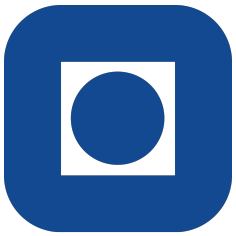

NTNU

Norwegian University of Science and Technology

Faculty of Medicine and Health Sciences

Department of Clinical and Molecular Medicine

# Bioinformatic analysis of regulatory regions in immediate-early response genes

*Author:*
Magnar Blindheim

Thesis written: Autumn 2019
Finished and delivered: January 15, 2020

**Abstract**

This thesis has looked at genetic and epigenetic regions which contribute in the overall regulation of immediate-early genes (IEGs). IEGs are genes that transcribe mRNA quickly after being stimulated, and are thus involved in a multitude of cellular processes. Some IEGs are classified as among others proto-onco genes, housekeeping genes, cellular repair genes and many more. The aim for the study has been to identify and quantify the strength for the enrichment of some regulatory regions in a set of IEGs compared to non-IEGs. Some transcription factor binding sites and some histone markers located in the vicinity of the transcription start site of the IEGs has been shown to be upregulated. Most of the regulatory regions that seem to be enriched have a function related to facilitating rapid transcription, which is expected in genes that in general respond quickly after being stimulated. The results support the current understanding that IEGs are primed for transcription by having more binding sites for transcription factors, such as the subunits of the cohesin complex. In addition, IEGs seem to be primed by having a favourable methylation and acetylation state of the histone complex, such as the enrichment of H3K4me3. IEGs seem to be under stricter regulation than other genes, which is indicated through the enrichment of binding sites for other transcription factors such as CTCF. The study has failed to provide any genetic markers that are depleted in IEGs, and finding depleted regions would further the current understanding of how IEGs are regulated.

Norwegian abstract:

Denne oppgaven har sett på genetisk og epigenetiske regioner i DNAet som bidrar til å regulere "immidiate-early gener" (IEG). IEG er gener karakterisert ved at de transkriberer mRNA raskt etter stimuli, og er involvert i en rekke cellulære prosesser. Ulike IEG fungerer blant annet som proto-onkogener, "housekeeping" gener eller cellulære reparasjonsgener. Målet for denne studien har vært å identifisere hvorvidt regulatoriske regioner i IEG er anriket sammenlignet med andre gener. Enkelte bindingsseter for transkripsjonsfaktorer, og noen histonmarkører som befinner seg i området rundt IEG var i denne studien oppregulert. De fleste av de oppregulerte regulatoriske regionene som har vært testet, har funksjoner som er knyttet til transkripsjonsregulering. Dette er forventet i gener som generelt sett responderer raskt etter stimuli. Resultatene støtter dagens bilde av at IEG er klargjort for rask transkripsjon ved blant annet å ha flere bindingsseter i promotorregioner hos IEG for transkripsjonsfaktorer, slik som subenhetene i cohesinkomplekset. IEG har også vist seg å være klargjort ved å ha et fordelaktig metylering- og acetyleringsmønster i histonkompleksene, eksempelvis oppreguleringen av H3K4me3. Det kommer også fram i studien at IEG kan være strengere regulert enn andre ikke-IEG. Dette sees gjennom oppreguleringen av en rekke transkripsjonsfaktor-bindingsseter slik som CTCF. Studien har ikke tilstrekkelig undersøkt genetiske regioner som har vært nedregulert, hvilket kan belyse videre aspekter rundt IEG reguleringsmønster.

# Contents

# 1 Introduction

Immediate early response genes is an interesting class of genes which has a wide variety of functions. Some of them are householding genes meaning that they are essential for the normal function of the cell. Others, like some in the MYC family genes are regulator genes and proto-oncogenes, and are thus important in understanding how cancer cells develop. Common for all IEGs is that they are transcribed quickly and transiently within a 90 minute time frame after stimulation. Many IEGs are already discovered as they respond to a wide variety of stimuli and have been described in multiple articles. The list of known IEGs is continuously expanding. This study will attempt to identify key genomic features which are abundant in most IEGs. The purpose will be to identify more IEGs in a future study, which may not necessarily respond to such a wide variety of stimuli as the current IEG set does. The data may also contribute to and consolidate the current understanding of genomic features of IEGs. Due to these genes' ability to reach a transient peak just ten minutes after stimulation, they are already widely used as a marker for early detection of neuronal plasticity[9][22]. The regulation of IEGs has also shown to be one of the primary mechanisms for regulation of the concentration of specific gene products[1].

All genes transcribed by RNA polymerases have a promoter region which is approximately 100-1000 base pairs long. This region consists of a binding site for the polymerase and participates in modulating the expression of the particular gene. The gene expression modulation can happen through a multitude of mechanisms, such as CpG islands - of which are found in approximately 70 percent of all promoter regions. CpG islands have a high content of cytoxine-guanine base pairs, making the two strings of the DNA helix loosely bound to each other. The effect of this structure is that the region will easily unwind, thereby facilitating transcription. Research has already shown that IEGs are enriched in CpG islands - meaning that IEGs have a large amount of CpG islands compared to the rest of the genome. This is an example of one way the rapid transcriptional response of IEGs can be explained by their genetic structure. This study will delve into other structural qualities with indicated relevance to the IEG response.

Epigenetic markers such as histone methylation and acetylation affects gene expression. This study illustrates epigenetic IEG regulation by testing some known activating markers.

IEGs have multiple interesting qualities making them able to respond quickly to different stimuli. For example transcriptional pausing, where RNA polymerase II (pol II) can stop transcribing until its signaled to continue. This is one of the suggested mechanisms for how IEGs can be transcribed without de novo protein synthesis, and can therefore partly explain how these genes respond so quickly. Most of these genes are also

represented in a wide variety of different species, including snails and mice, indicating their importance for survival[26]. In spite of this, many of them are poorly understood, and their role in a multitude of biological processes are largely unknown[10].

This study will primarily focus on mapping different genetic and epigenetic features which already seem to be of importance for the regulation of IEGs. Mostly the focus will be on transcription factors which are involved in some sort of genetic regulation in addition to chromatin status in the vicinity of IEGs.

