and the enrichment patterns of these fragments were investigated by all antibodies used. In total, 17 non-TSS containing enriched spots overlapped with a transfrag, and are listed in Table A.11a (HeLa S3) and Table A.11b (NTERA-D1) in Appendix A. Four and three polII enriched spots in HeLa S3 and NTERA-D1 overlapped with a transfrag. Also, seven and five CTCF-enriched spots in HeLa S3 and NTERA-D1 respectively overlapped with transfrags. It still remains unknown whether transfrags serve a function or are simply products of randomly assembled initiation machineries on sequences other than promoters. The fact that some enriched regions overlap with transfrags brings the possibility that certain protein assemblies or histone signatures might favour the assembly of polymerase II initiation machinery irrespective of the DNA sequence of the region, which will then produce such transcribed sequences. However this scenario cannot shed a light on functional role of transfrags.

This study produced a number of potential distal regulatory sequences, which need to be tested further in order to derive a regulatory map of the region. Inclusion of more cell lines to such studies will improve the coverage of such analysis since most probably many regions will produce signals only in the cells in which they function.

Transcript Type	HUGO Transcript ID	polII	H3K4me	H3K4me2	H3K4me3	H3Ac	H4Ac	H3K9me2	CTCF	Expression
RT	C20orf121-004	**7.44**	3.33	17.97	**36.72**	**14.44**	6.51			P
RT	NCOA3-001	**4.40**		6.66	**22.81**	**11.39**	3.39			P
RT	PPGB-004	**3.24**	2.94	14.29	**21.83**	**9.69**	2.53			P
RT	STK4-003	**3.68**	1.70	9.19	**20.95**	**12.13**	2.47			P
AT	UBE2C-003	**2.12**	1.97	10.55	**19.00**	**8.11**	0.00			P
RT	PIGT-003	**3.75**		7.28	**17.91**	**7.93**	3.70			P
RT	SERINC3-001	**1.95**		7.33	**17.51**	**5.66**	1.91			P
RT	UBE2C-001	**2.97**	1.97	4.21	**13.35**	**2.91**	0.00			P
RT	C20orf142-001	**0.00**		4.59	**8.96**	**3.30**	0.00			P
RT	ZSWIM1-002	**2.41**	1.89	3.04	**7.42**	**3.31**	1.74			P
AT	PPGB-001	**3.24**	2.70	4.59	**6.62**	**2.84**	2.55			P
RT	ACOT8-004	**0.00**	2.01	3.50	**6.28**	**2.19**	0.00			P
RT	ZSWIM3-001	**0.00**	2.01	3.50	**6.28**	**2.19**	0.00			P
RT	C20orf119-017	**0.00**		4.61	**6.22**	**2.90**	0.00			P
RT	SDC4-001	**5.28**	2.51	2.86	**4.46**	**2.07**	0.00			P
RT	PRKCBP1-007	**2.48**	3.10	7.18	**3.93**	**2.61**	3.28			P
RT	TOMM34-001	**0.00**	2.68	4.16	**3.73**	**2.14**	0.00			P
RT	TP53RK-001	**0.00**	2.05	4.37	**3.63**	**3.67**	0.00			P
RT	ELMO2-001	**0.00**	2.92	2.05	**2.63**	**1.59**	1.57			P
RT	C20orf67-001	**2.17**	0.00	2.16	**1.97**	**1.57**	0.00			P
RT	WISP2-002	**2.85**	4.45	2.67	**1.87**	**2.51**	2.05			P
AT	PRKCBP1-006		5.36	7.86	**1.82**	**3.41**	2.24			P
RT	YWHAB-001		2.07	2.05		**1.80**				P
RT	SLC35C2-006	**2.17**	1.81	1.66		**1.85**				P
RT	C20orf10-001									P
AT	C20orf119-004									P
AT	C20orf119-005									P
RT	DBNDD2-003									P
AT	DBNDD2-006									P
RT	CD40-001									P
AT	ELMO2-004									P
AT	ELMO2-006									P
AT	ELMO2-007									P
RT	NCOA5-001									P

