University of Natural Resources and

Applied Life Sciences, Vienna

Department of Biotechnology

Institute of Applied Microbiology

Identification and characterization of CHO endogenous gene regulatory elements

Master Thesis

Submitted by Andreas Michael Maccani

Vienna, February 2010

Supervisor:

Ao.Univ.Prof. Dipl.-Ing. Dr.nat.techn. Reingard Grabherr

Dipl.-Ing. Dr.nat.techn. Wolfgang Ernst

Assistant Supervisor:

Dipl.-Ing. Jens Pontiller

AUSTRIAN CENTER
of Biopharmaceutical
Technology

“Scientists are people of very dissimilar temperaments doing different things in very different ways. Among scientists are collectors, classifiers and compulsive tidiers-up; many are detectives by temperament and many are explorers; some are artists and others artisans. There are poet-scientists and philosopher-scientists and even a few mystics.”
Peter Medawar, Pluto’s Republic, Oxford University Press, New York, 1982, p. 116.

Acknowledgments

Writing this master thesis would not have been possible without the support of many people.

First of all, I would like to express my gratitude to Professor Reingard Grabherr and Dr Wolfgang Ernst for giving me the opportunity to conduct my study in their lab. Professor Grabherr has always had an open door for my concerns and her advice and encouragements were very helpful. Dr Ernst's scientific support and his suggestions were of great value for this thesis. I very much appreciate the collegial collaboration with both of them.

I am deeply grateful to Jens Pontiller for his generous guidance, continuing assistance, and the great teamwork, and to Martina Baumann for building the foundation of this research.

Furthermore, I owe a huge dept to all my other colleagues of the Baculo group, namely Esther Egger, Stefan Gross, Dr Stefan Heinl, Denise Herold, Florian Krammer, Silvia Löbsch, Jürgen Mairhofer, Dieter Palmberger, Katharina Spath, Philipp Tauber, Christopher Tauer as well as my former colleagues Jeny Cherukad, Maria Kern, and Dr Haruthai Thaisuchat for discussion, encouragement, and assistance. They created a positive atmosphere which made it a pleasure to work in this lab.

I am also indebted to the members of the ACBT team for sharing their knowledge and experience. Georg Hinterkörner deserves special thanks for taking care of our cultured Chinese hamster ovary cells.

Thanks are also due to Ingo Klancnik whose assistance was necessary for measuring bioluminescence.

Lastly, and with no less gratitude, I would like to thank my partner and my parents for their patience and enduring support.

Abstract

The Chinese hamster ovary (CHO) cell line is one of the most widely used mammalian expression systems for the production of therapeutic proteins. Today, strong viral promoters such as the cytomegalovirus (CMV) major immediate early promoter or the Simian virus 40 (SV40) immediate early promoter are commonly used for driving transcription. Beside high expression levels, several drawbacks are related to viral promoters as they are not regulated by the host cell. The permanent over-expression of a heterologous gene can lead to various stress reactions that affect correct posttranslational processing and may finally induce the premature activation of apoptosis. In addition, some viral promoters are cell cycle dependent and can also be silenced in certain stable cell lines resulting in a considerable heterogeneity within the cell population. These undesired effects might be avoided by using CHO endogenous gene regulatory elements. Modern approaches for genome-wide, high-throughput recognition of gene regulatory elements rely on whole genome sequence data, which are currently not available for the Chinese hamster, though. In this study, CHO endogenous promoters were directly identified starting from CHO genomic libraries. The applied method referred to as Library PCR enabled the *in vitro* amplification of the 5’ flanking region of a specific gene and thus the targeting of promoter sequences of highly expressed genes. Furthermore, Inverse PCR which allows the amplification of the unknown region that flank known genomic sequence could be demonstrated to be suitable for the CHO genome. In addition, Inverse PCR was successfully applied to elucidate the 5’ and 3’ flanking regions of previously identified gene regulatory elements by Library PCR. Promoter activity of obtained full-length fragments as well as of truncated promoter constructs was analyzed using a luciferase reporter assay. The transcriptionally active 5’ flanking region of the *Rps6* (ribosomal protein S6) gene or of a pseudogene thereof could be identified and characterized by employing these techniques.