# 2 Theory

## 2.1 Primary gene response

The primary response genes (PRGs) are the first responders to both cell intrinsic and cell extrinsic stimuli. They respond quickly, transiently and do not require de novo protein synthesis. This is a large group of genes consisting of IEGs and delayed primary response genes. Common for PRGs and thus also for IEGs, is their poised state. This means that they are heavily regulated, and primed for transcription through mechanisms described throughout this article. Genes being poised means that the chromatin structure is bivalent with simultaneous expression of repressive as well as activating markers, a key concept that will be examined further. PRGs play a key role in a multitude of different cellular processes, some of them being cell metabolism, neuronal plasticity and cellular maintenance[7].

### 2.1.1 Immediate early response genes

Within minutes after initial stimuli the transcription of IEGs is observed, despite the presence of protein synthesis inhibitors. The IEG response is seemingly the same in multiple different cell types, and a couple hundred different IEGs are assumed to exist in the human genome. Many IEGs are linked to the cell cycle and many are proto-oncogenes, of which the first discovered IEG, c-fos, is an example. Some of them are also transcription factors, indicating that they play a role in mediating the cellular response to different types of stress. An example of this is the immediate-early protein epidermal growth factor 1 (EGR1). This stress response is also tested through the expression of heat shock factor 1 (HSF1) binding sites in the promoter region. HSF1 is known as the primary mediator of transcriptional response to proteotoxic stress, and is also important in regulation of regular cellular development and metabolism[22].

### 2.1.2 Delayed Primary response genes

The PRGs are by definition the "first responders" to a multitude of stimuli. These genes differ from the secondary response genes not by how fast they respond, but whether or not they require de novo protein synthesis before they are transcribed. Some genes have shown to be transcribed without de novo protein synthesis, although significantly slower than the current list of IEGs. These are called delayed primary response genes and differ from IEGs in both genomic architecture and function, and will not be further addressed in this article[1].

## 2.2 General properties of IEGs

Currently there are more than 100 identifyed IEGs, and as previously mentioned they respond quickly and transiently to a wide variety of stimuli. For most IEGs peak mRNA-levels are detectable within 30-60 minutes after stimulation[22]. The activation of IEGs is described in interphasic cells where they are activated by extracellular signals such as growth factors (Platelet-Derived Growth Factor and Epidermal Growth factor), mitogens, phorbol esters, immunological and neurological and developmental stress. IEG products are usually degraded through proteolysis in the proteasome without prior ubiquitination. There has also been shown that IEG transcripts could be downregulated through the mechanism of targeted microRNA[18].

IEGs are on average of a shorter length than other genes (19 kb versus 58 kb), and they have significantly fewer exons. They have a high prevalence of TATA boxes and CpG islands. There is a known enrichment for some specific transcription factor binding sites within regulatory regions of IEGs, including serum-response factor (SRF), nuclear factor kappa B and cyclic AMP response element-binding protein binding sites. This has been suggested as a consistent and perhaps even redundant mechanism of transcription regulation[11].

## 2.3 Gene structure and regulation

The exons which are the protein coding part of the human genome are transcribed by pol II into a messenger RNA (mRNA) string, which is further translated by ribosomes in to proteins. All these steps are candidates for heavy regulation in many different ways. This article focuses specifically on regulation through transcription factor binding sites in the promoter region, transcriptional regulation and initiation, long range enhancer or repressor modulation and histone modification.

## 2.4 Promoter region

It has already been shown that the promoter regions of IEGs are enriched compared to other genes in TATA-boxes-binding protein (TBP) binding sites and there is also an enrichment of CpG-islands compared to non-IEGs. This has been linked to the fast activation of IEGs because of how these regions facilitate the fast unwinding of the DNA helix. This further facilitates the initiation of transcription. The TATA box helps the unwinding process by consisting of AT-rich sequences which break apart easily and CpG islands, by assmbling the DNA into unstable nucleosomes. One of the consequences of CpG-islands making the nucleosomes unstable is that the transcription becomes mostly independent of chromatin remodeling complexes (like SWI/SNF). SWI/SNF-independent genes are in general induced more quickly than SWI/SNF-dependent genes[19]. Other qualities of the promoter region contribute to the overall picture of how IEGs are poised for transcription. An example of such qualities is the amount of binding sites for activating transcriptional factors such as activating transcriptional factor B, 2 and 7 (BATF, ATF2 and ATF7).

## 2.5 Transcription initiation

Eukaryotic cells contain 3 nuclear RNA polymerases, with pol II being responsible for transcribing all mRNAs and numerous non-coding RNAs. Pol II does not recognize promoter DNA by itself, but rather as a part of the basal Pol II machinery that includes general transcription factors. Some transcription factors dissociate from Pol II when elongation starts. The C-terminal domain of Pol II contains multiple copies of a heptad repeat which is phosphorylated at serines 2 and 5. Serine 5 is phosphorylated by TFIIH at the promotor and serine 2 is phosphorylated by P-TEFb/CTK1 during elongation. C-terminal domain phosphorylation is important in coupling Pol II elongation to post-transcriptional steps such as mRNA capping, splicing, polyadenylation, export and chromatin modifications[14][24].

## 2.6 Distant enhancers and repressors

Distant enhancers and repressors are short strands of the genome which are located up to 1 million base pairs away from the gene in which they influence. Through the 3D structure of the DNA helix, enhancer and repressor regions can interfere with the transcription of other genes. Structures like the chromatin, which contribute to the overall 3D arrangement of DNA, will play a part in the effect from distant enhancers and repressors. CCCTC-binding factor (CTCF) is a transcription factor that is known to interfere with distant enhancer and repressors and their abundance in promoter

regions of IEGs is consequently of particular interest, and is more closely looked at under the subsection "CTCF".

## 2.7 General transcription factors

Binding sites of transcription factors (TFs) are abundant in the promoter regions of IEGs. As mentioned previously IEGs do not require de novo protein synthesis, which means that these factors have to be premade before transcription activation. The known abundance of binding sites of TFs in the promoter regions indicate that IEGs are well regulated in spite of their rapid reaction. Some TFs like TBP as described above are already reported to be found more frequently than others in the upstream regions of IEGs.