RT	PI3-001		3.23	2.27						P
RT	PKIG-001									P
RT	PLTP-001		3.47				2.20			P
AT	PRKCBP1-009									P
RT	SPINLW1-001									P
RT	SPINLW1-002									P
AT	SERINC3-002									P
RT	WFDC2-002									P
RT	ZNF335-001	**3.15**	3.44	9.15	**22.69**	**12.91**	6.68		**4.70**	A
RT	NEURL2-001*	**3.24**	2.94	14.29	**21.83**	**9.69**	2.53			A
RT	SLC12A5-001				**2.61**	**2.17**				A
RT	C20orf62-001[#]				**1.73**					A
RT	EYA2-003		9.82	5.11		**1.59**	2.51			A
RT	C20orf165-001	**2.25**	1.91							A
AT	EYA2-004									A
RT	GDAP1L1-001							**1.13**	**1.92**	A
RT	HNF4A-001									A
AT	HNF4A-003		5.13							A
RT	KCNS1-001									A
RT	MATN4-001								**6.49**	A
RT	MMP9-001									A
RT	R3HDML-001									A
RT	RBPSUHL-001								**8.67**	A
RT	RIMS4-002							**1.15**		A
RT	SULF2-002									A
AT	SULF2-003								**5.15**	A
RT	SEMG1-001									A
RT	SEMG2-001		1.51							A
RT	SLC13A3-001									A
AT	SLC13A3-002									A
AT	SLC13A3-004							**1.03**		A
RT	SLC2A10-001									A
RT	SPINT3-001									A
RT	TNNC2-003		1.57							A
RT	WFDC10A-001								**4.57**	A

RT	WFDC10B-001									A
RT	WFDC11-001									A
RT	WFDC12-001									A
RT	WFDC13-001									A
RT	WFDC3-006		2.99	1.64						A
RT	WFDC5-001								**2.26**	A
RT	WFDC6-002									A
RT	WFDC8-001									A
RT	WFDC9-001								**2.14**	A
RT	ZNF334-001									A
RT	C20orf123-001				**2.40**					No probe
RT	C20orf157-001									No probe
RT	C20orf168-001									No probe
RT	SPINT4-001									No probe

Table 5.15. ChIP Signals of 83 coding transcripts obtained from nine antibodies used in HeLa S3 cells. Since H3K27me3 did not show any enrichment on any start site, it is omitted. (*) Signals coming from *NEURL2* are attributed to *PPGB* since the latter is expressed while the former is not. (#) This signal is omitted since the corresponding spot has a high sequence similarity to a ubiquitously expressed gene elsewhere in the genome.

Transcript Type	HUGO Transcript ID	polII	H3K4me	H3K4me2	H3K4me3	H3Ac	H4Ac	H3K27me3	CTCF	expression
RT	C20orf121-004	**5.15**	1.75	25.77	**79.56**	**7.10**				P
RT	NCOA3-001	**4.75**		16.28	**54.36**	**5.18**				P
RT	STK4-003			18.35	**51.90**	**4.65**				P
AT	UBE2C-003	**3.81**		1.70	**43.81**	**4.32**				P
RT	PPGB-004	**3.34**		21.91	**42.93**	**4.79**				P
RT	SERINC3-001			12.24	**39.36**	**3.89**				P
RT	ZNF335-001	**2.74**		11.50	**31.97**	**3.43**			**3.86**	P
AT	PPGB-001	**3.34**		15.98	**30.33**	**3.63**				P
RT	UBE2C-001	**3.62**		6.47	**29.56**	**1.81**				P
RT	C20orf119-017			5.26	**26.42**	**2.76**				P
RT	GDAP1L1-001			14.25	**26.25**	**3.40**			**3.30**	P
RT	ZSWIM1-002	**2.62**		4.53	**21.51**	**1.66**				P
RT	CD40-001	**2.18**		12.59	**19.97**	**1.74**		1.93		P
RT	PIGT-003	**1.98**		7.57	**19.60**	**2.20**				P
RT	C20orf142-001			7.68	**18.47**	**2.66**				P
RT	ZNF334-001			7.49	**16.58**	**1.58**				P
AT	PRKCBP1-006	**2.21**	1.67	24.62	**12.53**	**3.84**				P
RT	ACOT8-004			4.02	**10.05**					P
RT	ZSWIM3-001			4.02	**10.05**					P
RT	PRKCBP1-007	**1.75**	2.50	7.75	**9.39**	**2.37**	3.46			P
RT	PLTP-001	**3.49**			**7.49**	**1.79**				P
RT	WFDC2-002			7.40	**7.29**	**1.51**				P
RT	ELMO2-001			1.62	**6.02**					P
RT	TOMM34-001			5.47	**5.70**			1.61		P
RT	SDC4-001	**2.03**		4.21	**5.65**	**1.84**				P
RT	C20orf67-001	**1.90**		3.93	**5.61**	**1.55**				P
RT	SLC13A3-001			7.60	**5.48**					P
RT	SULF2-002			9.40	**1.91**	**1.91**				P
RT	TP53RK-001			1.67	**1.51**					P
RT	DBNDD2-003							1.60		P
RT	NCOA5-001							1.55		P
AT	PRKCBP1-009		2.72	3.50		**1.87**				P
RT	SEMG1-001			2.95						P