Kurzfassung

Die CHO-Zelllinie (Chinese hamster ovary) ist eines der am häufigsten verwendeten Säuger-Expressionssysteme für die Produktion von therapeutischen Proteinen. Heute werden vor allem starke virale Promotoren wie der Cytomegalovirus (CMV) Major Immediate Early Promotor oder der Simian-Virus 40 (SV40) Immediate Early Promotor für die Transkription verwendet. Virale Promotoren ermöglichen zwar einerseits hohe Expressionslevel, andererseits sind auch einige Nachteile mit ihrer Verwendung verbunden, da sie nicht von der Wirtszelle reguliert werden. Die permanente Überexpression eines heterologen Gens kann viele verschiedene Stressreaktionen hervorrufen, welche eine korrekte posttranslationale Prozessierung beeinflussen und schlussendlich eine vorzeitigen Aktivierung der Apoptose auslösen können. Des Weiteren sind manche virale Promotoren abhängig vom Zellzyklus und können auch in gewissen stabilen Zelllinien stillgelegt werden, was eine beträchtliche Heterogenität innerhalb der Zellpopulation zur Folge hat. Durch die Verwendung CHO-endogener regulatorischer Elemente könnten diese unerwünschten Effekte vermieden werden. Moderne Ansätze für die genomweite Hochdurchsatzidentifizierung von Genregulationselementen beruhen auf kompletten Genomsequenzdaten, die allerdings für den Chinesischen Hamster derzeit noch nicht verfügbar sind. In dieser Studie wurden CHO-endogene Promotoren direkt ausgehend von genomischen Libraries identifiziert. Die angewandte Methode (hier als Library PCR bezeichnet) ermöglichte die *in vitro* Amplifizierung der 5‘-flankierenden Region eines spezifischen Gens und somit die gezielte Identifizierung von Promotoren hochexprimierter Gene. Darüber hinaus konnte gezeigt werden, dass die Inverse PCR für das CHO-Genom geeignet ist. Diese Methode ermöglicht die Amplifizierung von unbekannten Regionen die eine bekannte genomische Sequenz flankieren. Sie konnte erfolgreich angewandt werden um die 5‘- und 3‘-flankierenden Regionen von durch Library PCR identifizierten Genregulationselementen aufzuklären. Die Promoteraktivität der erhaltenen Fragmente sowie verkürzter Promotorkonstrukte wurde mittels Luciferase-Reporterassay analysiert. Durch Anwendung dieser Verfahren konnte die transkriptionsaktive 5‘-flankierende Region des *Rps6*-Gens (ribosomal protein S6) oder eines Pseudogens davon identifiziert und charakterisiert werden.

Contents

1 Introduction

1.1 The driving force behind CHO promoter identification

In recent years, cultivated mammalian cells have become the most significant host system for the production of recombinant proteins for therapeutic applications mainly due to their capacity for proper protein folding, assembly, post-translational modification like glycosylation, and product secretion. Because of these outstanding features, quality and efficiency of a protein can be far better when expressed in mammalian cells compared to other host systems like bacteria, yeast, or plants. The major factors of success for the process development for biopharmaceuticals are product quality, drug safety, speed of generating clinical phase 1–3 materials, economy of the manufacturing process, and regulatory acceptance. These factors also have great impact on the evaluation of mammalian cell expression systems to be used for the production of pharmacologically active proteins [**1**,**2**]. Today, the majority of all recombinant pharmaceutical proteins for mass production are expressed in mammalian cells, most of all in immortalized Chinese hamster ovary (CHO) cells, mainly because of their superior capacity for single-cell suspension growth [**3**]. But also other cell lines like mouse myeloma-derived NS0 cells [**4**], baby hamster kidney (BHK) cells, human embryo kidney (HEK-293) cells [**5**], and human retinal-derived PER.C6 cells [**6**] grow well in suspension and are regulatory approved for the production of recombinant proteins.

The maximization of therapeutic protein yield requires optimization of specific productivity and maintaining of a viable cell biomass as well as a stable protein production over an extended period of time [**7**]. Today, optimizations of mammalian expression systems mainly focus on process, media, and cell line improvements. However, the expression level of a heterologous gene is primarily determined by strength and efficiency of regulatory sequences directing its transcription and processing into messenger RNA (mRNA) as well as the chromosomal integration site, the copy number, the type of the particular protein, and the efficiency of translation into protein. A minimum mammalian expression cassette consists of a promoter which is 5’ upstream of the gene to be expressed and a 3’ untranslated region (3’ UTR) containing at least one polyadenylation (polyA) sequence (AATAAA) which is necessary for termination of transcription and polyadenylation of the mRNA 3’ end. Other elements such as enhancers, introns, or chromatin modifiers are frequently used to further boost expression [**8**]. However, the core promoter which consists of short regulatory sequences (e.g. TATA box, downstream promoter element) for transcription factor and RNA polymerase transcription complex binding is the central element of gene expression in mammalian cells [**9**,**10**].

High-level production of proteins in these cells requires strong constitutive promoters which are preferably active in a wide range of cell types. Today, the most common promoters fulfilling these criteria are of viral origin. For most host systems, the human and mouse cytomegalovirus (CMV) major immediate early promoter is the promoter of choice. But also the Simian virus 40 (SV40) immediate early promoter and the Rous sarcoma virus (RSV) long terminal repeat (LTR) promoter are frequently used for driving heterologous gene expression in mammalian cells [**8**,**11**]. Equally, inducible promoters might have value for the production of therapeutic proteins, especially if the expressed proteins inhibit growth of the host cell. For example, using the inducible mouse mammary tumor virus (MMTV) promoter in a CHO overexpression system could increase the product yield considerably in comparison to the same system using the strong constitutive SV40 promoter [**12**].