### 2.7.1 Cohesin complex

The cohesin complex regulate the separation of sister chromatids during cell division. The cohesin complex has also been mentioned in association with long range interaction between promoters. The complex consists of four different subunits, where two of the subunits, named RAD21 and SMC3 are analyzed in this study. The two other subunits are named SMC1 and SCC3, and all four are found with less than 0.5 percent amino acid divergence in prokaryotes as well as eukaryotes. This complex is interesting in regard to IEGs because of the already observed state of their promoter region. By already having a promoter region that is facilitating easy unwinding, an up-regulation of binding sites for RAD21 and SMC3 in promoter regions of IEGs would contribute to the easy unwinding of the DNA helix. An abundancy of CpG islands and TATA-boxes combined with enrichment of RAD21 and SMC3 binding sites could be interesting marks in the search for additional cell specific IEGs, or IEGs that responds to a narrower spectre of stimuli[13].

### 2.7.2 CTCF

CTCF is a zink-finger protein which is involved in multiple cellular processes like blocking distant enhancers, transcriptional regulation and regulation of chromatin structure. Cohesin and CTCF is linked through CTCF-CTCF/Cohesin loops and are working together to regulate long-range interactions. Acute depletion of CTCF has been shown to directly effect MYC regulation through loss of enhancer-promotor looping, indicating that they may play a part in the general regulation of IEGs[12][13]. The CTCF binding sites has been tested in multiple laboratories and are thus tested replicates.

These are marked CTCF1 and CTCF2 and their references are included in the last portion of the article.

### 2.7.3 Interesting proteins

DRB sensitivity inducing factor (DSIF) is a transcription factor which regulates the binding of pol II. It is known to interact with NELF to promote stalling of pol II during transcription of some genes. The stalling can be relieved by positive transcription elongation factor b (P-TEFb), which was mentioned earlier. Pol II stalling is a process described in further detail under the subsection "paused transcription".

EGR1 is a zink-finger protein which seems to have an important role in neuronal plasticity and functions as a transcriptional regulator. This protein functions as a TF and is fascinatingly also a well known immediate-early protein. This protein is interesting because of how it interacts with other early response genes. By being an immediate-early protein it responds quickly to a wide variety of stimuli before binding to the promoter regions of other late-response genes and further mediates their response[10].

Serum response factor (SRF) is a known gene regulator. It is a downstream target of many signaling pathways. It is important in the development of the embryonic mesoderm and essential in the formation of skeletal muscle.

Negative elongation factor (NELF) negatively impacts transcription by pausing pol II 20-60 nucleotides downstream from the transcription start site (TSS). NELF binds DSIF to pol II and is inhibited by P-TEFb. RDPB is a part of the NELF-complex. The mechanism of promoter proximal pausing is further explained under the section "Paused transcription"[8][17].

## 2.8 Chromatin structure

The chromatin structure in IEGs seems to facilitate the initiation of transcription. A genome-wide mapping of repressed intergenic and intragenic TSSs enriched with active chromatin marks and pol II has already shown strong association with IEGs[20]. Regions like these often have both activating and repressive histone modifications. Acetylation of Histone H3 Lysine 27 (H3K27me3) is for example an important repressive mark, whereas acetylation of Histone H3 Lysine 4 (H3K4ac) an important activating mark. If both are upregulated in a certain gene at the same time, the gene will therefore be in a poised state.

Histone acetylation has been shown to remain persistent before and after stimulation

which creates an open promoter structure [11][21]. The acetylation mark H3K4me3 is abundant across the promoter regions of IEGs. This is mostly found among the TSS of actively transcribed genes, and is often found with H3K36me3 in the coding region. Both of these marks signals that IEGs are transcribed actively[3][23].

As described in the article by Earnst and Kellis[6], different chromatin states are associated with different transcription states. As an example, H3K4me3 is associated with bivalent/poised state, active TSS and flanking TSS, whereas H3K27ac is associated with Active TSS and enhancer regions[2][25].

## 2.9 Paused transcription

Most IEGs are in an epigenetically poised state[1]. The fact that IEGs can be paused during elongation and resumed at a later stage is thought to be one of the reasons why IEGs do not need de novo protein synthesis before they are transcribed. If pol II does not have to assemble, and can restart transcription without classical transcription initiation, then protein transcription can initiate by signaling pol II to resume elongation. They may be reactivated through interaction with distant enhancers and it has been hypothesized that this interaction may happen through the production of eRNA from the sites of the enhancers[5].

As previously described, pol II can be stopped during elongation in a regulated process involving DSIF, P-TEFb and NELF among others. Most of transcription pausing happens proximal to where the transcription initiates, which is called promoter-proximal pausing. P-TEFb seems to be the main factor for resuming transcription. P-TEFb phosphorylates serine 2 of the C terminal domain of pol II as mentioned earlier, but other positive elongations factors include chromatin-modifying factors such as JIL-1 kinase, FACT, Paf1 complex and SPT6. Studies have shown that blocking P-TEFb will stop almost all transcription signaling, indicating that this is an essential process for transcriptional regulation[5].

## 2.10 DNA double strand breaks

Madabhushi et al. describe in their article published in Cell in 2015, that neuronal activity induces rapid IEG response in addition to double strand breaks in the promoters of some early response genes, including fos and EGR1. In this article they hypothezise that double strand breaks may contribute to resolve topological constraints to early-response gene expression[16]. Calderwood et al. showed one year later that topoisomerase IIb induced double strand breaks were necessary and sufficient for transcriptional activation of heat shock genes and serum-induced IEGs among others. They

also showed that this transcription was associated with initiation of DNA damage response pathways[4].

RAD51 is one of the proteins that play a major role in homologous recombination of DNA during double strand break repair, and is tested in this study. A recent experiment shows that the DNA repair-associated protein Gadd45 regulates the temporal coding of IEG expression within the prelimbic prefrontal cortex and is required for the consolidation of associative fear memory. This consolidates the current evidence for the importance of DNA double strand breaks in regulation of IEGs[15].

# 3 Method

## 3.1 IEG data set

A complete characterization of the IEG data set used in this study is described in the IER manuscript written by Bahrami and Drabløs[1]. A consensus for an IEG set has been developed through multiple experiments which have identified genes with some specific shared properties. The shared properties for all IEGs in the general consensus data set is the rapid response under 60 minutes. This also means that the current set for IEGs is quite general and responds to a wide variety of stimuli in multiple different cell types. As some of the IEGs are not tested with blocked protein synthesis, some of the IEGs may not fulfill the criteria of not requiring de novo protein synthesis. The properties which have been examined in this study were chosen through a process of reading source material and identifying different regions that would be expected to be enriched in IEGs.

