4 IDENTIFICATION AND CHARACTERIZATION OF CORE PROMOTER ELEMENTS BY GENE REPORTER ASSAYS

Promoters carry the central regulatory information of genes, therefore their correct annotation and characterization is vital to understand gene function. The idea that gene expression might be controlled by specific regions in the genome came from Scaife and Beckwith in 1966 (Scaife and Beckwith, 1966); they identified several genomic mutations decreasing or abolishing the activity of the lac operon genes in *E. coli*. They also confirmed that these mutations act at the sequence level as the insertion of a second lac region into the genome did not relieve the mutational effect. Two years later, Ippen and co-workers used bacterial strains with mutations at the start of the lac operon and located a control region that they call promoter where the transcription initiates (Ippen et al., 1968). An elegant study from Block et al. in 1971 successfully confirmed that, indeed, transcription initiates on lac promoter, using *in vitro* transcription (Eron and Block, 1971).

In vitro transcription was the first method of choice for promoter identification in mammalian cells (Weil et al., 1979) (Manley et al., 1980). Although this method is powerful for confirmation of promoters, it is not suitable for studying its specific regulation. Another approach to assess promoter activity is to measure the RNA levels of its gene. However, RNA quantitation can be a tedious task; template amplification may not be always successful and it is often difficult to have accurate and reproducible measurements for genes with low expression.

In late 1970’s, an alternative method emerged to study promoters *in vivo* where the promoter of interest is joined with a “reporter” gene whose product can be assayed to monitor the activity of the promoter controlling the reporter gene (Ota et al., 1979). This method has been successfully adapted to mammalian cells using suitable reporter gene such as chloroamphenicol acetyltransferease (Gorman et al., 1982), luciferase (de Wet et al., 1987), green fluorescent protein (GFP) (Kain et al., 1995). In such assays, the reporter gene is encoded by a plasmid of between 4-6 kb in size with a multiple cloning site placed immediately upstream of the reporter gene and an antibiotic resistance gene for selection of recombinant clones (see section 2.1.3.1). The plasmid carrying a candidate promoter fragment in front of the reporter is transiently transfected into mammalian cells to detect promoter activity (see section 2.1.5). Typically, transfected cells are incubated for 24 – 72 h depending on the cell type, doubling time of the cells or regulatory nature of promoter of interest. Detection of the reporter protein in the cells and its expression level is then correlated with promoter activity. Alternatively, cells can be stably transfected with plasmid which integrates into the genome of the host cell and is retained beyond division (Pellicer et al., 1980). Reporter assays allow us to do large-scale promoter screening due to its rapid and scalable nature (Trinklein et al., 2003; Cooper et al., 2006). In this study, I have screened 74 candidate promoter fragments using dual luciferase reporter assay in two different human cell lines in HeLa S3 and NTERA2 clone-D1 (NTERA-D1).

4.1 Reporter Genes

The main considerations in choosing a reporter gene are (i) its activity ideally should be absent in the host cell; (ii) other enzymatic activities of the host cell should not interfere with the reporter activity; (iii) its detection assay should be rapid, reproducible and sensitive without using any hazardous chemical to the host cells.

Reporter genes are employed for assessing promoter activity (Gorman et al., 1982; de Wet et al., 1987; Kain et al., 1995), subcellular protein localization (Kain et al., 1995), to assess gene delivery methods into cells (Ewert et al., 2004) and *in vivo* detection of protein-protein interactions (Mitra et al., 1996). There are several alternative reporter genes to monitor promoter activity *in vivo* and each has advantages and disadvantages depending on its purpose of use. Here, three common reporter genes for promoter assays will be described in brief and compared.

4.1.1 Chloroamphenicol Acetyltransferease (CAT)

CAT is encoded by a bacterial drug-resistance gene and inactivates the antibiotic chloroamphenicol by acetylating at one or both of its hydroxyl groups (Shaw, 1975). First, the cells are lysed and incubated with radioactively labelled chloroamphenicol in the presence of the cofactor n-Butyryl Coenzyme A. CAT activity can then be assayed by autoradiography of the lysate subjected to thin layer chromatography (TLC). Acetylated and non-acetylated chloroamphenicol are separated by TLC, and the presence of the acetylated form correlates to the expression of CAT. In order to avoid using radioactivity, an enzyme-linked immunosorbent assay (ELISA) has been adapted to detect CAT activity (Gao et al., 2002).

4.1.2 Green Fluorescent Protein (GFP)

GFP is a 27-kDa monomeric protein from the jellyfish *Aequorea* widely used as a reporter protein in fixed and live tissues and no substrate is required for its visualization. Wild type GFP when excited by violet (395 nm) light emits green (509 nm) light which can be detected by a fluorometer. For protein localization and interactions, GFP is the reporter of choice since it is possible to detect its signal in intact cells. GFP signal cannot be enhanced enzymatically like the luciferase signal (see section 4.3.1), which leads to lower sensitivity. Yet, it has been shown that GFP is as sensitive as luciferase when the GFP expression is quantified by flow cytometry where the emitted fluorescence intensity can be measured at the single-cell level without any processing steps (Ducrest et al., 2002).

4.1.3 Luciferase

Luciferase is a 62 kDa beetle enzyme that catalyses the reaction of luciferin, to oxyluciferin in the presence of ATP and O_2 and Mg^{2+} and yellow light (560 nm) is produced as a result. It is a chemiluminescent assay since it requires a chemical modification to give the luminescence that can be detected by a luminometer. It can detect very low levels of gene expression (Promega, 2006). But, the delivery of luciferin to different cell types is difficult and unlike GFP, it is not possible to detect directly.

Nowadays, luciferase reporter assays are commonly used in monitoring promoter activity in mammalian cells and are also the method of choice in this study. Figure 4.1 shows that the bioluminescent reactions of firefly and renilla luciferase enzymes which catalyse the two luciferin molecules used in this study. Firefly luciferase, using ATP, catalyses the two-step oxidation of luciferin to oxyluciferin, which yields light at 560 nm. Renilla luciferase catalyses the oxidation of coelenterazine to coelenteramide, which yields light at 480 nm.

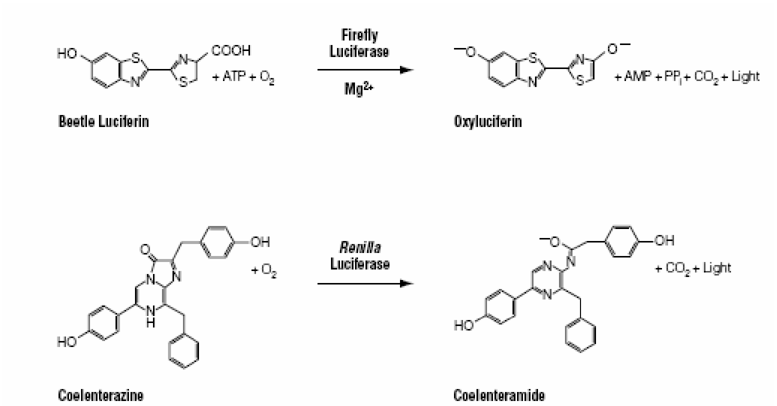

Figure 4.1. Bioluminescence reactions catalysed by firefly and renilla luciferase. This figure is reproduced from Promega® Product Technical Manual No 0.40.

Luciferase assays are faster than any other available reporter assay requiring cell lysis and it has been shown that they are up to 100 times more sensitive than CAT assays (Shaw-Jackson and Michiels, 1999). Figure 4.2 shows the linear ranges of firefly and renilla luciferases that can detect concentrations as low as 10^{-20} and 3×10^{-19} moles respectively.

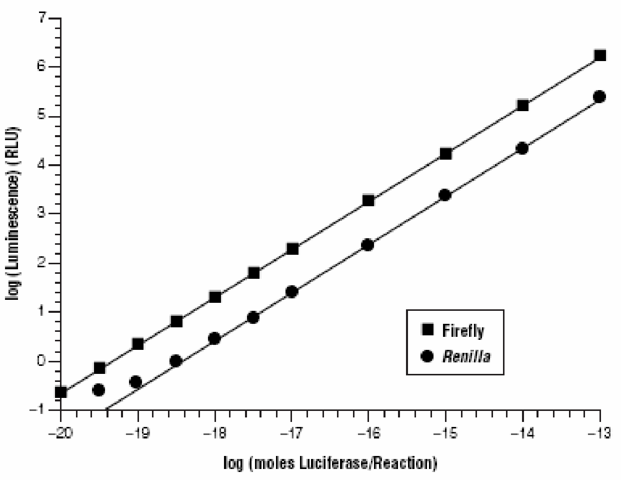

Figure 4.2. Linear ranges of firefly and renilla luciferases. The linear range of the firefly luciferase assay is seven orders of magnitude, providing detection sensitivity of 1 femtogram (approximately 10^{-20} mole) of experimental reporter enzyme. The renilla luciferase assay has a linear range of greater than five orders of magnitude and allows for the detection of approximately 30 femtograms (approximately 3×10^{-19} moles) of control reporter enzyme. This figure is reproduced from Promega® Product Technical Manual No 0.40.

Another advantage of luciferase is its short half-life (~3 h) in mammalian cells compared to CAT (~50 h) (Promega, 2006). Due to its short half-life, it can reflect better changes in its promoter activity.

4.2 Dual Luciferase Assays

Reproducibility in reporter assays mainly depends on the efficiency of plasmid delivery into host cells assuming a constant amount of transfected plasmid. There might also be factors interfering with transfection or gene expression that are inherent to the experimental system. In such cases, an internal control plasmid can be used to normalize experimental variations. This control plasmid carries a different reporter gene under the control of a constitutively active promoter and it is co-transfected with the main reporter plasmid to the host cell. Activities of these reporter genes can then be detected by separate means.

In this study, dual luciferase reporter assays have been employed where the internal control plasmid pRL-SV40 (see map in section 2.1.3.1) carries the coelenterazine reporter gene under the control of the SV40 promoter. The internal control plasmid was co-transfected with the pGL3-basic reporter plasmid (see map in section 2.1.3.1) carrying the beetle luciferin under the control of a putative promoter fragment. First, the signal obtained by the oxidation of beetle luciferin catalysed by firefly luciferase was detected, a reagent was then added to quench the firefly luciferase action and catalyse the renilla luciferase reaction which oxidizes coelenterazine. The signal from the renilla luciferase was then detected (see section 2.1.6). The normalized signal was calculated by dividing the firefly to the renilla signal. Figure 4.3 shows the signals generated by both luciferase enzymes used in dual reporter assays and the residual activity of luciferin molecule (middle column) where the firefly luciferase activity is quenched by greater than five orders of magnitude.

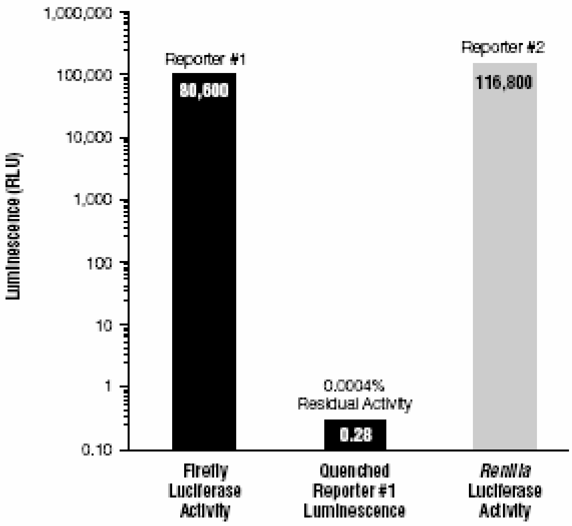

Figure 4.3. Measurement of luciferase activities before and after the addition of Stop& Glo® Reagent which quench the activity of beetle luciferase and initiate the renilla luciferase reaction. Beetle luciferase luminescence was quenched by greater than 5 orders of magnitude. This figure is reproduced from Promega® Product Technical Manual No 0.40.