Beside high expression levels of the genes of interest, several drawbacks are related to viral promoters as they are not regulated by the host cell. Thus, the permanent over-expression of the recombinant protein can lead to various excessive stress reactions like the unfolded protein response (UPR) or the endoplasmatic reticulum (ER) stress response. These phenomena affect the correct posttranslational processing of the recombinant protein and may finally induce premature activation of cellular apoptotic pathways. Furthermore, viral promoters are cell cycle dependent showing the highest transcriptional activity in the S phase, and they can also be silenced in certain stable cell lines resulting in a considerable heterogenic population of transfectants regarding the amount of protein expressed [**8**]. These undesired properties of viral promoters could be avoided by using of cell endogenous transcription regulatory elements as they are under the control of the host cell's regulatory network.

Outstanding examples corroborating this hypothesis are the transcription regulatory sequences from the Chinese hamster elongation factor-1α (EF-1α) which have been described as driving high-level expression in CHO cells as well as in non-hamster mammalian cells [**13**]. In this study, the 5' and 3' flanking transcription control regions of the Chinese hamster EF-1α (CHEF1) gene were used for the expression of heterologous genes in various mammalian cell lines revealing expression levels which were significantly higher than CMV promoter driven vectors. Furthermore, the authors of this study hypothesized that transcription control regions from highly expressed CHO genes might reduce the requirement for gene amplification which is a stepwise, labor-intensive and time-consuming procedure typically required for establishing a high-level mammalian expression cell line.

All of these potential benefits push on scientific research for the identification of endogenous gene regulatory elements. In the master thesis at hand, I describe a method for the discovery of 5’ flanking regions of highly expressed CHO endogenous genes and the characterization of the identified transcriptionally active elements.

1.2 The CHO expressions system

Recombinant protein therapeutics have revolutionized modern medicine in the past decades as they provide innovative and effective therapies for many previously incurable diseases, ranging from cancer to infertility. In 1987, the tissue plasminogen activator (r-tPA, Activase) was the first recombinant therapeutic protein produced in mammalian cells that gained approval for clinical application. The production hosts used for synthesizing r-tPA were Chinese hamster ovary (CHO) cells. Today, a variety of approved biologics produced in CHO cell lines are available on the market, such as Erythropoietin, Interferon-β, Factor IX, and Factor VIII to name but a few. Because of the amassed knowledge and expertise in applying CHO cell lines for the expression of recombinant glycoproteins over the past decades, CHO cells will remain the most dominant host system for the production of recombinant proteins, at least in the near future [**14**].

1.2.1 History: From hamster to cell culture

Chinese hamsters (*Cricetulus griseus*) are hamsters belonging to the *Cricetidae*, a family of rodents. They are native to the desert of northern China and Mongolia. Already in 1919, Chinese hamsters were used in laboratory for typing pneumococci. In the 1950s, it was mainly George Yerganian of the Boston Children's Cancer Research Foundation who pioneered the research in hamster genetics. At that time it was found out that Chinese hamsters have a rather low chromosome number of 22 which made them very suited models for radiation cytogenetics and tissue cultures [**15**]. It was then in 1957, when Theodore T. Puck of the University of Colorado Medical Center in Denver isolated an ovary of a female Chinese hamster and succeeded to establish the original Chinese hamster ovary (CHO) cell line [**16**]. In 1980, Gail Urlaub and Lawrence A. Chasin of the Columbia University in New York isolated mutants of Chinese hamster ovary cells lacking the metabolic enzyme dihydrofolate reductase (DHFR) after mutagenesis and selection of thymidine auxotrophic cells [**17**]. DHFR catalyzes the formation of intracellular tetrahydrofolic acid, which is an essential cofactor of the *de novo* pathways for the nucleic acid biosynthesis. Fully deficient CHO mutants require thymidine, glycine, and hypoxanthine for growth and can be used for the selection of cells expressing exogenous proteins. Because of the great adaptive ability and the ease of maintenance, the CHO dhfr$^-$ cell line has become the most widely used mammalian expression system for the production of biologically active heterologous proteins.

1.2.2 Recombinant protein production in CHO cells

The characteristics and maximum achievable yield of a recombinant protein are primarily affected by the choice of host cells. Correct protein folding and post-translational modifications like glycosylation determine the solubility, stability, biological activity, and half-life time in human bodies and thus the efficacy of a therapeutic protein. Another key issue in regard to the selection of the most suited host system is product safety. Production cell lines must be free from any human pathogenic agents. From an industrial point of few, host cells should be able to grow in suspension allowing volumetric scale up by using large bioreactors. Furthermore, the host cells must be accessible to genetic modifications allowing the introduction of foreign DNA and the high-level expression of the desired protein [**14**].