## 3.2 Genome browser

University of California - Santa Cruz Institute of Genomics has developed a tool called UCSC Genome Browser which has been used to extract data. Data has primarily been sampled through the ENCODE project to be used in the enrichment analysis. The UCSC Table Browser has been utilized to extract .bed files for further analysis. The BED format used to format the data for analysis in this experiment consists of three required fields: in which chromosome the track is found, the start position and end position of the track. A BED file consisting of CpG islands in our genome will consist of a long list of start and stop positions of all CpG islands in the human genome. This data can be coupled with a similar list of known IER genes and compared to other non-IEGs to see if there are more or less CpG islands in vicinity of the IER genes than what would be expected by chance.

## 3.3 ENCODE-project

The Encyclopedia of DNA elements (ENCODE) consortium is an ongoing international collaboration of research groups with a common goal of building a comprehensive list of functional elements in the human genome. This includes elements that act on protein and RNA-levels. All of the data used in this project is sampled from the ENCODE 3 project. The position of all genes are defined by the position of the TSS for the given genes which marks position 0. The sampling for this study being 1800 nucleotides upstream and 200 nucleotides downstream refers to the relative position to the TSS for the given genes. Upstream refers to the direction from TSS away from the transcribed part whereas downstream refers to the direction of which the gene is transcribed.

The ENCODE 3 project annotates the data on three different levels. Ground, middle and top level, where ground level data are derived directly from the data such as TSS, chromatin marks and gene expressions. Top level is for example annotation of chromatin states. Only ground level data are being used in this study.

## 3.4 Cell lines

The cell lines used in this study stem from the UCSC Genome Browser, sampled through the ENCODE 3 project. The cell lines used in this enrichment analysis are GM12878, which is a well documented B-lymfocyte cell line, K562 and H1-hESC. The analysis was first done on the GM12878 cell line and the data was later validated through other cell lines. All data collected as a part of this study is included in the results.

There are different tiers of cell lines used in the ENCODE 3 project, and GM12878 belongs to tier 1. The tier classification is made to aid in choosing which cell lines to study. The two other cell lines that are classified as tier 1 are H1-hESC and K562. Most of the tissue specific cell lines are from tier 3. The other two cell lines used in this experiment are H1-hESC which are embryonic stem cells and K562 derived from leukemia cells during blast crisis.

## 3.5 Data analysis

The comparison between IEGs and the general gene set has been done using a local tool which utilizes the Fisher exact test. Fisher exact test is a method for analyzing whether or not there is a non-random association between two variables. The data is analyzed from a 2 by 2 contingency table and returns odds ratio with the 95 percent confidence interval along with the p-value. The odds ratio quantifies whether the attribute in

question is enriched or depleted by comparing different regions in the genome. In this analysis the overlap of different .bed-files has been tested. The 200 downstream nucleotides and the 1800 upstream nucleotides of the genes have been analyzed. 2000 base pairs were chosen because most regulatory TF binding sites and histone complexes are found within 2000 nucleotides from the TSS, and primarily upstream as the area downstream mostly consists of the transcribed gene. The Fisher exact test shows how many times a .bed file is checked for overlaps with both the IEG and the non-IEG data set. The odds ratio gives information on whether or not the quality overlaps more or less in IEGs.

The analysis has been done on UCSC reference genome hg19. This version of the genome was published in feburary 2009, is widely used and it is the 19th version of the reference genome.

# 4 Equipment

For this study, a computer operating with Linux operative system has been used. The computer has access to the UCSC Genome Browser and the local tool "stat over" which has been developed by Finn Drabløs for the enrichment analysis. The enrichment analysis is based on the Fisher exact test and utilizes .bed files consisting a known set of IER genes and a set of known protein coding genes.

# 5 Results

The results are as described previously presented with the 95 percentage confidence interval and the odds ratio for expression of multiple TF binding sites, and histone markers in the set of IEGs compared to a set of non-IEGs. Firstly, data from relevant histone markers are presented. Data from transcription factor binding sites are then presented. The cell lines are color coded, and the p-values for the different measurements are shown in the graphs. A line displays the odds ratio of 1, and indicates where there is no enrichment of the quality. The data from the CpG analysis is not included in the graphs, but showed a enrichment with an odds ratio of 2.00 (1.31 - 3.14) with a p-vaule of less than 0.001.