All transfections were carried out in 48-well plate format. Each plasmid carrying a putative promoter fragment (construct) was transfected in triplicate in each experiment and transfection experiment was performed in duplicate. The methodology is described in section 2.1.5 in detail.

4.2.1 Positive and Negative Controls

A plasmid carrying the beetle luciferin reporter gene under the control of the strong SV40 promoter is used as positive control (pGL3-promoter plasmid, see Figure 2.5 for the map) to confirm transfection success and assess its efficiency.

A promoter is stripped of its genomic context in reporter assay studies and this may lead to non-specific promoter activity. That is because the complete regulatory information of a promoter is encoded in its histone code and that of distal genomic regulatory elements. To estimate the degree of non-specific promoter activity, five randomly selected inter- or intra-genic fragments with no known or predicted promoter activity were also included in the study as negative controls. Their observed activation levels were used to determine the degree of noise in the applied reporter assay methodology.

4.2.2 Determining Promoter Activities

Transfection of each promoter construct (with or without SV40 enhancer) was performed in triplicate per experiment and two experiments were performed per construct. Thus, in total six replicates were generated per construct. The signal from each well is calculated by dividing the luciferin (Luc) to coelenterazine (renilla, Ren) signal (Equation 4.1). The mean raw signal of ith promoter in jth experiment (Mean_Raw_Signal$_{ij}$) is calculated as in Equation 4.2. (k is the replicate index within one specific experiment).

$$Raw\_Signal_{ijk} = \frac{Luc_{ijk}}{\mathrm{Re}\,n_{ijk}} \qquad \text{Equation 4.1}$$

$$Mean\_Raw\_Signal_{ij} = \left(\sum_{k}^{3} Raw\_Signal_{ijk} \right) \times \frac{1}{3} \qquad \text{Equation 4.2}$$

Then, the mean signal of each negative control is calculated according to Equation 4.2 and the negative control construct that gives the highest signal is taken as the background signal (Background$_{ij}$). The background signal was subtracted from the signal of each promoter to remove the background noise (Equation 4.3).

$$Signal_{ij} = Mean\_Raw\_Signal_{ij} - Background_{ij} \qquad \text{Equation 4.3}$$

The signals taken from each experiment (Signal$_{i1}$, Signal$_{i2}$) are averaged out to calculate the mean signal of the ith promoter (Equation 4.4).

$$Mean\_Signal_{i} = \frac{\left(\sum_{j=1}^{2} Signal_{ij} \right)}{2} \qquad \text{Equation 4.4}$$

The standard deviation of the mean raw signal of the ith promoter at the jth experimental replicate is calculated according to (Equation 4.5).

$$\sigma_{ij} = \sqrt{\sum_{k=1}^{3} \frac{(Raw\_Signal_{ijk} - Mean\_Raw\_Signal_{ij})^2}{2}} \qquad \text{Equation 4.5}$$

Note that this expression employs the unbiased estimate method where the square root of squared deviations of each measurement from the mean is divided by *n-1* where n is the number of measurements (n=3 in this case) (Kenney, 1962)

To determine confidence intervals for whether a fragment is a promoter or not, we assume that the measured firefly/renilla luciferase signal from a promoter inherently contains some background activity. This background activity is set as the activity of the negative control fragment that gives the highest firefly/renilla luciferase signal. Since we assume that we can measure the background independently, the errors in their measurement will be independent as well. This means that the standard deviation (σ) of a real promoter activity will be the sum of the standard deviation of the measured signal and the standard deviation of the background (Equation 4.6) (Abramowitz, 1972).

$$\bar{\sigma}_{ij} = \sqrt{\left(\sigma_{Signal_{ij}}\right)^2 + \left(\sigma_{Background_{ij}}\right)^2} \qquad \text{Equation 4.6}$$

Equation 4.6 will give the standard deviation of the mean signal for each promoter in one experiment. The standard deviation of the replicated measurement will be simply the sum of the standard deviations of each measurement in each experiment (Equation 4.7).

$$\sigma_i = \frac{\sqrt{\sum_{j=1}^{2} (\bar{\sigma}_{ij})^2}}{2} \qquad \text{Equation 4.7}$$

If we assume that all the errors in measurements follow a Gaussian distribution, then our confidence intervals to decide whether a fragment is a promoter or not, i.e. showing an activity higher than the background are as follows (Abramowitz, 1972);

- Mean_Signal$_i$ ± σ_i → 68.3 % confidence that it is a promoter
- Mean_Signal$_i$ ± $2\sigma_i$ → 95.4 % confidence that it is a promoter
- Mean_Signal$_i$ ± $3\sigma_i$ → 99.7 % confidence that it is a promoter.

4.3 Cloning of candidate promoter fragments

Each candidate promoter fragment cloned to pGL3-basic plasmid was approximately 300 bp in size including 250 bp upstream and 50 bp downstream of the annotated TSS. Here, only the activity of extended core promoters was investigated since proximal promoter elements (especially between 500 bp to 1000 bp upstream of TSS) often have potential silencer elements (Cooper et al., 2006). Although core promoter region is typically accepted as 100 bp upstream to 50 bp downstream of TSS, here a longer fragment size was used to include possible activatory binding sites to increase the chance of obtaining a promoter activity from the selected fragments.

The 300 bp putative promoter fragments do not contain the complete proximal promoter region. Therefore, fragments that require enhancer elements located in proximal regions for activation cannot be detected using gene reporter assays. In an attempt to recover the activity of such fragments, 300 bp putative promoter fragments were also cloned to pGL3-enhancer plasmid which carries SV40 enhancer (see map in section 2.1.2.9). SV40 enhancer is a ubiquitously active regulatory element which has binding sites for several common activator transcription factors (see section 1.2.2). SV40 enhancer may recover the activity of candidate core promoter fragments that do not contain activatory proximal promoter elements. Also the differential response of candidate promoters to SV40 enhancer was investigated for further characterization of binding motifs in these putative fragments.

To investigate the activation recovery ability of SV40 enhancer, I cloned 17 putative promoter fragments that were 600 bp long (550 bp upstream and 50 bp downstream of the annotated TSS) to pGL3-basic plasmid and compared their activity to those of 300 bp with or without SV40 enhancer (Figure 4.4) in HeLa S3 cells. Four 300 bp constructs did not give activity, but showed activity when they were cloned in front of SV40 enhancer or when the proximal promoter was included (600 bp construct). Two constructs did not show any activity in any configuration and the remaining 11 constructs showed activity with all types of constructs. This suggested that SV40 enhancer can successfully replace the proximal promoter elements of core promoters.

As mentioned earlier, there are 103 protein coding genes and 177 coding transcripts utilizing different promoters in the 20q12-13.2 region (section 3.1). Primers were designed to amplify extended core promoter regions of 132 coding transcripts for cloning (see section 2.1.1). Of these, 109 were successfully amplified by PCR (see section 2.1.2.1). In total, 74 and 76 fragments were successfully cloned to pGL3-basic and pGL3-enhancer plasmids respectively and 71 candidate promoter fragments were cloned to both pGL3-basic and pGL3-enhancer plasmids.

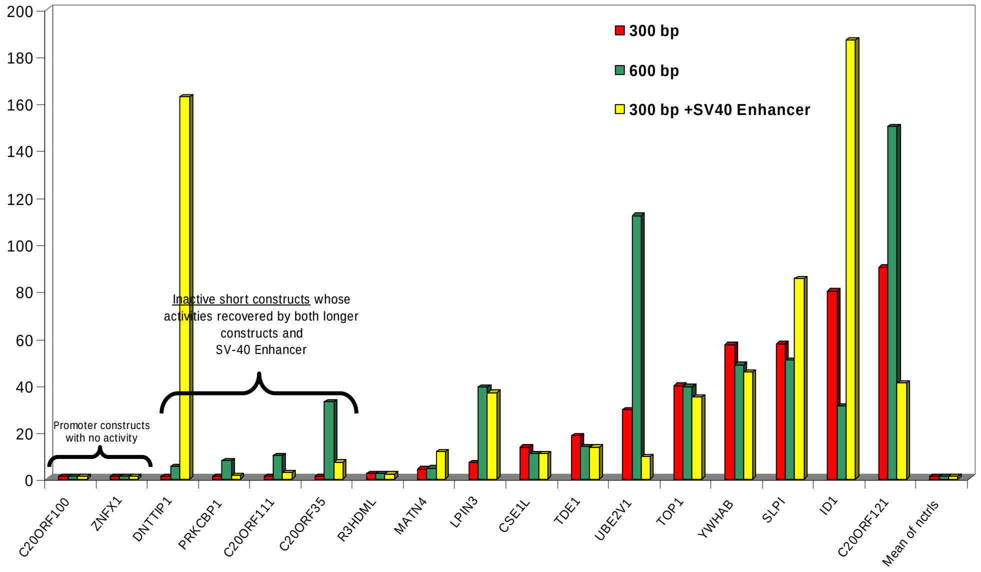

Figure 4.4. Comparison of activities between candidate promoter fragments of 300 and 600 bp in size, and 300 bp fragments cloned together with SV40 enhancer.

4.3.1 Optimisation of Reporter Assays

Selection of Reporter Gene

The use of GFP was considered as a reporter gene first since it can be detected in intact cells using a fluorescent microscope and its signal can be detected using a fluorometer without lysing cells (see section 4.1). Six promoter constructs, whose details are given in Table 4.1, were cloned to pEGFP-1 (Clontech, #6086-1) plasmids and pEGFP-N1 (Clontech, #6085-1), which carries GFP under the control of the strong CMV promoter and was used as positive control.

Four fragments showed activity higher than the background (nctrl2), but the activities between these promoters did not differ significantly, which suggested that the GFP reporter assay can only tell promoter status but it is not sensitive enough to determine the fold difference between promoters of different strength. Therefore GFP reporter assays were abandoned and dual luciferase assays were used instead.

Gene Name	Promoter prediction	CpG island prediction	BAC Clone Name (Ensembl)	Clone coordinates of the cloned region for promoter activity	Coordinates of the cloned region relative to TSS of the gene
ZHX3	yes	yes	DJ796I11	65473..67029	-940..+616
LPIN3	no	no	DJ450M14	11949..12990	-398..+643
ZNFX1	yes	yes	DJ66I20	78596.. 80221	-743..+882
C20orf130	yes	yes	DJ620E11	8596.. 9636	-546..+236
C20orf10	no	no	DJ453C12	89414..90400	-737..+249
nctrl2	no	no	DJ148H17	53971..54733	-

Table 4.1. Details of promoter fragments cloned to pGFP-1 reporter vector. TSS is denoted as +1.

Transfections were performed according to protocol in section 2.1.5 and results are shown in Figure 4.5.