The gained experience over the past decades showed that CHO cells comprise many of these properties. Thus, CHO cells are able to produce glycoproteins that are compatible and biological active in humans. Moreover, they have been proven as safe and actually harboring no human pathogens. A study demonstrated that a multitude of human pathogenic viruses including human immunodeficiency virus (HIV), influenza, polio, herpes, and measles viruses do not replicate in CHO cells [**18**]. The most convincing factors for the industry are that CHO cells can be easily genetically manipulated and have the ability to high density growth in suspension cultures allowing bioreactor scales already exceeding 10,000 l. Furthermore, the availability of CHO dhfr$^-$ cells enables the effective selection of stable production clones and amplification of the desired gene leading to high specific productivity levels.

CHO dhfr$^-$ cells are auxotrophs for glycine, hypoxanthine, and thymidine, and hence so these nutrients must be supplemented for growth. However, transfection of cells using the heterologous gene combined with a functional copy of the DHFR gene allows a clonal selection when grown in media lacking these supplements. This system also facilitates the amplification of the introduced gene of interest. For this purpose, the cells need to be cultured in media containing high amounts of methotrexate (MTX), which is a folic acid analog blocking DHFR activity. In order to survive, the cell responds with amplifying the copy number of the DHFR gene. This effect also leads to the co-amplification of the transfected gene of interest and thus enables the generation of a high producing cell line.

A considerable disadvantage in using CHO as well as other mammalian cell lines for the production of recombinant proteins is the rather low volumetric yield of product. In comparison to microbial host systems, the productivity of mammalian cell cultures is generally about

10 to 100-fold lower and hence industrial production requires very large and expensive production facilities. Over the past decades, process improvement was mainly determined by the optimization of culture strategies, media formulation, process monitoring and control as well as by screening and cell line development. Today, fed-batch cultivation is most widely used allowing a higher cell density and longer culture duration, which enables final product titers of 1–5 g l^{-1} [**14**].

1.2.3 Novel strategies in CHO cell engineering

One of the main costs in the production of recombinant therapeutic proteins is related to downstream purification. These costs can be reduced by maintaining a high-density cultivation of viable high-level expressing cells for an extended period of time. Hence, novel strategies for the engineering of such cells are needed.

Currently, cell line development is a predominantly empirical process which is very labor-intensive and time-consuming. Difficulties in proper cell line engineering are primarily related to the poor understanding of the biology and physiology of mammalian cells. Therefore, efforts to understand the underlying mechanism are required in order to develop metabolically engineered cell lines with enhanced productivity. Various aspects of cellular mechanisms, including metabolism, protein processing, and the balancing pathways of cell growth and apoptosis need to be considered as they have influence over well growth and production characteristics [**19**].

Viability of mammalian cells can be considerably improved by limiting the lactate production in culture. Lactate accumulates as a result of the consumption of glucose and other nutrients in excess of the requirement for cell growth [**20**]. Engineered CHO cells over-expressing the anti-sense lactate dehydrogenase A (LDH-A) RNA and the glycerol-3-phosphate dehydrogenase (GPDH) facilitate the decrease in lactic acidoses and cell death due to cell-cell contact inhibition as well as the increase of growth rate and oxidoresistance [**21**]. In another study, the GLUT5 fructose transporter was expressed in CHO cells allowing them to utilize fructose instead of glucose. As the GLUT5 fructose transporter has a high K_m value for its substrate, fructose is supplied at a more moderate rate into the cells. When GLUT5 expressing clones were cultured in media containing fructose in place of glucose, sugar consumption and lactate production rates were drastically reduced [**22**].

Other efforts have been made to increase the sensitivity of production cell lines to apoptosis using metabolic engineering strategies. A considerable number of cells die following a genetically defined program known as apoptosis or programmed cell death during standard bioreactor cultivation [**23**]. Serum components are known to be chiefly responsible for apoptosis protection. However, the demand of serum-free media has exacerbated the problem of premature apoptosis initiation [**24**]. A considerable number of antiapoptosis engineering studies have been published with main focus on over-expression of *bcl-2* family members which regulate induction of the caspase-9-dependent apoptosis pathway at the outer membrane of mitochondria [**24**,**25**].

Genomics and proteomics tools such as DNA microarrays and mass spectrometry can make a significant contribution towards a greater understanding of genetic regulatory circuitries. DNA microarrays enable the analysis of expression levels of thousands of genes in parallel. By using this tool, relevant apoptosis signaling genes in batch and fed-batch CHO cell cultures have already been identified allowing for specific targeting of these genes to prolong cell viability in culture [**26**].

RNA interference (RNAi) is further tool which can be applied to increase productivity and quality of recombinant proteins expressed in CHO cells. This novel technology enables the silencing of gene expression in cells or organisms. Several approaches including the silencing of apoptosis-associated gene expression, protein glycosylation-associated gene expression, lactate dehydrogenase, and dihydrofolat reductase are described in the literature [**27**].