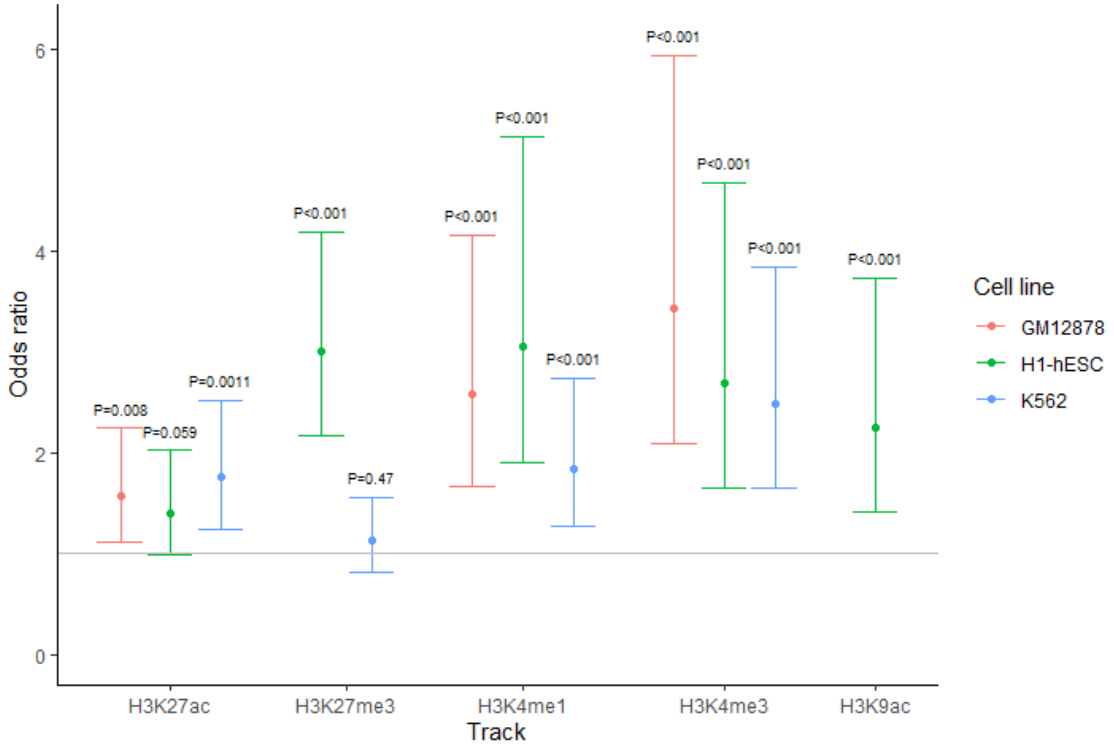


Figure 1:
This figure shows the histone markers tested in this experiment. The odds ratio is displayed on the y-axis, and the different histone markers on the x-axis. The color marks the cell line. A line for when odds ratio equals 1 is displayed as well. The p-value for the specific analysis is displayed until the p-value is less than 0.001. The accurate p-value is displayed in table 1 later in this section.

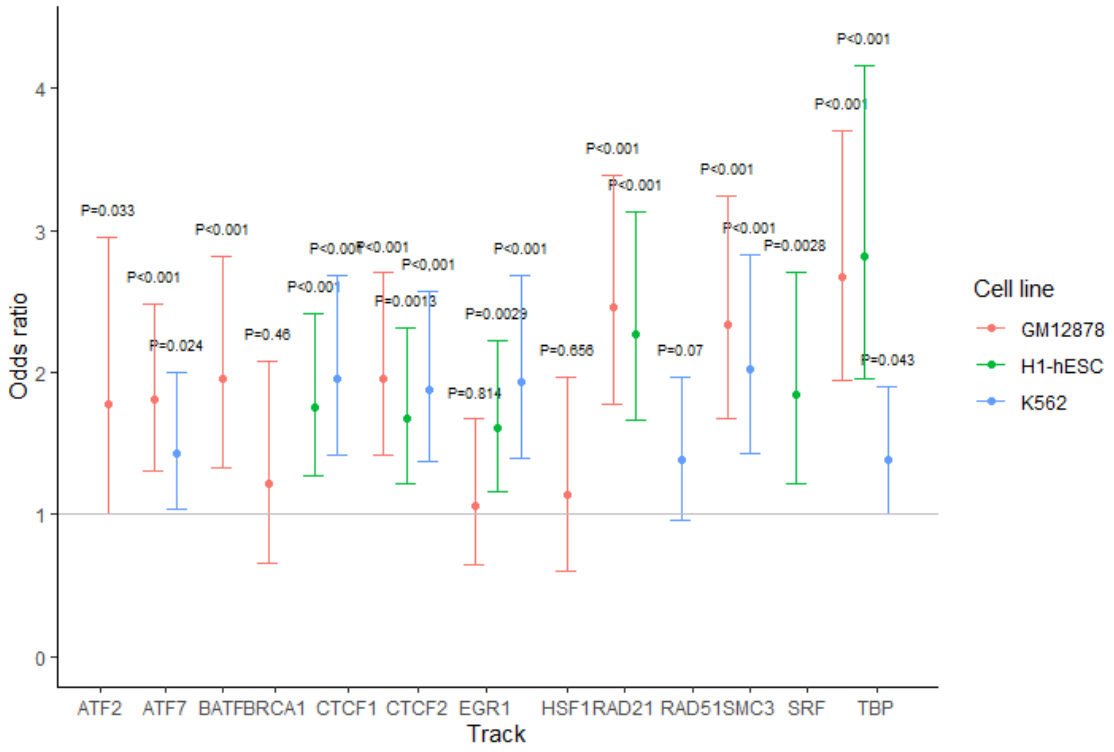


Figure 2:
This figure shows the transcription factor binding sites tested in this experiment. The odds ratio is displayed on the y-axis, and the different transcription factor binding sites on the x-axis. The color marks the cell line. A line for when odds ratio equals 1 is displayed as well. The p-value for the specific analysis is displayed until the p-value is less than 0.001. The accurate p-value is displayed in table 1 later in this section.