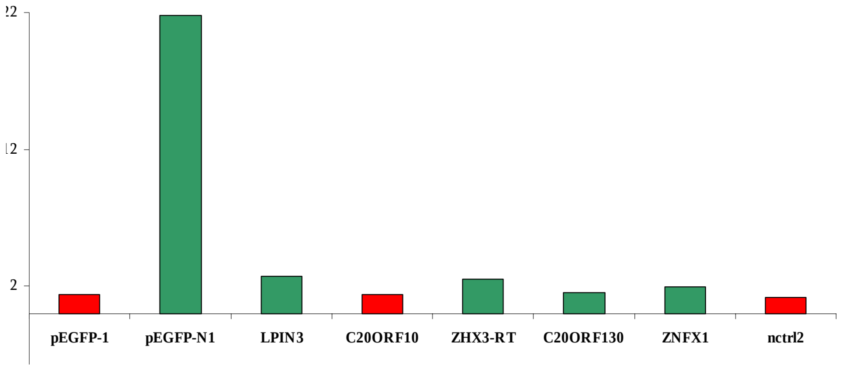

Figure 4.5 Results of transfection with GFP reporter gene. PEGFP-N1 is the positive control plasmid which carries strong CMV promoter. Fragments that did not give significant activity over background (nctrl2) were shown in red.

Optimisation of Dual Luciferase Assays

A key step in reporter assays is the transfection. A series of exprements were carried out to optimise transfection conditions. These included tests for optimising:

- the amount of internal control plasmid
- plate format and
- the number of cells to be transfected per well.

In brief, optimised conditions include the transfection of 50 µg of pRL-SV40 plasmid (internal control plasmid) with 100 µg of pGL3-basic or enhancer plasmids carrying a putative promoter fragments. Transfections were performed in 48-well plate format instead of 96-well format since the former gave better reproducibility between replicates with 2x10^4 cells per well (HeLa S3 or NTERA-D1). The method is detailed in section 2.1.5.

4.4 Cell lines

4.4.1 HeLa S3

HeLa S3 is a sub-clone of parent HeLa cell line which is a cervical carcinoma cell line. HeLa S3 is an epithelial cell line and it grows as monolayers. This cell line is commonly used and is chosen to be used for ENCODE consortium as well. HeLa S3 is also relatively easy to transfect with foreign DNA. Therefore this cell line has been selected for this study.

4.4.2 NTERA2 clone D1

NTERA-2 clone D1 (NTERA-D1) is a pluripotent embryonal carcinoma cell line. It is a sub-clone of NTERA-2 cells which were originally isolated from a lung metastasis of 22 year old patient with primary embryonal carcinoma of the testis. Treatment of this cell line with retinoic acid (RA) results in differentiation to neuronal and other cell types (Mavilio et al., 1988; Pleasure and Lee, 1993; Segars et al., 1993). A testis cell line was desired as the second cell line in this study since testis cDNA libraries contained the highest number of transcribed sequences in the region (Stavrides, 2002). NTERA-D1 is one of the two commercially available testis cell line. The other line (Hs1-tes, ATCC) is a normal human testis cell line. Hs1-tes cells were the first choice but they have a doubling time of ~3 days and they also have finite life span. Therefore this cell line was abandoned due to its incompatibly with large-scale experimental approaches, and NTERA-D1 cell was used instead.

4.5 Transfection Results

4.5.1 Promoter Activities in HeLa S3

As mentioned 74 candidate promoter fragments were successfully cloned to pGL3-basic plasmid. Their promoter activities were assessed using dual luciferase assays in HeLa S3 cells. Out of the 74 constructs, 30 (40.5%) showed an activity with 99.7% confidence.

Figure 4.6 shows background subtracted activities of the 74 (candidate) promoters as calculated in section 4.2.2. In Figure 4.7, confidence levels are shown for each promoter; each bar represents the background subtracted mean signal in units of its standard deviation. In this study, a candidate fragment was accepted as a promoter if its activity is higher than three times of its standard deviation (corresponding to 99.7% confidence level), which means that fragments with an acceptance rate (background subtracted mean signal/its standard deviation) of greater than 3 in these units are accepted as promoters.

The confidence level of a fragment does not depend only on its actual signal, it also takes the uncertainty of the actual signal (its standard deviation) into consideration. This means that the more consistent the activity of a fragment among replicates, the higher a confidence level it will have.

4.5.2 Promoter Activities in NTERA-D1

As in 4.5.1, candidate promoters were also assessed in NTERA-D1 cells where 45 of the 74 fragments tested (60.8%) showed a significant activity. This figure is higher than HeLa S3 cells where only 40.5% of the constructs showed activity which suggests that the testis cells indeed exhibit a diversified transcriptome (Jongeneel et al., 2005). Background subtracted activities of all fragments is shown in Figure 4.8. Also, Figure 4.9 displays confidence intervals for each candidate promoter.

4.5.3 Comparison of promoter activities between the cell lines

Out of 74 candidate promoters, 30 (40.5%) showed activity in both cell lines. There are 16 inactive promoters in HeLa S3 which showed activity in NTERA-D1 cell line. All promoters that showed activity in HeLa S3 were also active in NTERA-D1.

4.5.3.1 Transcript Type and Promoter Activity

Of the 74 candidate promoters, 50 and 24 correspond to representative and alternative transcripts respectively. Table 4.2 summarizes the activity status of both transcript types in the two cell lines. While the promoters of 14 representative transcripts inactive in HeLa S3 were active in NTERA-D1, this was the case for only promoters of two alternative transcripts. This observation is supportive of alternative transcripts being tightly-regulated with restricted expression although the explanation may simply be incomplete annotation.

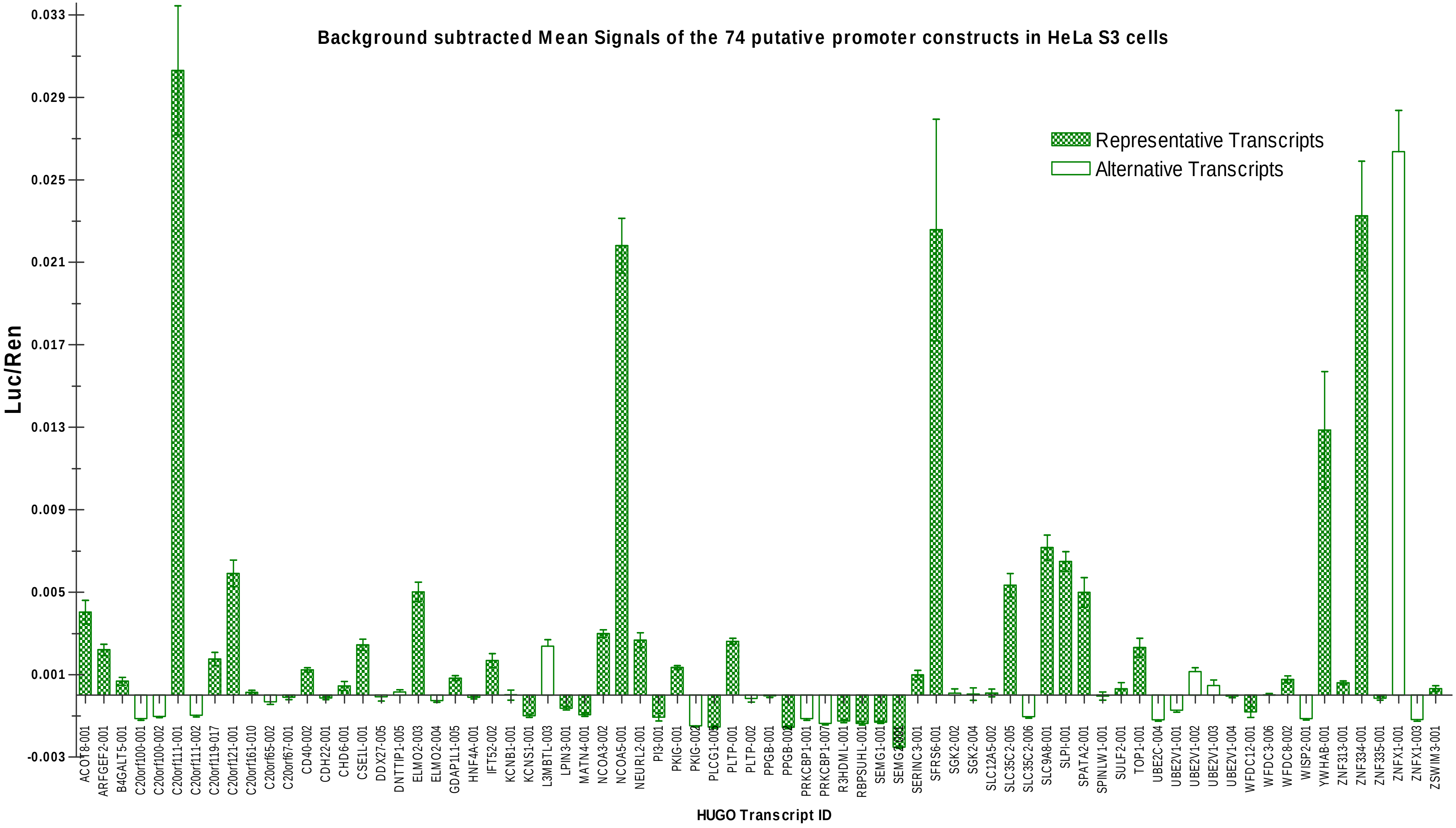

Figure 4.6 Background subtracted Luciferase/Renilla signals of 74 candidate promoters in HeLa S3 cell line. Representative transcripts (RTs) are shown with columns with crossed pattern and alternative transcripts (ATs) are shown with unfilled columns.