1.2.4 Exploring the CHO genome

Over the past years several mammalian genomes like human, mouse, or rat have been completely sequenced. However, the Chinese hamster, despite the great importance of CHO cells for recombinant protein production, is not among them. The comprehensive sequence data available for mouse and human have pushed the development of tools displaying the ability to investigate gene and protein expression in a high-throughput manner, which revolutionized the research of these organisms [**28**].

However, to facilitate gene discovery and genetic engineering cDNA libraries were constructed and expressed sequence tags (ESTs) were sequenced starting in 2004 [**29**]. The founding of the Consortium of Chinese Hamster Ovary Cell Genomics in 2006 [**30**] could intensify the efforts in generating CHO sequence data. To date, more than 68,000 ESTs have been sequenced and been assembled into over 28,000 unique CHO transcripts [**31**]. Sequence alignment analyses with orthologous genes from other species revealed that CHO transcripts are generally most similar to mouse, but also a significant number of genes have the highest similarity to rat or human, whereas the identity between mouse genes and CHO genes ranges from 75 – 97% [**32**]. With regard to the number of chromosomes, Chinese Hamster and mouse are quite different (mouse 2n = 40, Chinese hamster 2n = 22), however due to large chromosome segment rearrangements, the relative positions of most genes to each other is comparable [**29**].

The generated sequence repertoire has also been used to design both CHO cDNA and Affymetrix microarrays for transcriptome profiling which can be greatly useful for analyzing gene expression of CHO cells cultured under conditions important in bioprocessing [**29**,**31**].

Furthermore, CHO and Chinese hamster ESTs have been mapped onto the mouse genome in order to create a genomic scaffold for the Chinese hamster which can contribute to future genome sequencing efforts [**28**]. However, the whole genome sequence of the Chinese hamster is still not available inhibiting studies of the structural and regulatory characteristics of the CHO genome. Just recently, Omasa et al. reported the construction of a CHO genomic bacterial artificial chromosome (BAC) library which presumably covers the CHO genome five times [**33**]. Additionally they have estimated the size of the Chinese hamster genome to about 2.8 Gb, using a calculation method described by Dolezel et al. [**34**].

Although the availability of CHO cDNAs has facilitated various approaches for the identification of gene regulatory sequences such as described in the master thesis at hand, the existence of the completely sequenced CHO genome would be a tremendous benefit concerning this matter. However, as high-throughput sequencing technologies have become widely available recently, the complete genomic sequence of the Chinese hamster should be on hand in the near future.

1.3 Eukaryotic transcription regulation

In the eukaryotic genome, the genetic information encoding for proteins is transcribed into mRNA by a large multisubunit enzyme called the RNA polymerase II (Pol II). The activity of Pol II is highly divers for individual genes and is specifically regulated by combinatorial molecular interactions of various transcription factors with each other and with gene specific DNA sequences. Thousands of factors regulating transcription have already been identified, whereas most of them are proteins but also RNAs are involved. They recruit Pol II to the gene's promoter in order to initiate the synthesis of a full-length RNA transcript at the transcription start site (TSS) [**35**].

1.3.1 Gene promoters

When, where, and at what level a gene is transcribed, is determined by the DNA sequence in and around the promoter. There are mainly three parts, the core promoter, proximal promoter elements, and distal regulatory elements like enhancers, silencers, insulators, and locus control regions (LCRs) that regulate transcription (Figure 1-1) [**36**,**35**]. These *cis*-acting transcriptional regulatory elements are composed of binding sites for *trans*-acting transcription factors, which can either contribute to the activation or repression of transcription.

The Pol II core promoter is generally defined as the DNA sequence minimally required for the accurate initiation of transcription [**9**]. The elements of the core promoter sequence determine the assembly of distinct preinitiation complexes (PICs) which consist of the general transcription factors (GTFs) and the Pol II. Promoter-proximal regions as well as enhancer regions bind specific transcription factors, called activators or repressors. They can directly interact with general transcription factors, but mainly they conduct regulation via coregulators. Some of these mostly multiprotein complexes can influence expression by direct interaction with Pol II or GTFs, others however by modifying nucleosomes or the chromatin structure [**35**].

Transcription initiation can be divided into two types – focused and dispersed [**37**,**38**]. Promoters with focused initiation have either a single transcription start site or several start sites within a small number of nucleotides. In contrast, promoters having a dispersed initiation comprise multiple weak start sites over a broad region of about 50 to 100 nucleotides. Focused transcription can be found in all organisms and is probably the most common type of transcription in simpler organisms. However, in vertebrates approximately 70% of genes use the dispersed mode of transcription initiation. Such promoters are generally located in CpG islands and are typically associated with constitutive genes, whereas focused promoters are usually connected with regulated genes [**38**].