| $Track$ | $Odds$ | $Left$ | $Right$ | $Cellline$ | $P-value$ |
|---|---|---|---|---|---|
| $ATF2$ | $1,77$ | $1,00$ | $2,95$ | $GM12878$ | $3,3E-2$ |
| $ATF7$ | $1,81$ | $1,31$ | $2,48$ | $GM12878$ | $2,1E-4$ |
| $ATF7$ | $1,43$ | $1,04$ | $2,00$ | $K562$ | $2,39E-2$ |
| $BATF$ | $1,96$ | $1,33$ | $2,82$ | $GM12878$ | $4,5E-4$ |
| $BRCA1$ | $1,22$ | $0,66$ | $2,08$ | $GM12878$ | $4,6E-1$ |
| $CTCF1$ | $1,75$ | $1,27$ | $2,41$ | $H1-hESC$ | $1,35E-3$ |
| $CTCF1$ | $1,96$ | $1,42$ | $2,68$ | $K562$ | $2,18E-5$ |
| $CTCF2$ | $1,96$ | $1,42$ | $2,7$ | $GM12878$ | $2,98E-5$ |
| $CTCF2$ | $1,68$ | $1,22$ | $2,31$ | $H1-hESC$ | $1,35E-3$ |
| $CTCF2$ | $1,88$ | $1,37$ | $2,57$ | $K562$ | $7,7E-5$ |
| $EGR1$ | $1,06$ | $0,64$ | $1,67$ | $GM12878$ | $8,14E-1$ |
| $EGR1$ | $1,61$ | $1,16$ | $2,22$ | $H1-hESC$ | $5,89E-2$ |
| $EGR1$ | $1,93$ | $1,40$ | $2,68$ | $K562$ | $3,1E-5$ |
| $H3K27ac$ | $1,57$ | $1,11$ | $2,24$ | $GM12878$ | $8,5E-3$ |
| $H3K27ac$ | $1,40$ | $0,99$ | $2,02$ | $H1-hESC$ | $5,89E-2$ |
| $H3K27ac$ | $1,75$ | $1,24$ | $2,51$ | $K562$ | $1,11E-3$ |
| $H3K27me3$ | $3,00$ | $2,17$ | $4,18$ | $H1-hESC$ | $3,48E-12$ |
| $H3K27me3$ | $1,12$ | $0,81$ | $1,55$ | $K562$ | $4,7E-1$ |
| $H3K4me1$ | $2,57$ | $1,66$ | $4,15$ | $GM12878$ | $3,6E-6$ |
| $H3K4me1$ | $3,04$ | $1,90$ | $5,13$ | $H1-hESC$ | $1,87E-7$ |
| $H3K4me1$ | $1,84$ | $1,26$ | $2,73$ | $K562$ | $7,5E-4$ |
| $H3K4me3$ | $3,42$ | $2,09$ | $5,93$ | $GM12878$ | $1,57E-8$ |
| $H3K4me3$ | $2,69$ | $1,64$ | $4,67$ | $H1-hESC$ | $1,15E-5$ |
| $H3K4me3$ | $2,48$ | $1,65$ | $3,84$ | $K562$ | $2,26E-6$ |
| $H3K9ac$ | $2,24$ | $1,41$ | $3,73$ | $H1-hESC$ | $2,54E-4$ |
| $HSF1$ | $1,14$ | $0,6$ | $1,97$ | $GM12878$ | $6,56E-1$ |
| $RAD21$ | $2,46$ | $1,78$ | $3,39$ | $GM12878$ | $4,87E-8$ |
| $RAD21$ | $2,27$ | $1,66$ | $3,13$ | $H1-hESC$ | $1,68E-7$ |
| $RAD51$ | $1,38$ | $0,96$ | $1,97$ | $K562$ | $7,0E-2$ |
| $SMC3$ | $2,34$ | $1,68$ | $3,24$ | $GM12878$ | $5,11E-7$ |
| $SMC3$ | $2,02$ | $1,43$ | $2,83$ | $K562$ | $6,52E-5$ |
| $SRF$ | $1,84$ | $1,22$ | $2,71$ | $H1-hESC$ | $2,78E-3$ |
| $TBP$ | $2,67$ | $1,94$ | $3,7$ | $GM12878$ | $4,13E-10$ |
| $TBP$ | $2,82$ | $1,96$ | $4,16$ | $H1-hESC$ | $1,62E-9$ |
| $TBP$ | $1,38$ | $1,00$ | $1,90$ | $K562$ | $4,3E-2$ |

Table 1: Data for both transcription factor binding sites and histone markers.

# 6 Discussion

## 6.1 Cell line differences

As shown in the results, there are big differences in expression between different cell lines for some of the measurements like TBP, EGR1 and H3K27me3. This is particularly visible in the difference between K562 and the other 2 cell lines in binding sites for TBP. In general the structure of the human genome should be fairly similar, especially within these structures which has proven to be conserved in prokaryotes as well as eukaryotes. These differences could be a result of multiple sources of error as described under sources of error. However, in most of the data all cell lines support the same results such as RAD21, CTCF1 and 2, ATF7, H3K27ac, H3K4me1 and H3K4me3. In some of the data such as CTCF2 and H3K27ac the testing of additional cell lines support the data. The data from CTCF1 and CTCF2 overlap for H1-hESC and K562, which is consistent with them being measurements of the same TF in the same cell line.

## 6.2 Histone markers

To summarize the results, there is a slight observable enrichment of H3K27ac for all cell lines. All cell lines show enrichment of H3K4me1/3, and H1-hESC shows an enrichment of H3K27me3 whereas K562 fails to reproduce the results. This data supports the hypothesis that IEGs are under considerable regulation. All the histone markers used in this study apart from H3K27me3 are associated with active transcription. Most of them being significantly enriched supports the current evidence that shows that IEGs are epigenetically primed for active transcription. H1-hESC being enriched in H3K27me3 is especially interesting as it is the only repressive marker used in this study. Here H1-hESC provides an odds ratio of 3.00 (2.17 - 4.18) whereas K562 completely fails to reproduce the result. These results could indicate that methylation in this area changes over time depending on the current cell status. If that is the case and the data can be reproduced, it could support the theory of IEGs being in a poised state containing repressive as well as activating markers at the same time.

## 6.3 Transcription factor binding sites

The enrichment of ATF2, ATF7, BATF, SRF, TBP and EGR1 all signal that IEGs are genetically primed by having a lot of binding sites for TFs which facilitate transcription. Having more TF binding sites contribute to the the binding sites reacting with TFs

at lower concentrations, and thus making the gene more responsive to changes in TF concentration. HSF1 is a regulatory TF that is not enriched in this study.

Enrichment in CTCF is indicative of the highly regulated IEGs. CTCF is known to work with the cohesin molecule to regulate long-range interactions. The upregulation of CTCF binding sites may therefore indicate that IEGs are targets for a high level of long range interactions. This evidence is supported by the previously described epigenetic poised state of which IEGs seem to be.

The enrichment in RAD21 and SMC3 is by far significant, meaning that they are some of the enrichments found in this analysis with the highest odds ratio and lowest p-value. This is interesting as it indicates that transcription is further facilitated by easy unwinding of the DNA helix. This evidence indicates that the cohesin complex has a high affinity to IEGs, and is involved in the total regulation of many different IEGs. RAD51 was tested as an indicator of DNA double strand breaks in IEGs. The result from this test is borderline significant with an odds ratio of 1.38 (0.96 - 1.97) and a p-value of 0.07. It consolidates the current evidence supporting that DNA double strand breaks might play an important role in regulation of IEGs, especially as DNA double strand breaks alone has been shown to be sufficient for IEG transcriptional activation.

## 6.4 Sources of error

Differences between cell lines has been partly discussed in the section "Cell line differences". Some of the differences may be caused by errors in the data sampling and analysis. This is because the total number of data points is low enough that small errors in the ENCODE data may cause large errors in the analyzed data. The cell lines used in tier 1 ENCODE data are cancer cell lines, and multiple mutations in the genome are prone to affect the data for this analysis. Some of the IEGs are proto-oncogenes and it is likely that there are some mutations both in regulatory regions and exons of these genes in particular.