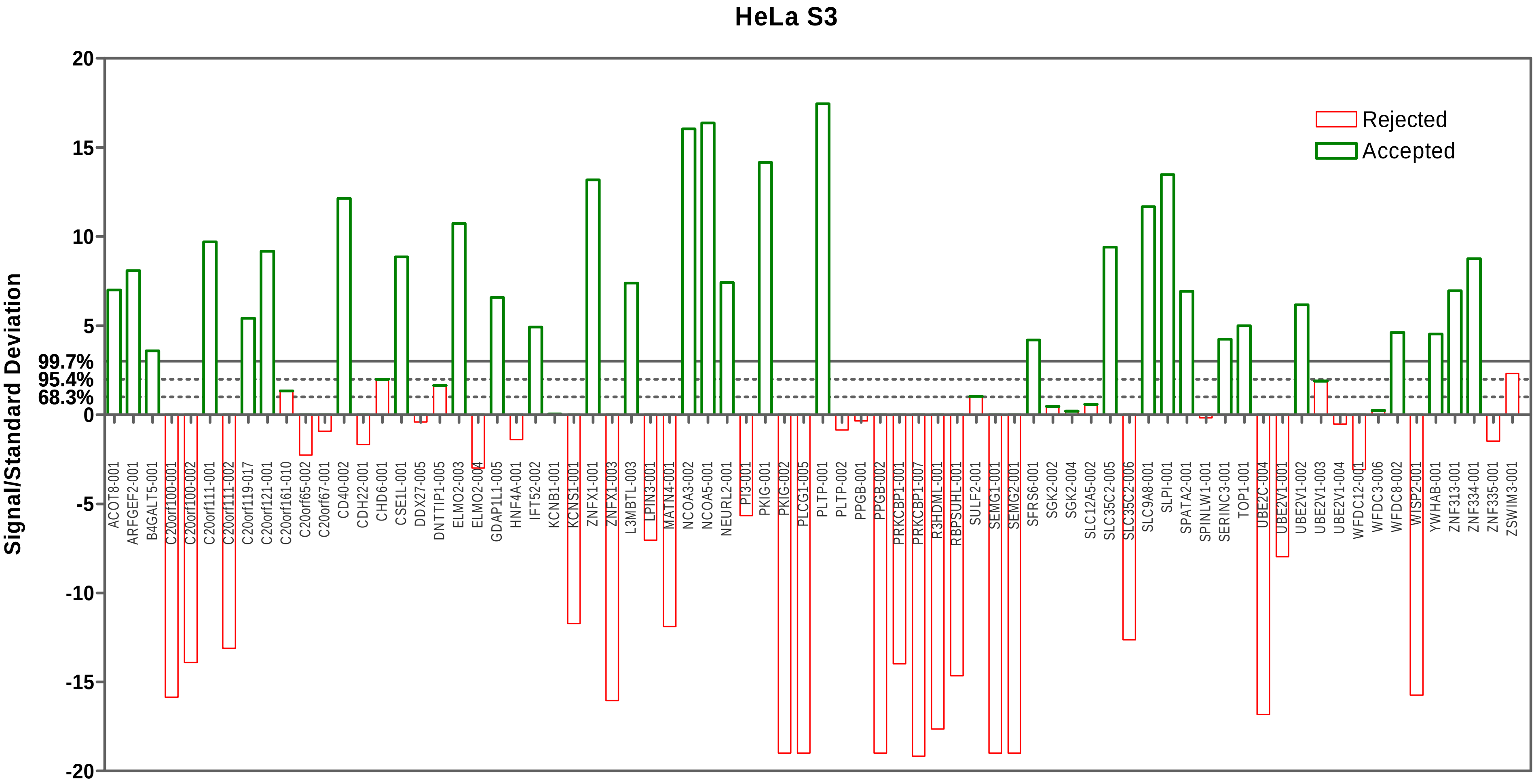

Figure 4.7 Confidence levels to accept (green bars) or reject (red bars) a fragment as promoter in HeLa S3. Here, fragments which shows 3*σ higher than the background are accepted as a promoter and the confidence level for this decision is 99.7% assuming the error distribution follows Gaussian. Dotted lines show lower confidence levels. Note that rejection rates smaller than -10 is not shown on the plot.

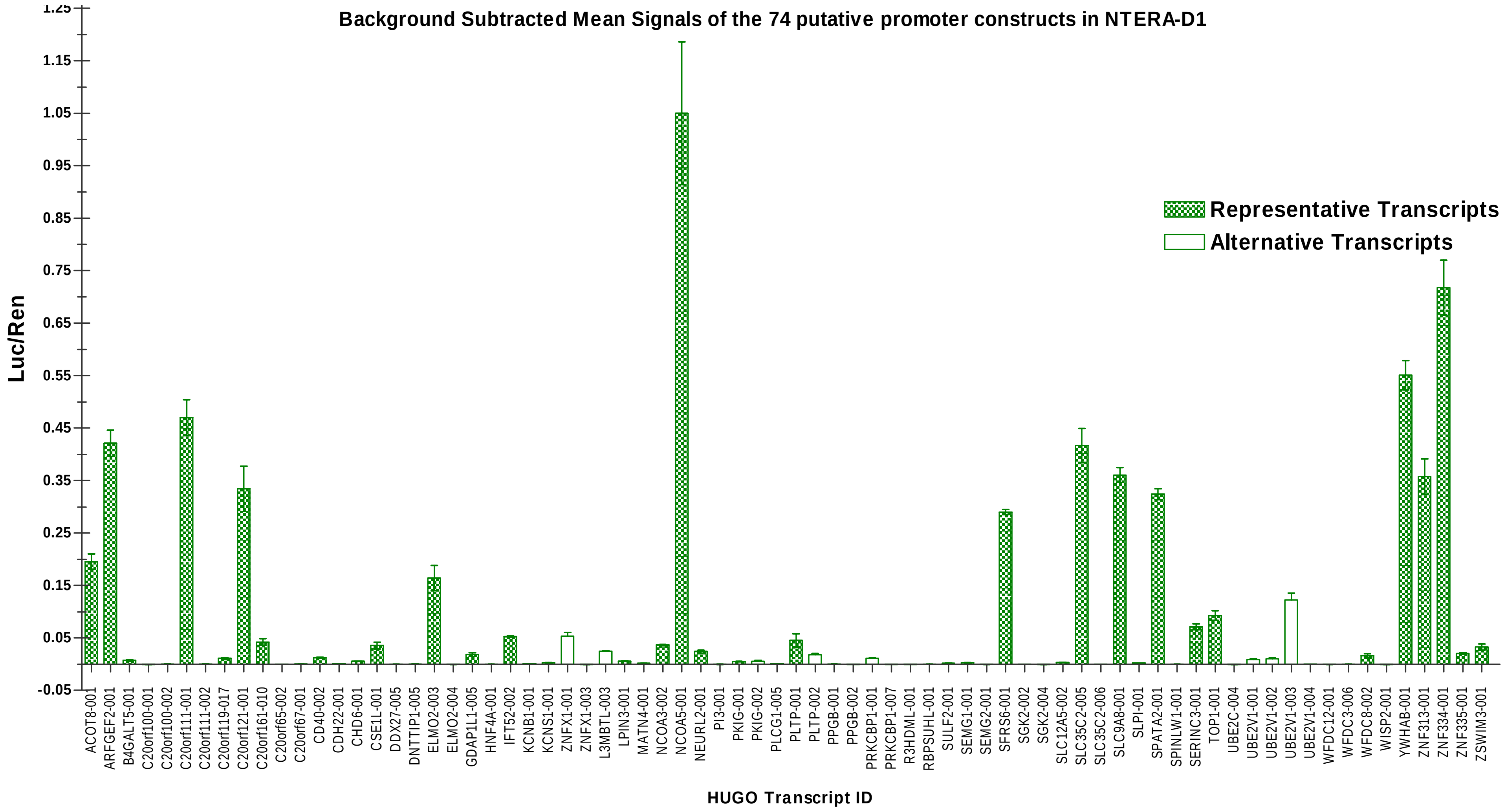

Figure 4.8 Background subtracted Luciferase/Renilla signals of the 74 candidate promoters in NTERA-D1 cell line. Representative transcripts (RTs) are shown with columns with crossed pattern and alternative transcripts (ATs) are shown with unfilled columns.

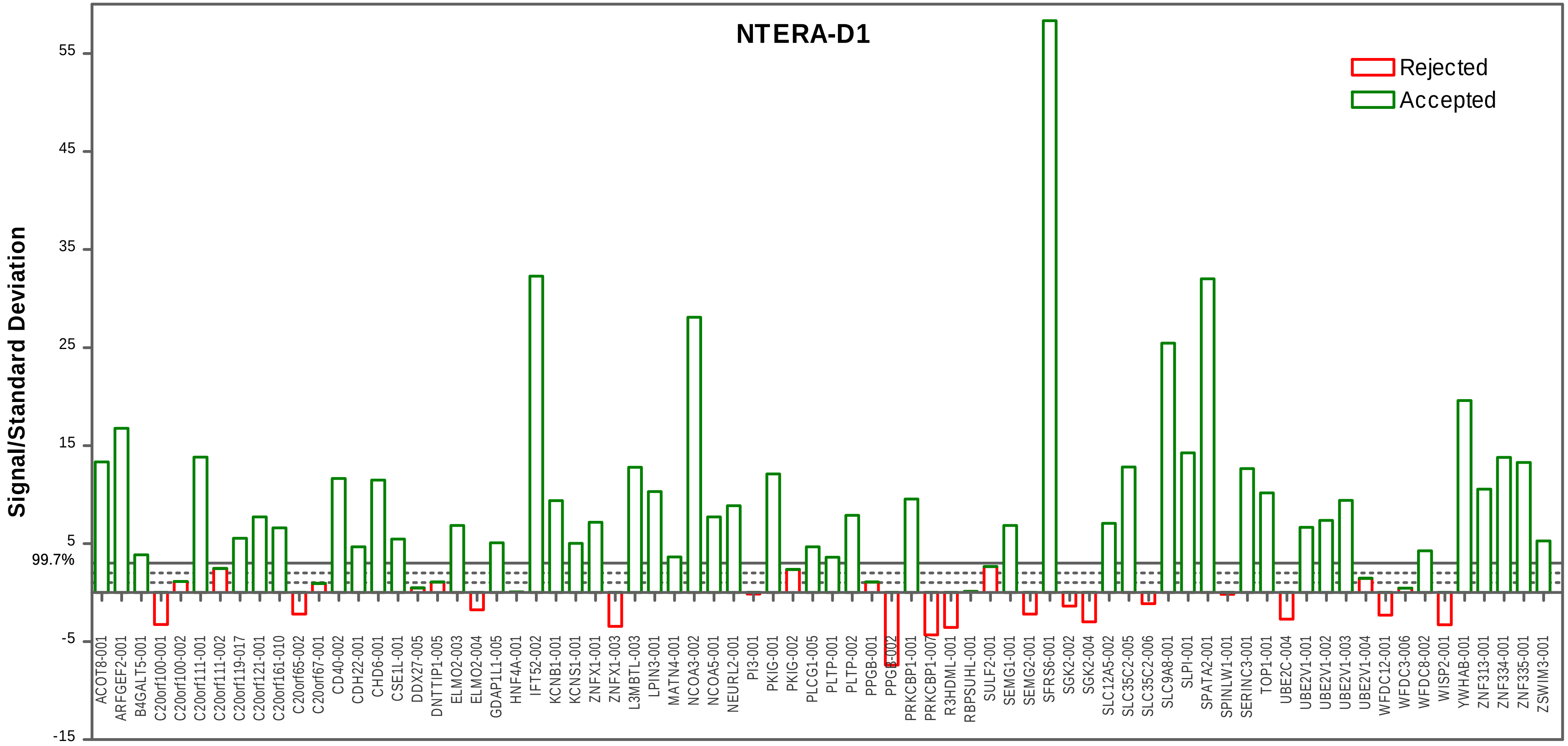

Figure 4.9 Confidence levels to accept (green bars) or reject (red bars) a fragment as promoter in NTERA-D1 cell line. Here, fragments which shows 3*σ higher than the background are accepted as a promoter and the confidence level for this decision is 99.7% assuming the error distribution follows Gaussian. Dotted lines show lower confidence levels.

	HeLa S3			**NTERA-D1**		
Transcript Type/Promoter Status	Active	No activity	Total	Active	No activity	Total
Representative Transcripts	27 (54%)	25	52	42 (81%)	10	52
Alternative Transcripts	3 (13%)	19	22	5 (29%)	17	22
All Transcripts	30 (42%)	44	74	47 (64%)	27	74

Table 4.2 Promoter activities and transcript types

4.5.3.2 Expression Profile and Promoter Activity

As discussed in section 3.4, both cell lines were profiled with Affymetrix Expression Arrays. Expression profiles of the 50 genes (corresponding to 74 promoter fragments) were correlated with their promoter activity and the results are summarized in Table 4.3. While 70% of the expressed genes in HeLa S3 showed an activity in reporter assays, this number is 89% for NTERA-D1 cells.