Sequence motifs like TATA box, BRE (TFIIB recognition element), Inr (initiator), MTE (motif ten element), DPE (downstream promoter element), DCE (downstream core element), and XCPE1 (X core promoter element 1) are generally found in focused core promoters (discussed in more detail below) [**37**,**38**]. However, TATA box, BRE, DPE, and MTE motifs are generally absent in dispersed promoters [**39**,**40**].

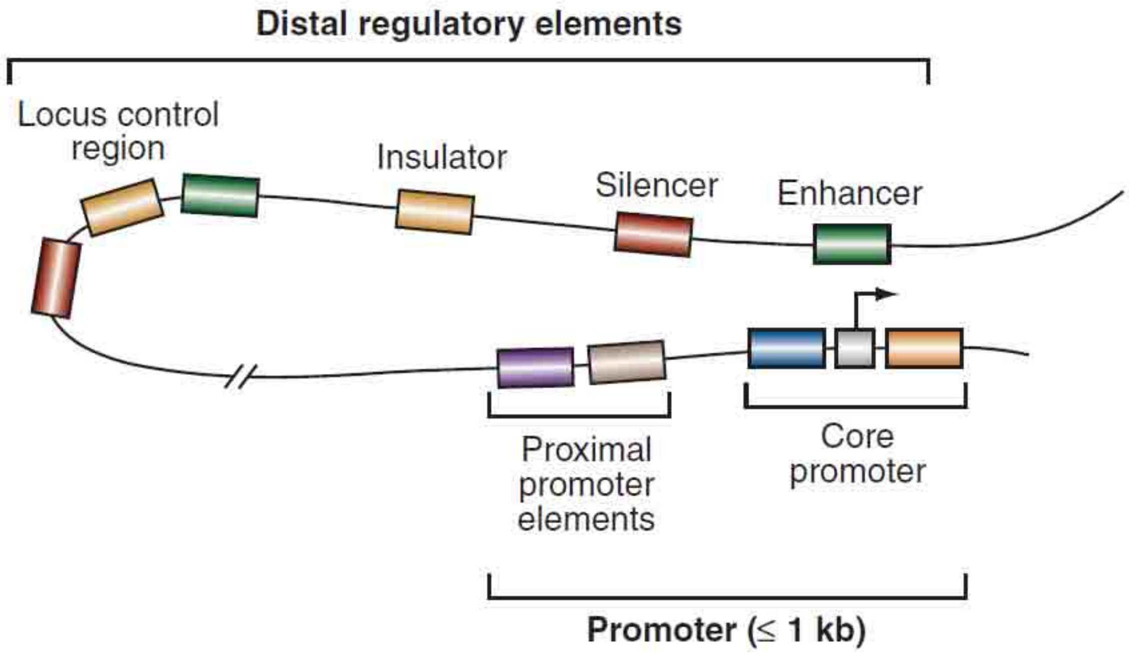


Figure 1-1: Schematic illustration of a typical gene regulatory region
The promoter is composed of a core promoter and proximal promoter elements and is typically less than 1 kb in length. Distal regulatory elements including enhancers, silencers, insulators, and locus control regions can be located up to several 100 kb from the promoter. These distal elements may interact with the core promoter or proximal promoter site by DNA looping [**36**].

1.3.2 General transcription factors

The focused core promoter is typically located -40 to +40 relative to the +1 transcription start site and serves as binding site for the RNA polymerase II machinery which initiates transcription. However, Pol II can synthesize RNA from DNA itself, but is not able to recognize the core promoter [**41**,**38**]. This event requires the assembly of a transcription preinitiation complex (PIC) which is additionally composed of general (or basic) transcriptions factors

(GTFs) including TFIIA (transcription factor for RNA polymerase II A), TFIIB, TFIID, TFIIE, TFIIF, and TFIIH. However, these factors interact differently depending on the core promoter. In case of the TATA box-driven core promoter the GTFs can assemble into a PIC in the order TFIID, TFIIB, RNA polymerase II-TFIIF complex, TFIIE, and TFIIH [**9**]. In contrast, the same factors cannot handle the transcription initiation from a DPE-dependent core promoter [**42**].

A key factor in the recognition process of focused core promoters is TFIID, a multisubunit complex composed of the TATA-binding protein (TBP) and 13 – 14 TBP-associated factors (TAFs). Several studies indicate that TFIID is required for initiation of transcription, but not any more once synthesis of mRNA is already in process [**43**].

Another essential factor for transcription is TFIIB as it plays a crucial role in the formation of the PIC. TFIIB recruits Pol II and provides the link between the DNA-bound TFIID and Pol II. Furthermore, TFIIB affects the catalytic activity of Pol II and seems to be a target for transcriptional activator proteins as well [**44**].

TFIIA has a stabilizing effect to the TFIID promoter complex as it may facilitate the binding of TBP to the TATA box. However, it is not essential for the assembly of the PIC. TFIIF is pre-bound to Pol II and can interact with several GTFs. Its function comprises mainly the prevention of an unspecific binding of Pol II. TFIIE and TFIIH bind to the core promoter at the very end completing the PIC assembly and contribute to the unwinding of DNA and the early steps of mRNA synthesis [**44**,**38**].