For histone data the methylation structure changes over time and the genomic state of the cell in question influences the data material in a way that makes the data not represent actual human cells in vivo.

For TF binding sites the data analyzed represent 1800 upstream base pairs and 200 downstream base pairs of the gene in question. Data outside these parameters is not represented in these results, even though they may be of importance.

# 7 Conclusion

The results from this analysis contributes to the current evidence that supports that IEGs are regulated through multiple mechanisms in a way that differs from other genes. This study shows that IEGs are enrichment in CpG-islands, TATA-boxes, cohesin complex binding sites, CTCF and other long range interaction indicators, RAD51 as an indicator of DNA double strand breaks, histone methylation and acethylation status. All these contribute to create a complex image of IEGs being under high, and perhaps redundant regulation, and that this regulation probably is necessary for the rapid response of the IEGs. For the genes to respond in such a quick manner to different stimuli they seem to be primed for activation and transcription through all of the mechanisms described above. In the search to discover more IEGs many of these different qualities of general IEGs can probably be used as they in combination seem to be vital for the rapid gene response. It would be very interesting to further use this data in combination with machine learning to attempt to identify similar genes which may be IEGs. Furthermore this study has primarily looked at TFs and histone markers that were already expected to be enriched in IEGs. Many TFs and histone markers that are not consistently enriched has thus not been identified.

# 8 References

# References


[1] Shahram Bahrami and Finn Drabløs. "Gene regulation in the immediate-early response process". In: *Advances in Biological Regulation* 62 (2016), pp. 37–49. ISSN: 2212-4926. DOI: https://doi.org/10.1016/j.jbior.2016.05.001. URL: http://www.sciencedirect.com/science/article/pii/S221249261630001X.

[2] Swneke D. Bailey et al. "ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters". In: *Nature Communications* 6 (Feb. 2015). Article, 6186 EP -. URL: https://doi.org/10.1038/ncomms7186.

[3] Jung S. Byun et al. "Dynamic bookmarking of primary response genes by p300 and RNA polymerase II complexes". In: *PNAS* (Nov. 2009). URL: https://www.pnas.org/content/106/46/19286.short.

[4] Stuart K. Calderwood. "A critical role for topoisomerase IIb and DNA double strand breaks in transcription". In: *Transcription* 7.3 (May 2016). PMC4984685[pmcid], pp. 75–83. ISSN: 2154-1272. DOI: 10.1080/21541264.2016.1181142. URL: https://www.ncbi.nlm.nih.gov/pubmed/27100743.

[5] Malka Cohen-Armon, Adva Yeheskel, and John M Pascal. "Signal-induced PARP1-Erk synergism mediates IEG expression". In: *Signal transduction and targeted therapy* (Apr. 2019). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6459926/.

[6] Jason Ernst and Manolis Kellis. "Chromatin-state discovery and genome annotation with ChromHMM". In: *Nature protocols* (Dec. 2017). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5945550/.

[7] Trent Fowler, Ranjan Sen, and Ananda L. Roy. "Regulation of Primary Response Genes". In: *Molecular Cell* 44.3 (2011), pp. 348–360. ISSN: 1097-2765. DOI: https://doi.org/10.1016/j.molcel.2011.09.014. URL: http://www.sciencedirect.com/science/article/pii/S1097276511008045.

[8] Toshitsugu Fujita, Isabelle Piuz, and Werner Schlegel. "The transcription elongation factors NELF, DSIF and P-TEFb control constitutive transcription in a gene-specific manner". In: *FEBS Letters* 583.17 (2009), pp. 2893–2898. DOI: 10.1016/j.febslet.2009.07.050. eprint: https://febs.onlinelibrary.wiley.com/doi/pdf/10.1016/j.febslet.2009.07.050. URL: https://febs.onlinelibrary.wiley.com/doi/abs/10.1016/j.febslet.2009.07.050.

[9] Francisco T Gallo et al. "Immediate Early Genes, Memory and Psychiatric Disorders: Focus on c-Fos, Egr1 and Arc". In: *Frontiers in behavioral neuroscience* (Apr. 2018). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5932360/.

[10] Francisco T. Gallo et al. "Immediate Early Genes, Memory and Psychiatric Disorders: Focus on c-Fos, Egr1 and Arc". In: *Frontiers in behavioral neuroscience* 12 (Apr. 2018). PMC5932360[pmcid], pp. 79–79. ISSN: 1662-5153. DOI: 10.3389/fnbeh.2018.00079. URL: https://www.ncbi.nlm.nih.gov/pubmed/29755331.

[11] Shannon Healy, Protiti Khan, and James R. Davie. "Immediate early response genes and cell transformation". In: *Pharmacology Therapeutics* (Sept. 2012). URL: https://www.sciencedirect.com/science/article/pii/S0163725812002033.

[12] Judith Hyle et al. "Acute depletion of CTCF directly affects MYC regulation through loss of enhancer-promoter looping". In: *Nucleic acids research* (July 2019). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6648894/.

[13] Xiong Ji et al. "3D Chromosome Regulatory Landscape of Human Pluripotent Cells". In: *Cell stem cell* 18.2 (Feb. 2016). S1934-5909(15)00505-6[PII], pp. 262–275. ISSN: 1875-9777. DOI: 10.1016/j.stem.2015.11.007. URL: https://www.ncbi.nlm.nih.gov/pubmed/26686465.

[14] Iris Jonkers and John T. Lis. "Getting up to speed with transcription elongation by RNA polymerase II". In: *Nature News* (Feb. 2015). URL: https://www.nature.com/articles/nrm3953?draft=collection.

[15] Xiang Li et al. "The DNA Repair-Associated Protein Gadd45 Regulates the Temporal Coding of Immediate Early Gene Expression within the Prelimbic Prefrontal Cortex and Is Required for the Consolidation of Associative Fear Memory". In: *The Journal of neuroscience : the official journal of the Society for Neuroscience* (Feb. 2019). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6363930/.

[16] Ram Madabhushi et al. "Activity-Induced DNA Breaks Govern the Expression of Neuronal Early-Response Genes". In: *Cell* 161.7 (June 2015). S0092-8674(15)00622-4[PII], pp. 1592–1605. ISSN: 1097-4172. DOI: 10.1016/j.cell.2015.05.032. URL: https://www.ncbi.nlm.nih.gov/pubmed/26052046.