	HeLa S3			**NTERA-D1**		
Promoter Activity	Active	No activity	Total	Active	No activity	Total
Expressed Genes	21	9	30	31	4	35
Not Expressed Genes	6	14	20	8	7	15

Table 4.3 Expression Profile and Promoter Activity of 50 genes in HeLa S3 and NTERA-D1 cell lines.

A higher proportion of expressed genes compared to non expressed genes showed promoter activity in reporter assays. Genes that are expressed but with no promoter activity reported here may correspond to promoters with an epigenetic or distal activation mechanism; their promoter cannot be activated when stripped from its genomic environment. Also, 30% and 53.3% of promoters of non-expressed genes showed a significant activity in HeLa S3 and NTERA-D1 cell lines respectively. This result is expected since reporter assays cannot fully mimic the endogenous status of a promoter since the promoter lacks its histone code and cis-acting distal regulatory signals that might have a repressive effect.

4.6 Promoter Activities in synergy with SV40 Enhancer

In section 4.5, I described the activity of 74 candidate promoter fragments in pGL3-basic plasmids. Out of these, 46 showed promoter activity in at least one of the two cell lines. As described in section 4.3, 76 candidate promoter fragments (includes 71 of the above) were cloned in to pGL3-enhancer plasmid which contains the SV40 enhancer, in order to examine changes in promoter activity due to the enhancer.

4.6.1 Promoters in synergy with SV40 Enhancer in HeLa S3 cells

First, the effect of SV40 enhancer was assessed in HeLa S3 cells. Out of 76, 46 (60.5%) fragments drove the expression of the luciferase reporter gene under the effect of enhancer. The background subtracted mean activities for these constructs are shown in Figure 4.10. As expected, the magnitude of signals obtained from promoter-enhancer constructs is much stronger (approximately six fold) than of those obtained without enhancer. Figure 4.11 displays the confidence levels applied to accept whether a given construct gave or not activity under the effect of the enhancer. Constructs accepted as responsive to the enhancer if they showed activity 3*σ higher than the background signal (representing 99.7% confidence level assuming the errors follows Gaussian distribution).

As mentioned, 71 promoters were tested both with and without enhancer. Out of these 71, 24 (corresponding to 21 representative and 3 alternative transcripts) showed activity with and without enhancer, whereas 20 (16 representative and 4 alternative transcripts) showed activity only in the presence of the enhancer. Also, 4 active promoters did not show any activity in synergy with the enhancer, and 23 promoters (9 representative and 14 alternative transcripts) did not show any activity with or without the enhancer.

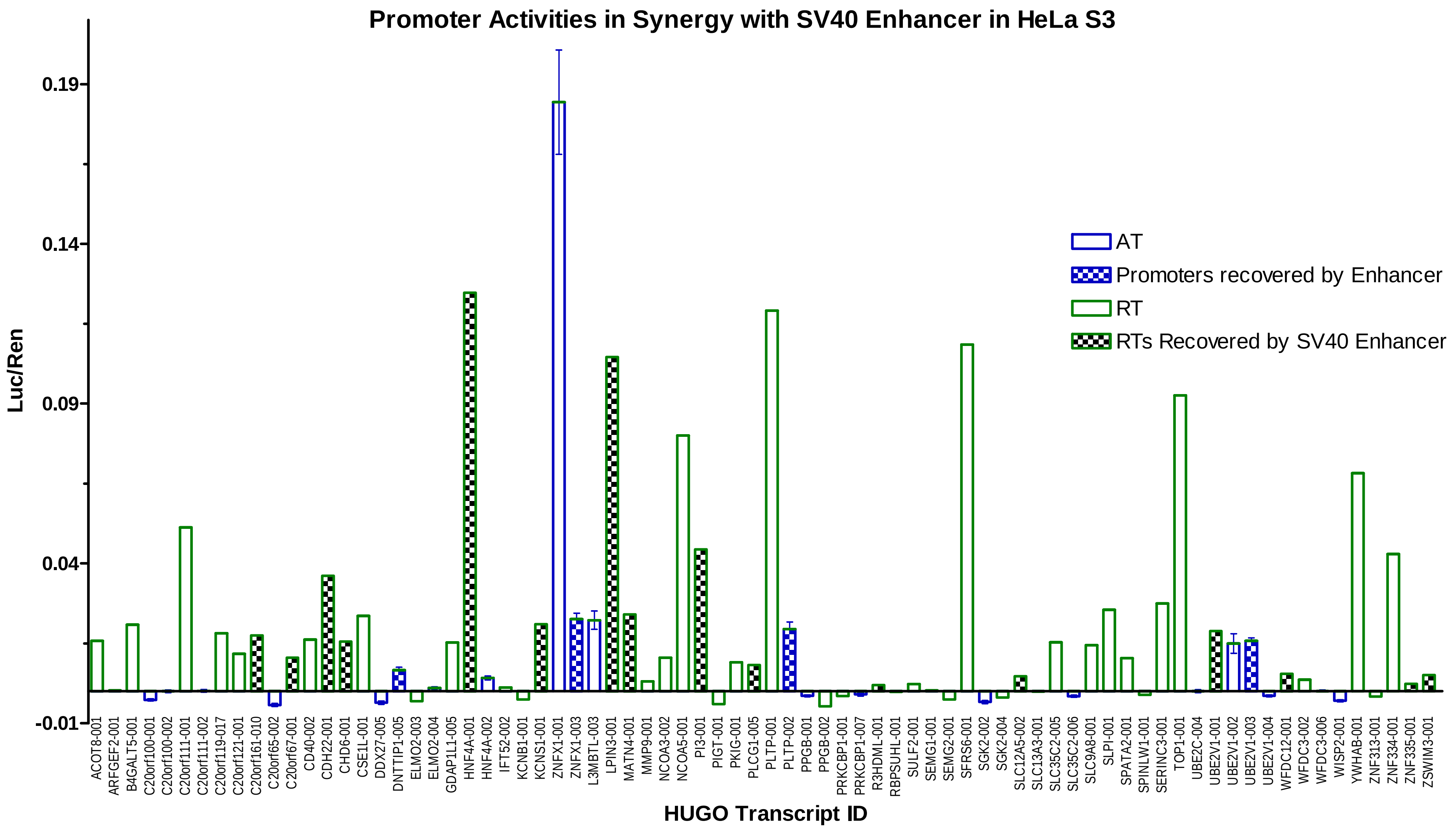

Figure 4.10 Background subtracted activity of 76 candidate promoter fragments in synergy with SV40 Enhancer in HeLa S3 cells. The blue bars represents the promoter activities of alternative transcripts and the bars with check pattern are the fragments which showed activity only in synergic with the enhancer.

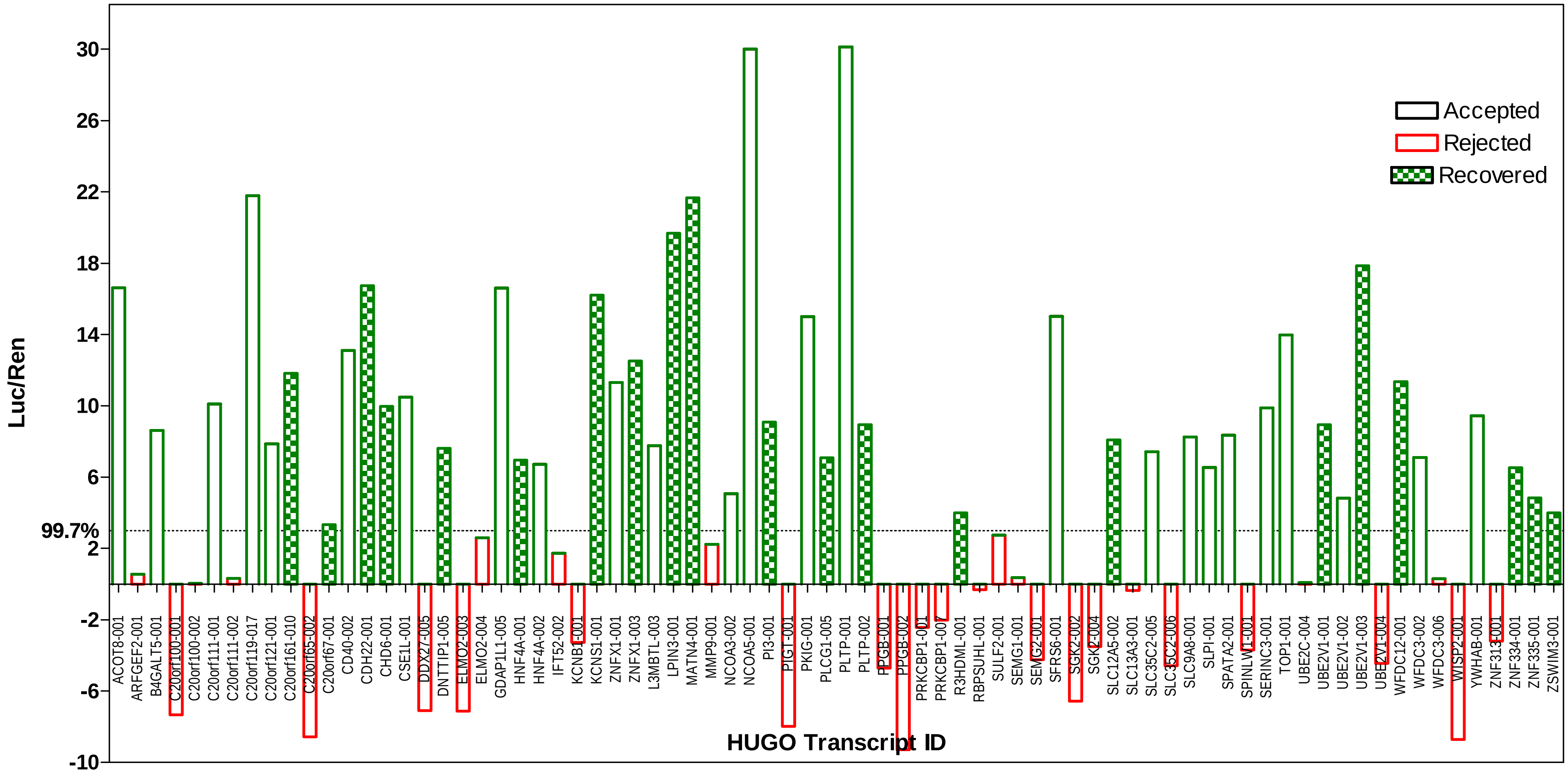

Figure 4.11 Confidence levels to accept (green bars) or reject (red bars) the activation response in HeLa S3 cell line. Here, fragments which shows 3*σ higher than the background are accepted as activated by the enhancer and the confidence level for this decision is 99.7% assuming the error distribution follows Gaussian. Bars with squared patterns denote the promoters that gave activity only in the presence of the enhancer.

4.6.2 Promoters in synergy with SV40 Enhancer in NTERA-D1 cells

The promoter activities of the 76 constructs were also assessed under the effect of the SV40 enhancer in NTERA-D1 cells and 61 (80%) of them showed activity. The background subtracted mean activity of these candidate promoters is shown in Figure 4.12, and Figure 4.13 displays confidence levels of the constructs to the responsiveness to the enhancer.

Of the 76 constructs, 71 were also assessed without enhancer. Of these, 42 showed activity with the enhancer. There are 13 putative promoters (representing 6 representative and 6 alternative transcripts) that showed activity only in synergy with the enhancer and 14 promoters (representing 5 representative and 9 alternative transcripts) that showed no activity with or without the enhancer. Also, 2 constructs (ELMO2-003, WFDC3-006) that showed activity without the enhancer were not active in the presence of the enhancer.