Beside these classical GTFs, the large multiprotein complex Mediator seems to play a crucial role in the activation and repression of Pol II transcription. Mediator is composed of more than 20 subunits and functions as adaptor that mediates transcriptional signals from DNA binding transcriptions factors bound at the proximal promoter or at distal regulatory elements to Pol II and the GTFs bound at the core promoter [**45**].

1.3.3 Core promoter elements

The DNA sequence of the core promoter comprises the site of transcription initiation and in the majority of cases just about 40 nucleotides. Several sequence motifs like the TATA box, Inr, BRE, or DPE (Figure 1-2) that are typically found in focused core promoters have been identified, however they are not universal. For instant, TATA boxes are not obligatory present in core promoters [**9**].

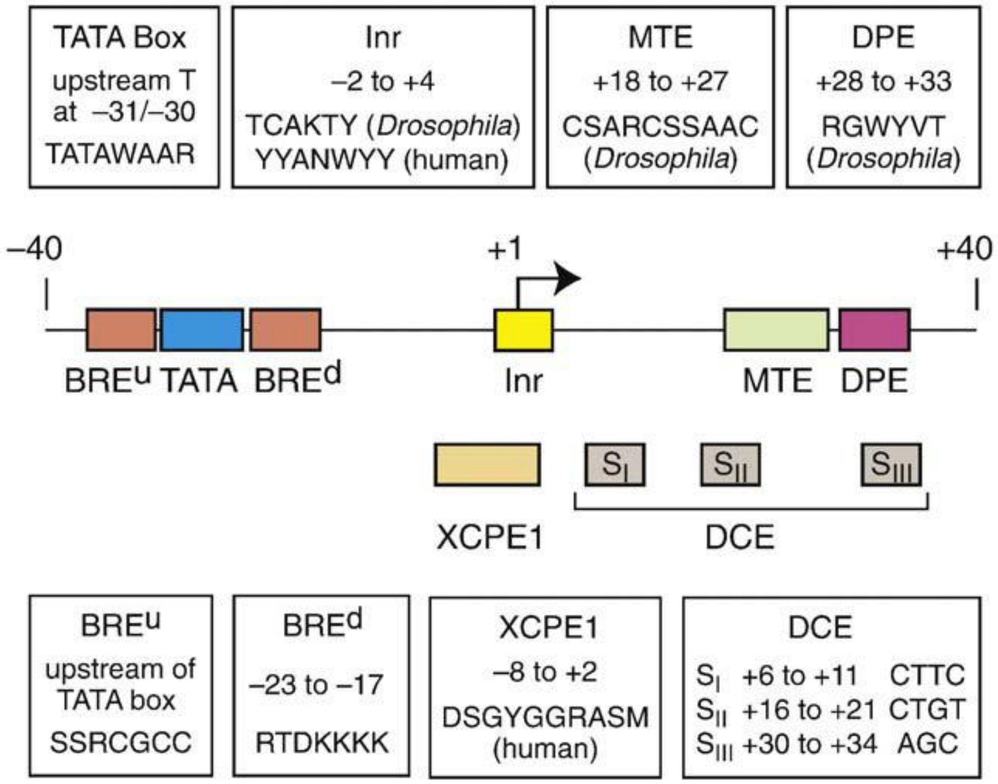


Figure 1-2: Common focused core promoter motifs for transcription by RNA polymerase II
Motifs typically found in focused core promoters inclusive consensus sequence and position relative to the transcription start site (A+1 in the Inr) are shown. However, there are no universal core promoter elements and it is likely that there are additional unknown core promoter motifs [**38**].

The **INITIATOR (INR)** encompasses the transcription start site and is probably the most common core promoter element [**46**,**10**,**47**]. The Inr consensus sequence was determined to be YYANWYY in human and TCAKTY in *Drosophila*, whereas the A nucleotide of the consensus sequence often represents the +1 start site. Several factors interact with the Inr motif, whereas the binding of TFIID seems to be of great relevance as the sequence binding specificity of TFIID, or more precisely of the TAF1 and TAF2 subunits, to Inr is identical to the Inr consensus sequence [**48**,**49**].

The **TATA BOX** was discovered in 1979 and was consequently the first eukaryotic promoter motif to be identified [**50**]. Its consensus sequence in metazoan was determined to be TATAWAAR, whereas the distance to the transcription start site is correlated with the tissue specificity of respective gene [**51**]. In general, the upstream T of the TATA box consensus sequence is located at position -31 or -30 relative to the A+1 in the Inr, which are the optimal positions for a high tissue-specific transcription. The general transcription factor associated with the TATA box is TFIID, whereas the TBP subunit specifically binds to the TATA box sequence. The TATA box is probably the most investigated core promoter element, however it is only present in about 10 – 15% of all mammalian core promoters [**39**,**52**].