RT	YWHAB-001			2.82						P
RT	MMP9-001			1.59						P
RT	SLC35C2-006	**2.19**								P
RT	C20orf10-001									P
AT	C20orf119-004									P
AT	C20orf119-005									P
AT	DBNDD2-006									P
AT	ELMO2-004									P
AT	ELMO2-006									P
AT	ELMO2-007									P
RT	MATN4-001								**4.11**	P
RT	PKIG-001									P
RT	RBPSUHL-001								**5.46**	P
AT	SULF2-003								**3.79**	P
AT	SLC13A3-002									P
AT	SLC13A3-004									P
RT	SLC2A10-001									P
AT	SERINC3-002									P
RT	NEURL2-001*	**3.34**		21.91	**42.93**	**4.79**				A
RT	C20orf165-001	**1.89**		5.84	**9.51**				**2.21**	A
RT	KCNS1-001			7.24	**8.33**					A
RT	WFDC10A-001			7.99	**6.89**				**4.50**	A
RT	WFDC9-001									A
RT	C20orf62-001[#]				**3.80**					A
RT	SLC12A5-001			8.55	**2.54**			9.52		A
RT	RIMS4-002							3.86		A
RT	EYA2-003							2.04		A
AT	EYA2-004			1.61						A
RT	TNNC2-003		2.40							A
RT	HNF4A-001									A
AT	HNF4A-003									A
RT	PI3-001									A
RT	R3HDML-001									A
RT	SEMG2-001									A
RT	SPINLW1-001									A

RT	SPINLW1-002									A
RT	SPINT3-001									A
RT	WFDC10B-001									A
RT	WFDC11-001									A
RT	WFDC12-001									A
RT	WFDC13-001									A
RT	WFDC3-006									A
RT	WFDC5-001									A
RT	WFDC6-002									A
RT	WFDC8-001									A
RT	WISP2-002									A
RT	C20orf123-001			1.56	**3.91**					No probe
RT	C20orf157-001									No probe
RT	C20orf168-001									No probe
RT	SPINT4-001									No probe

Table 5.16 ChIP Signals of 83 coding transcripts obtained from nine antibodies used in HeLa S3 cells. Since H3K9me2 did not show any enrichment on any start site, it is omitted. Signals coming from *NEURL2* are attributed to *PPGB* since the latter is expressed while the former is not. (#) This signal is omitted since the corresponding spot has a high sequence similarity to a ubiquitously expressed gene elsewhere in the genome.

Cell Type	HUGO Gene ID	polII	H3K4me	H3K4me2	H3K4me3	H3Ac	H4Ac	H3K9me2	H3K27me2	expression
HeLa S3	SLC12A5				**2.61**	**2.17**				A
	RIMS4							**1.15**		A
	EYA2		9.82	5.11		**1.59**	2.51			A
	CD40									P
	TOMM34		2.68	4.16	**3.73**	**2.14**				P
	DBNDD2									P
	NCOA5									P
Cell Type	**HUGO Gene ID**	**polII**	**H3K4me**	**H3K4me2**	**H3K4me3**	**H3Ac**	**H4Ac**	**H3K9me2**	**H3K27me2**	**expression**
NTERA-D1	SLC12A5			8.55	**2.54**				9.52	A
	RIMS4								3.86	A
	EYA2								2.04	A
	CD40	**2.18**		12.59	**19.97**	**1.74**			1.93	P
	TOMM34			5.47	**5.70**				1.61	P
	DBNDD2								1.60	P
	NCOA5								1.55	P

Table 5.17 Enrichment profiles of nine genes whose start sites were enriched with H3K27me3 in NTERA-D1 cells in both cell lines.