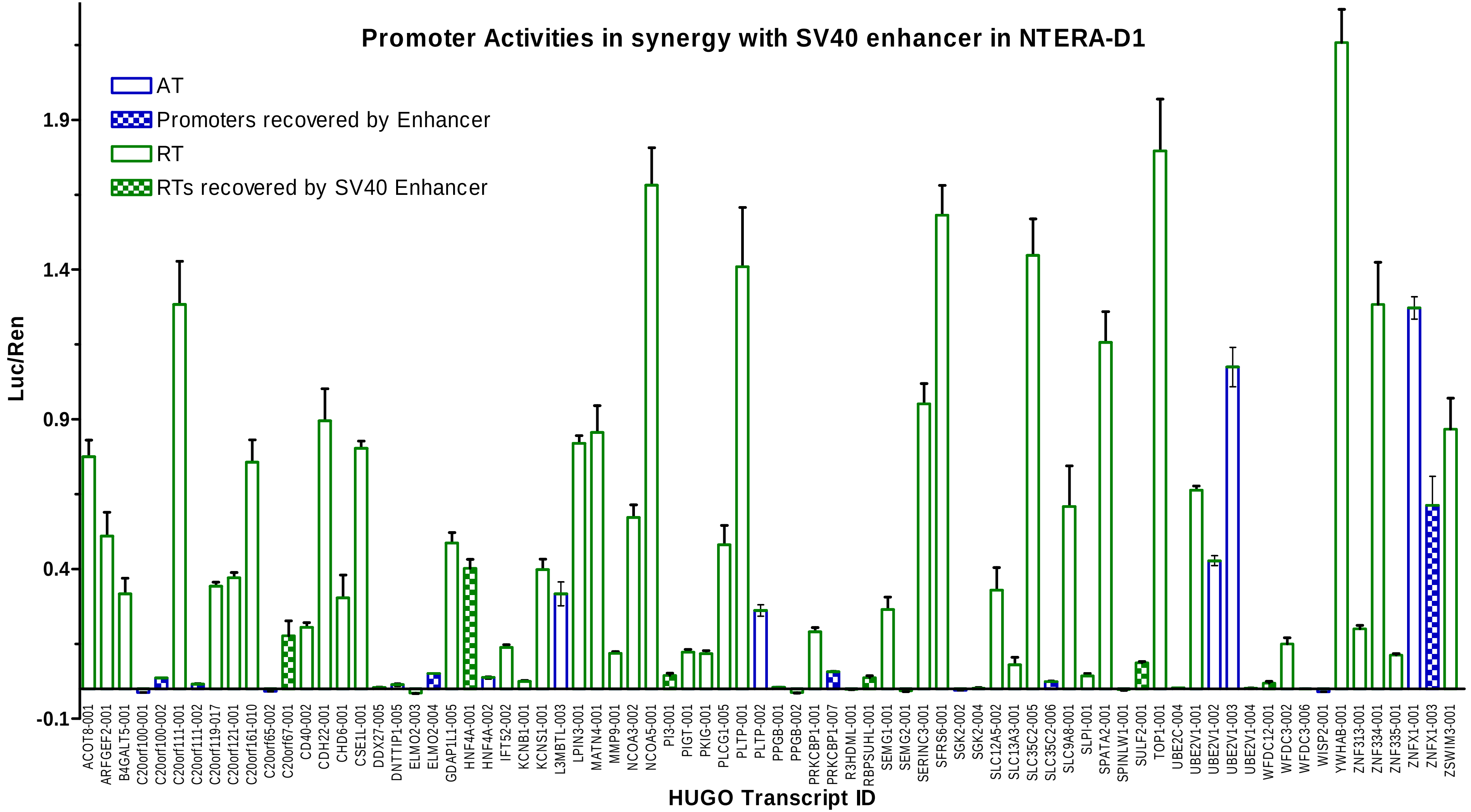

Figure 4.12 Background subtracted activity of 76 candidate promoters in synergy with SV40 enhancer in NTERA-D1 cells. The blue bars represents the promoter activities of alternative transcripts and the bars with checked pattern are the fragments which showed activity only in synergy with the enhancer.

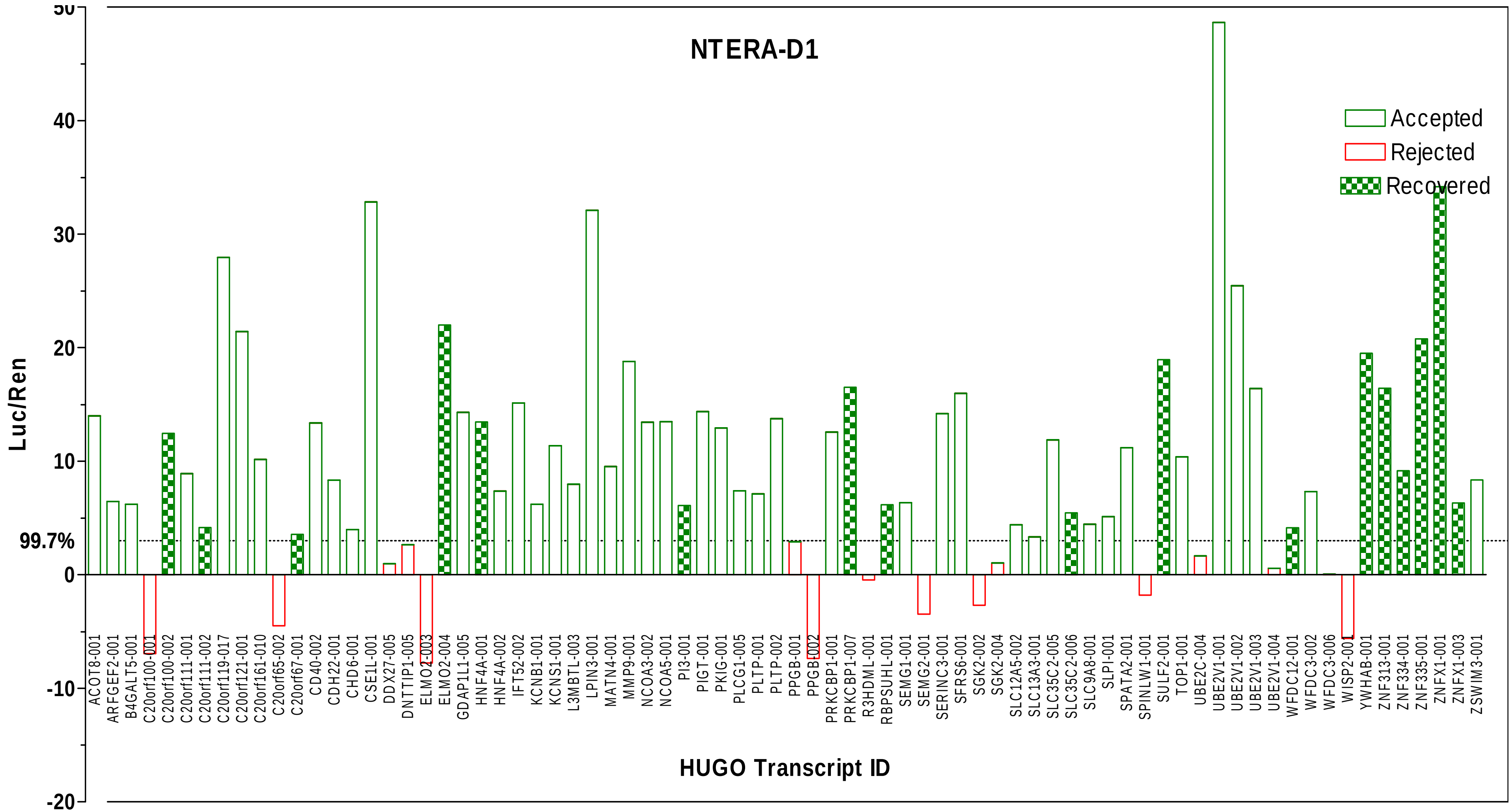

Figure 4.13 Confidence intervals to accept (green bars) or reject (red bars) the activation response in NTERA-D1 cells. Here, fragments which shows 3*σ higher than the background are accepted as activated by the enhancer and the confidence level for this decision is 99.7% assuming the error distribution follows Gaussian. Bars with squared patterns denote the promoters that gave activity only in the presence of the enhancer.

The response status of the candidate promoters to SV40 enhancer is summarized in Table 4.4 where 44 constructs (37 of them belong to representative transcripts) showed in both cell lines activity in synergy with the enhancer and 14 (5 of them belong to representative transcripts) did not. Also, 75% of constructs activated by the enhancer in both cell lines are associated with a promoter (FirstExon) and/or TSS (Eponine) prediction whereas only 14% of the inactive constructs are associated with a promoter and/or TSS prediction. This may mean that prediction programs perform better on sequences carrying higher number of transcription factor binding motifs since the activation by the enhancer mainly depends on this feature.

Construct Summary	Number of constructs activated by enhancer in NTERA-D1	Number of constructs that did not show activity with enhancer in NTERA-D1	Total Number of Constructs
Number of constructs activated by enhancer HeLa S3	**44**	**2**	**46**
Number of constructs that did not show activity with enhancer in HeLa S3	**16**	**14**	**30**
Total Number of Constructs	**60**	**16**	**76**

Table 4.4 Summary of the results obtained with SV40 Enhancer in HeLa S3 and NTERA-D1 cell lines

4.6.3 Sp1 binding sites on 71 promoters

There are 71 candidate promoter fragments, whose activities were assessed both in the absence and presence of SV40 enhancer in each cell line. These constructs are grouped as non-responsive, the ones that did not show any activity in synergy with the enhancer, and responsive. SV40 enhancer contains binding sites for Sp1, POU2F1 and NFKB1 transcription factors that interact with Sp1 transcription factor to mediate promoter activation. So, I searched for Sp1 binding sites in these 71 sequences using the program MAPPER (Marinescu et al., 2005) and attempted to correlate the Sp1 binding site profiles to their activation levels. Table A7 (Appendix A) lists the Sp1 binding site coordinates in the 71 constructs and each construct's response to SV40 enhancer. Out of 71, 13 constructs did not have any Sp1 binding sites.

Since Sp1 acts as an activator in a cooperative manner, a higher number of Sp1 bound to a promoter means a better chance for activation (Anderson and Freytag, 1991). Therefore, I generated the number of putative Sp1 binding sites on each of the 71 candidate promoter fragments activated or not by SV40 in both cell lines to generate frequency distribution plots for the number of putative binding sites and fitted curves for the frequency distribution plots. Figure 4.14 shows that promoters which are not activated in synergy with the enhancer, have a lower number of Sp1 binding sites.