[17] Tyler E. Miller et al. "Transcription elongation factors represent in vivo cancer dependencies in glioblastoma". In: *Nature* 547.7663 (July 2017). nature23000[PII], pp. 355–359. ISSN: 1476-4687. DOI: 10.1038/nature23000. URL: https://www.ncbi.nlm.nih.gov/pubmed/28678782.

[18] O'Donnell et al. "Immediate-early gene activation by the MAPK pathways: what do and don't we know?" In: *Portland Press* (Jan. 2012). URL: https://portlandpress.com/biochemsoctrans/article/40/1/58/66323/Immediate-early-gene-activation-by-the-MAPK.

[19] Vladimir R. Ramirez-Carrozzi et al. "A Unifying Model for the Selective Regulation of Inducible Transcription by CpG Islands and Nucleosome Remodeling". In: *Cell* (July 2009). URL: https://www.sciencedirect.com/science/article/pii/S0092867409004450.

[20] Morten Rye et al. "Chromatin states reveal functional associations for globally defined transcription start sites in four human cell lines". In: *BMC Genomics* 15.1 (Mar. 2014), p. 120. ISSN: 1471-2164. DOI: 10.1186/1471-2164-15-120. URL: https://doi.org/10.1186/1471-2164-15-120.

[21] Ana Soloaga et al. "MSK2 and MSK1 mediate the mitogen[U+2010] and stress[U+2010]induced phosphorylation of histone H3 and HMG[U+2010]14". In: *The EMBO Journal* (June 2003). URL: https://www.embopress.org/doi/full/10.1093/emboj/cdg273.

[22] Frank M. J. Sommerlandt et al. "Immediate early genes in social insects: a tool to identify brain regions involved in complex behaviors and molecular processes underlying neuroplasticity". In: *Cellular and molecular life sciences : CMLS* 76.4 (Feb. 2019). PMC6514070[pmcid], pp. 637–651. ISSN: 1420-9071. DOI: 10.1007/s00018-018-2948-z. URL: https://www.ncbi.nlm.nih.gov/pubmed/30349993.

[23] John W. Tullai et al. "Immediate-Early and Delayed Primary Response Genes Are Distinct in Function and Genomic Architecture*". In: *Journal of Biological Chemistry* (Aug. 2007). URL: http://www.jbc.org/content/282/33/23981.short.

[24] Joseph T Wade and Kevin Struhl. "The transition from transcriptional initiation to elongation". In: *Current opinion in genetics  development* (Apr. 2008). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2563432/.

[25] L. Ashley Watson and Li-Huei Tsai. "In the loop: how chromatin topology links genome structure to function in mechanisms underlying learning and memory". In: *Current opinion in neurobiology* 43 (Apr. 2017). S0959-4388(16)30249-5[PII], pp. 48–55. ISSN: 1873-6882. DOI: 10.1016/j.conb.2016.12.002. URL: https://www.ncbi.nlm.nih.gov/pubmed/28024185.

[26] Chuan Xu et al. "Identification of Immediate Early Genes in the Nervous System of Snail Helix lucorum". In: *eNeuro* 6.3 (May 2019). ENEURO.0416-18.2019[PII], ENEURO.0416–18.2019. ISSN: 2373-2822. DOI: 10.1523/ENEURO.0416-18.2019. URL: https://www.ncbi.nlm.nih.gov/pubmed/31053606.

# 9 References to data used in the study

ATF2 - GM12878 - encTfChipPkENCFF127GYQ
ATF7 - GM12878 - encTfChipPkENCFF726VEK
ATF7 - K562 - encTfChipPkENCFF868QLL
BATF - GM12878 - encTfChipPkENCFF593FBF
BRCA1 - GM12878 - encTfChipPkENCFF082JDH
CTCF1 - H1-hESC - encTfChipPkENCFF093VEE
CTCF1 - K562 - encTfChipPkENCFF085HTY
CTCF2 - GM12878 - encTfChipPkENCFF710VEH
CTCF2 - H1-hESC - encTfChipPkENCFF402JJK
CTCF2 - K562 - encTfChipPkENCFF738TKN
EGR1 - GM12878 - encTfChipPkENCFF618EFD
EGR1 - H1-hESC - encTfChipPkENCFF004WYV
EGR1 - K562 - encTfChipPkENCFF220IXP
H3K27ac - GM12878 - wgEncodeBroadHistoneGm12878H3k27acStdPk
H3K27ac - H1-hESC - wgEncodeBroadHistoneH1hescH3k27acStdPk
H3K27ac - K562 - wgEncodeBroadHistoneK562H3k27acStdPk
H3K27me3 - H1-hESC - wgEncodeBroadHistoneH1hescH3k27me3StdPk
H3K27me3 - K562 - wgEncodeBroadHistoneK562H3k27me3StdPk
H3K4me1 - GM12878 - wgEncodeBroadHistoneGm12878H3k04me1StdPkV2
H3K4me1 - H1-hESC - wgEncodeBroadHistoneH1hescH3k4me1StdPk
H3K4me1 - K562 - wgEncodeBroadHistoneK562H3k4me1StdPk
H3K4me3 - GM12878 - wgEncodeBroadHistoneGm12878H3k04me3StdPkV2
H3K4me3 - H1-hESC - wgEncodeBroadHistoneH1hescH3k4me3StdPk
H3K4me3 - K562 - wgEncodeBroadHistoneK562H3k4me3StdPk
H3K9ac - H1-hESC - wgEncodeBroadHistoneH1hescH3k9acStdPk

HSF1 - GM12878 - encTfChipPkENCFF308ZGG
RAD21 - GM12878 - encTfChipPkENCFF753RGL
RAD21 - H1-hESC - encTfChipPkENCFF732XGR
RAD51 - K562 - encTfChipPkENCFF096XMD
SMC3 - GM12878 - encTfChipPkENCFF686FLD
SMC3 - K562 - encTfChipPkENCFF041YQC
SRF - H1-hESC - encTfChipPkENCFF725LDD
TBP - GM12878 - encTfChipPkENCFF490VYC
TBP - H1-hESC - encTfChipPkENCFF629MBV
TBP - K562 - encTfChipPkENCFF380FJL