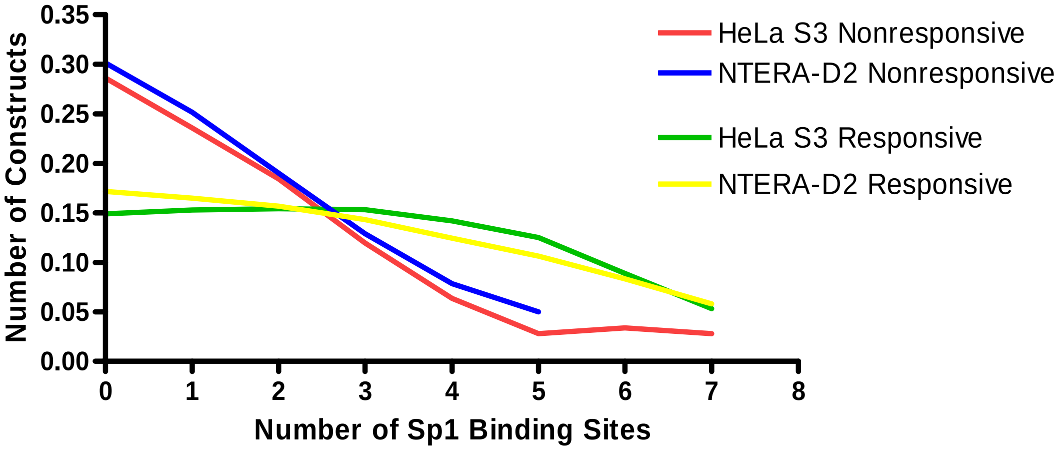

Figure 4.14. Frequency distribution of putative Sp1 binding sites on promoters responding or not responding to SV40 Enhancer.

This is not to say that higher number of putative Sp1 binding sites means activation since the promoters which were activated by the SV40 enhancer have variable number of putative Sp1 binding sites. Interestingly, only a few promoters with a high number of putative Sp1 binding sites (>4) are not activated by the enhancer, which suggests that collective Sp1 binding may indeed increase a promoter's chance for activation by the SV40 enhancer. This is in agreement with Sp1 playing a pivotal role in recruiting several activator proteins due to its diverse interactions (see Table 4.5).

4.6.4 Promoters active only in the presence of SV40 Enhancer

Out of 71 candidate promoters cloned to both vectors (with and without SV40 enhancer), 20 (16 representative and 4 alternative transcripts) showed activity only in the presence of the enhancer in HeLa S3 cells. Of these 20, half belong to genes that were expressed in HeLa S3 according to the Affymetrix Expression Array. In NTERA-D1 cells, 12 candidate promoters (6 representative and 6 alternative transcripts) were active only in the presence of the enhancer of which 9 were expressed in this cell line.

The activities of the promoters recovered by the action of the enhancer are shown in Figure 4.15. in a scale between -100 to 100 in order to compare different data sets. The candidate promoters that did not show activity above background were divided by the lowest activity within that dataset and multiplied by 100 (so that the lowest activity will be -100) and the candidate promoters that showed activity above background were divided by the highest activity within that dataset and multiplied by 100 (so that the highest activity will be 100). Five candidate promoters (denoted by a star in Figure 4.15) whose activities recovered in both cell lines.

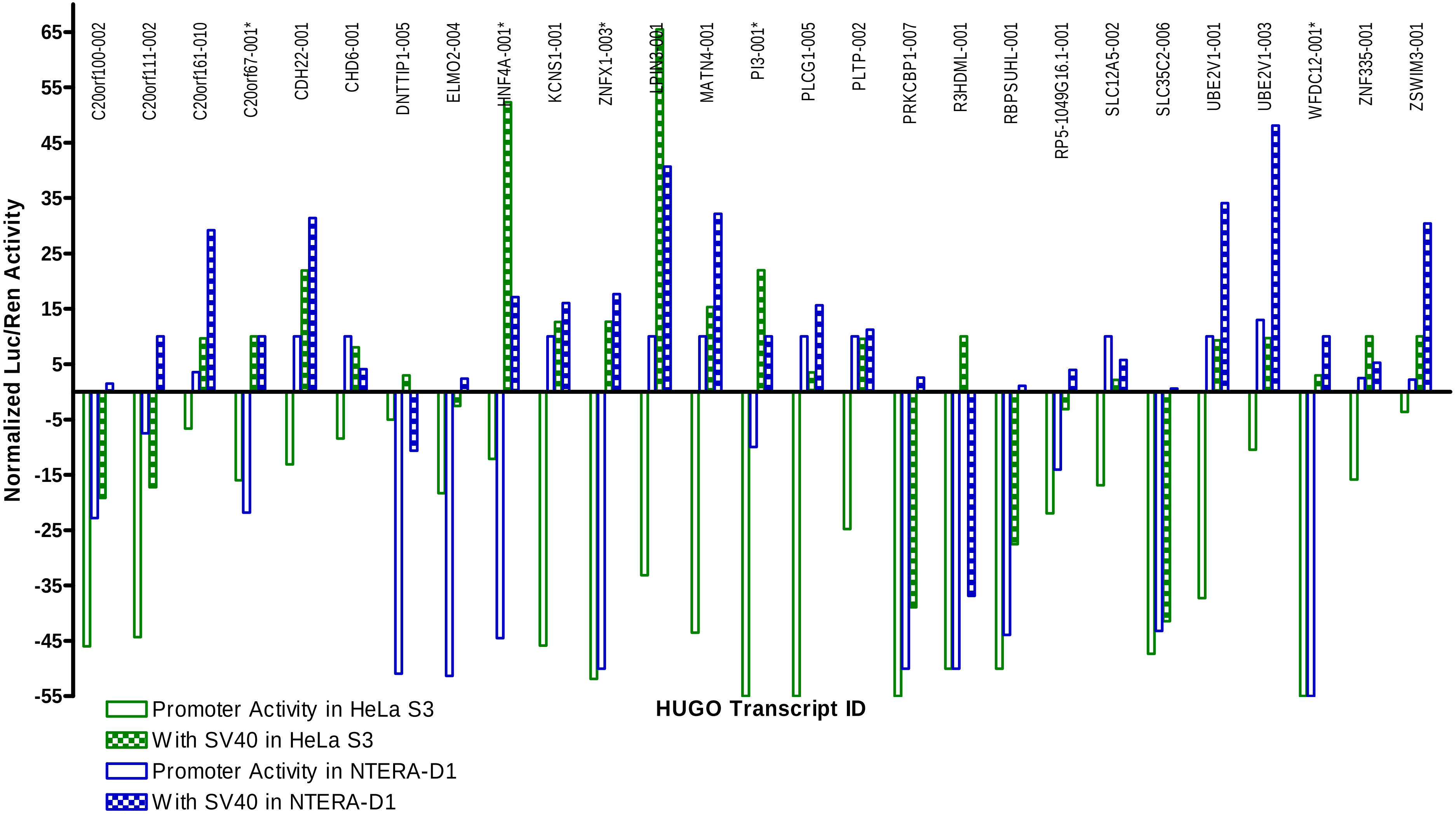

Figure 4.15 Scaled activities of promoters active only in synergy with SV40 enhancer in HeLa S3 and/or NTERA-D1 cells. Scaling was performed by dividing the activity of each promoter to the highest activity within all constructs. Constructs with (*) are recovered by the enhancer in both cell lines.

For a promoter to be responsive to an enhancer, it should contain binding site(s) for the proteins interacting with factors that bind directly or indirectly to the enhancer. The transcription factors that bind to SV40 enhancer are listed in Table 4.5. The 27 candidate promoters active only in the presence of SV40 enhancer were searched with MAPPER for binding sites of CREB, YY1, Sp1,TBP, NFKB1, AP1 and MYC transcription factors for an attempt to detect any correlation between the presence of the binding sites of these factors and the recovery ability of the promoter by the enhancer.

SV40 Enhancer Binding Factors	Interacting Factors
p300	CREB, YY1
Sp1	Sp1, YY1, TBP, NFKB1
AP-1	AP-1, AP-2
TEF-1	TFIID
POU2F1 (oct-1)	Sp1, GR-alpha, TBP
AP-2	AP-2, AP-4
NFKB1	Sp1
MAX	AP-1

Table 4.5. The left column lists transcription factors that have binding sites on SV40 enhancer and right column lists its interaction partners.

Also, the binding site profiles of the promoters not activated by the enhancer were generated and compared. Only putative sites with an 85% confidence level were included in the analysis to avoid false positive hits.

The transcription binding site profiles of the candidate promoters which gave activity only in the presence of the SV40 enhancer were given in Table A8 (HeLa S3) and Table A9 (NTERA-D1) (Appendix A). Also, Table A10 (Appendix A) lists the transcription binding site profiles of the candidate promoters that did not give activity in any cell line. No clear correlation between the binding sites and the ability of the enhancer to recover activity can be derived. Note that presence of a putative binding site in a promoter does not mean that the promoter is under the effect of this protein since the position of the binding site relative to TSS is also crucial. Therefore, rather than looking only for the presence of a putative binding site, I also looked at its position relative to the TSS. Accordingly, out of 27 candidate promoters recovered by the enhancer in at least one cell line, 7 carry a putative CREB protein binding site around 250 bp upstream of the TSS. None of the 12 inactive putative promoter fragments in both cell lines had such binding site in that position. The promoter of C20orf100-001 transcript, which did not give any activity with or without enhancer in any cell line, also contains a putative CREB binding site around 70 bp upstream of the TSS.

YY1 (Ying Yang 1) is a zinc finger protein that can activate, repress or initiate transcription in different gene contexts (Shrivastava and Calame, 1994). It initiates transcription when present around the TSS (Shi et al., 1997). In other positions, it can behave as an activator or a repressor according to the promoter context. Putative binding sites on 71 candidate promoters were plotted with respect to their position relative to TSS in Figure 4.16. There are two peaks at 250 downstream of TSS, and around TSS where YY1 may play a role in transcription initiation.

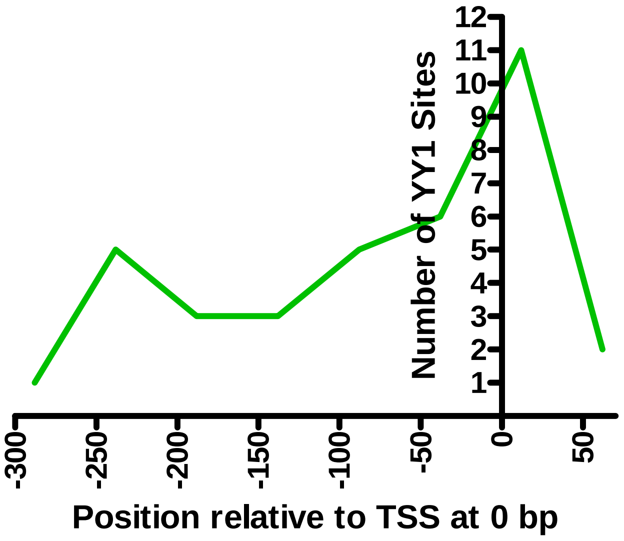

Figure 4.16 Number of putative YY1 binding sites plotted against their position relative to the TSS at 0 bp.

YY1 interacts with Sp1 and p300 that have binding sites on SV40 enhancer (Seto et al., 1993) (Austen et al., 1997). Six promoters recovered by the SV40 enhancer which contain putative YY1 binding sites between 90 to 250 upstream of the TSS, whereas none of the inactive promoters contain YY1 putative binding sites in this region. Since YY1 has an ability to produce sharp bends on the sequence it binds, such positional constraints for it to exert its activation or repression function on the promoter is expected (Kim and Shapiro, 1996).

4.6.5 Promoters Repressed by SV40 Enhancer

Four candidate promoter fragments were repressed by SV40 enhancer in HeLa S3 cells. The promoter of ELMO2-003 was also repressed in NTERA-D1 cells. The sequence of ELMO2-003 was searched with MAPPER (Marinescu et al., 2005) for putative transcription binding sites that may be responsible for the observed silencing effect. Interestingly, there is a ZFP161 (Zinc finger protein 161) binding site located at around 36 bp upstream of the TSS. ZFP161 contains a POZ (Poxvirus and zinc finger) domain found at the amino termini of several zinc finger transcription factors (Bardwell and Treisman, 1994). This domain has been shown to mediate protein oligomerization which subsequently prevents high-affinity DNA binding (Numoto et al., 1999). It can form dimers but also interacts with non-POZ domain containing proteins (Kaplan and Calame, 1997). It acts as a repressor on MYC and β-actin promoters, but also activates the HIV-1 long terminal repeats (LTR). ELMO2-003 has 6 binding sites for Sp1 (one overlapping with ZPF161 binding site) but no binding sites for the activators interacting with SV40 enhancer binding proteins. It is shown that Sp1 and ZPF161 can act together to repress transcription (Kaplan and Calame, 1997). There are 2 other transcripts (SLC12A5-002, SNX21-010) containing ZPF161 binding sites downstream of TSSs (46 and 50 bp respectively), their transcription factor binding motif profile for Sp1, YY1 and NFKB1 activator proteins are given in

Table 4.6. Unlike ELMO2-003, promoters of above two transcripts contain YY1 and NFKB1 binding sites that can act as activators by interacting with SV40 binding factors p300 and Sp1, overcoming the repression effect of ZPF161.

Interestingly a putative binding site cyclic AMP element responsive binding (CREB) protein binding site was found around 140 bp upstream of the TSS site of the three candidate promoters repressed in HeLa cells (active in NTERA-D1 cells) by SV40 enhancer. CREB protein interacts with p300 that has a binding site on SV40 enhancer. It is known that CREB protein mediate repression by forming homodimers, or heterodimers with potential activators to prevent their activation function (Costa and Medcalf, 1996) (Van et al., 2001). In this case, CREB protein binding stabilized by SV40 enhancer on these candidate promoters might mediate recruitment of certain repressor proteins depending on the transcription factor binding site motif profiles of these promoters. Note that there are other candidate promoters carrying a CREB protein binding site around the same location which means that this repression effect is most likely dependant on the nature of the promoter sequences.

Table 4.8 lists some of the promoters whose activities are greatly enhanced by SV40 enhancer. As can be seen from the table, except HNF4A-001, remaining promoters have putative binding sites for at least one activator protein interacting with factors that have binding site(s) on SV40 enhancer.

	HeLa S3			NTERA-D1			Binding Site Coordinates relative to TSS			
HUGO Transcript ID	**Promoter Activity**	**Enhancer Response Activity**	**Response to Enhancer**	**Promoter Activity**	**Enhancer Response Activity**	**Response to Enhancer**	**Sp1**	**NFKB1**	**YY1**	**ZPF161**
ELMO2-003	0.172	-0.708	repressed	0.143	-1.000	repressed	-226;-182R;-84;-63;35;	0	0	36
SLC12A5-002	-0.169	0.022	recovered	0.003	0.058	active	-202R;-93;-44;33;45;	0	-241	46
SNX21-010	-0.066	0.097	recovered	0.036	0.292	active	-223;-75;-63;29;49;58R	-160	0	50

Table 4.6 Transcription Binding Sites of three constructs carrying ZPF161 binding site downstream of TSS. Binding site coordinates are given relative to TSS and "R" denotes for the binding motif on the opposite strand. Note that ELMO2-003, which was repressed under the effect of SV40 enhancer, does not contain activators such as YY1 or NFKB1.

	HeLa S3		NTERA-D1		Transcription Factor Binding Site Coordinates relative to TSS				
HUGO Transcript ID	Promoter Activity	Response to Enhancer	Promoter Activity	Response to Enhancer	**Sp1**	**CREB**	**YY1**	**MYC**	**ZF161**
ARFGEF2-001	6.595	-18.510	53.784	14.938	-240R;-116;-38R;-22R;	-167R;	-156;-57R;	-226R;-225;-181;-174R;	
ZNF313-001	1.641	-51.291	39.882	8.968	-107;-87R;-70R;-56R;47;	-125;	0;	80R;81;	
IFT52-002	3.143	-13.918	7.422	6.079	-123R;	-115R;34;			
ELMO2-003	0.172	-0.708	0.143	-1.000	-226;-182R;-84;-63;35;				36

Table 4.7 Transcription Binding Site profile of the constructs that are repressed in HeLa S3 cells. Binding site coordinates are given relative to TSS and "R" denotes for the binding motif on the opposite strand.

HUGO Transcript ID	**Promoter Activity**	**Response to Enhancer**	**Promoter Activity**	**Response to Enhancer**	**Transcription Factors and their binding sites**
LPIN3-001	inactive	0.65	inactive	0.60	1 AP1 binding site at +51 bp 6 Sp1 binding sites NFKB1 binding site at -125 bp
HNFA-001	inactive	0.53	inactive	0.17	No sites found
SFRS6-001	0.30	0.64	0.43	0.70	YY1 binding sites at -146R bp 66R bp MYC binding site at +5 bp
TOP1-001	0.05	0.54	0.10	0.70	CREB binding site at -93R bp
ZNFX1-001	0.97	0.99	0.04	0.64	6 Sp1 binding sites CREB binding site at -188 bp
PLTP-001	0.10	0.79	0.01	0.45	TATA-box at -41 bp NFKB1 binding site at -196R bp MYC binding site at -162 bp 3 Sp1 binding sites
MATN4-001	inactive	0.15	0.01	0.32	YY1 binding site at -91 bp

Table 4.8 Transcription factor binding site profile of promoters whose activities are greatly enhanced in synergy with SV40 enhancer. The response levels to enhancer is normalized to map onto 0 to 1 range. Binding sites were given relative to the TSS at +1 and “R” denotes the opposite strand.

4.7 Summary

This study investigated 74 putative promoter fragments in HeLa S3 and NTERA-D1 cell lines. Of these, 30 and 47 showed activity in HeLa S3 and NTERA-D1 cells respectively. The activity of promoter fragments is shown in Figure 4.17.

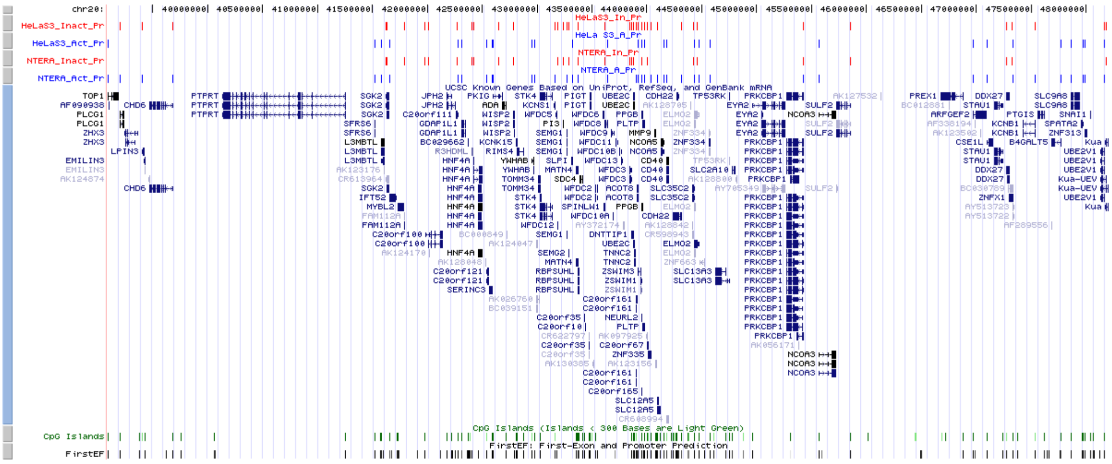

Figure 4.17 Promoter activities of 74 putative promoter fragments in HeLa S3 and NTERA-D1 cell lines. Inactive promoters are shown in red and active promoters were shown in blue. The annotation is taken from UCSC Genome Browser.

Then, 71 of these 74 putative fragments were investigated under the effect of SV40 enhancer in both cell lines. Of the 71 fragments, 46 and 60 fragments showed activity in synergy with the enhancer, and the activities of 16 and 13 putative fragments were recovered using SV40 enhancer in HeLa S3 and NTERA-D1 cells respectively. The activity of putative promoter fragments in synergy with the enhancer is shown in Figure 4.18.

Of the 103 representative transcripts, core promoter activities of 50 were investigated in two cell lines and 35 (70%) showed activity in at least one cell line. When the 50 core promoters were analysed in synergy with SV40 enhancer, 46 (92%) showed activity in at least one cell line. Therefore, the promoter activity of 92% of the investigated putative fragments were verified using two cell lines. In another study, gene reporter assays were employed to verify the activities of 921 putative promoter fragments in Ensembl region using 16 different cell lines (Cooper et al., 2006).

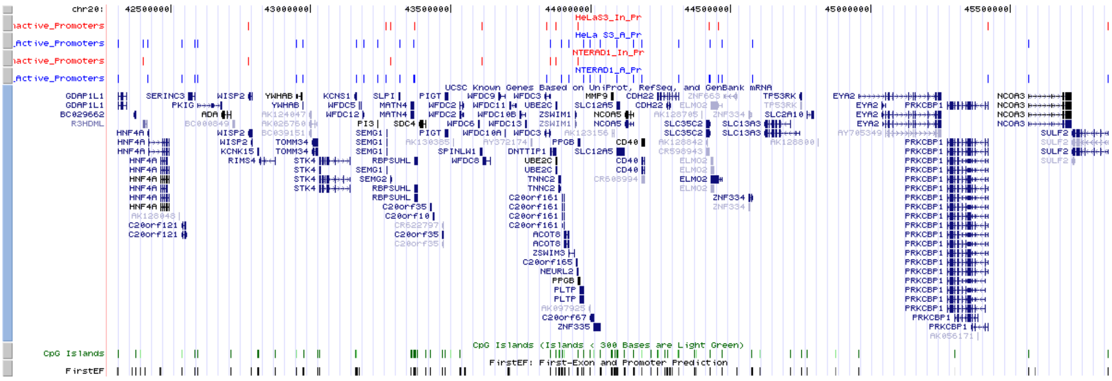

Figure 4.18. Promoter activities of 71 putative promoter fragments in synergy with SV40 enhancer in HeLa S3 and NTERA-D1 cell lines. Inactive promoters are shown in red and active promoters were shown in blue. The annotation is taken from UCSC Genome Browser.

Of these, 387 (42%) fragments showed activity in at least one cell line. Based on the high promoter activity recovery rate obtained in this study with the help of SV40 enhancer in two cell lines, such an approach appears to be more cost-effective to investigate the promoter activity of a putative fragment. If a given genomic fragment has a tissue-specific promoter activity, it may still give activity in a cell line in which it is inactive with the help of a strong enhancer since most enhancers helps recruiting ubiquitously expressed activators onto core promoter regions. Promoters that did not show activity in any configuration (with or without SV40 enhancer) can be further tested by using a fragment longer than 300 bp in a reporter assay. If such fragments are indeed promoters, they might still need their specific enhancer elements located in proximal promoter regions. However, in this case, it is possible that the promoter has a tissue specific expression, therefore more cell lines should be employed as well.

In this study, candidate promoter fragments were selected based on the previous annotation of the region and since 92% showed activity in at least one configuration, it conforms the accurate annotation status of transcripts in this region.