The **TFIIB RECOGNITION ELEMENT (BRE)** was initially identified as a sequence element located immediately upstream of the TATA box, that represents the binding site for the general transcription factor TFIIB [**53**]. Later, an additional binding site for TFIIB was discovered binding immediately downstream of the TATA box [**54**]. The consensus sequence of the upstream BRE (BRE^u) was identified to be SSRCGCC [**53**], whereas the downstream BRE (BRE^d) has a consensus of RTDKKKK [**54**]. BRE^u as well as BRE^d enhance the formation of the TFIIB-TBP-promoter complex and consequently increase transcription activity. However, both can act on transcription in a negative manner as well, whereas the negative effect of BRE^d correlates with the present of a BRE^u [**44**]. Furthermore, BRE^u may also contribute to transcription regulation [**55**,**38**].

The **DOWNSTREAM CORE PROMOTER ELEMENT (DPE)** is located downstream of the transcription start site and acts as a TFIID binding site [**56**]. Its exact location is +28 to +33 relative to the A+1 in the Inr, having a consensus sequence of RGWYVT in *Drosophila* [**57**]. However, the *Drosophila* consensus was also found in mammalian core promoters showing DPE activity [**37**]. The precise distance between the Inr and DPE is very important for optimal transcriptional activity as TFIID binds cooperatively to both motifs [**56**,**57**]. The DPE has the same function as the TATA box, because both motifs act as recognition site for TFIID and are interchangeable to gain basal transcription activity. DPE-dependent promoters generally have just the DPE and the Inr motif and seem to be as common as TATA box-dependent core promoters at least in *Drosophila*. However, core promoters comprising TATA box, Inr, and DPE elements have also been identified [**57**,**38**].

The **MOTIF TEN ELEMENT (MTE)** was identified as a potential core promoter element via computational analysis revealing it as an overrepresented sequence motif in a vast number of *Drosophila* core promoters [**58**]. The MTE consensus is CSARCSSAAC and is located precisely from +18 to +27 relative to A+1 in the Inr and is conserved from *Drosophila* to human. The nucleotides from +18 to +22 are most crucial for transcriptional activity mediated by the MTE, though [**59**,**37**]. Like for DPE-depended promoters, the distance between the MTE and Inr is important for optimal transcription, as the MTE functions cooperatively with the Inr as well. The

MTE requires Inr for proper function, but can act independently of the DPE and the TATA box. Furthermore, addition of an MTE can compensate the loss of transcription due to mutation of the DPE or TATA box. In addition, strong synergism between the MTE and the TATA box as well as the DPE has been observed [**59**].

The **DOWNSTREAM CORE ELEMENT (DCE)** was initially discovered in the human β-globin promoter as a downstream core promoter motif which is different from the DPE [**60**]. Later studies revealed that the DCE element is present in a variety of core promoters especially such containing a TATA box [**61**]. The DCE comprises three subelements which are distinct from the DPE sequence each: S_I is CTTC, S_{II} is CTGT, and S_{III} is AGC. S_I is located from +6 to +11, S_{II} from +16 to +21, and S_{III} from +30 to +34 relative to the transcription start site. The DCE function requires the binding of the TAF subunits of the TFIID, whereas TAF1 interacts with the DCE in a sequence-specific manner.

The **X CORE PROMOTER ELEMENT 1 (XCPE1)** was identified and characterized based on analyses of the hepatitis B virus (HBV) X gene promoter [**62**]. However, XCPE1 is present in about 1% of human core promoters and most notably in TATA-less ones. XCEP1 has a consensus sequence of DSGYGGRASM and is located from -8 to +2 relative to the transcription start site. XCEP1 itself drives transcription at a very low level but can be activated by transcriptional activators like NFR1, NF1, or Sp1. Moreover, it was found that the majority of XCPE1-containing promoters contain Sp1-binding sites (GC boxes), NF1-binding sites (CAAT boxes), and NFR1-binding sites. The TBP is essential for the function of XCPE1, however TAF1 of TFIID is not required to drive transcription.

The **X CORE PROMOTER ELEMENT 2 (XCPE2)** was just recently discovered as a new core promoter element that drives the transcription from the second transcription start site of the HBV X gene [**63**]. XCPE2 has a consensus sequence of VCYCRTTRCMY and can also be found in human promoter region where it generally drives transcription from one of the start sites of TATA-less dispersed promoters having multiple TSSs. Like XCEP1, XCEP2 is located around the start site (-9 to +2) and is functionally similar to XCEP1 as well. However, XCEP2 can show a basal level of transcription by itself whereas XCEP1 requires the binding of activators. Transcription driven by XCPE2 needs at least Pol II, TFIIB, MED26-containing Mediator, and either TFIID or free TBP.

CpG islands (CGIs) are stretches of DNA that are rich of GC nucleotides and overrepresented in mainly unmethylated CpG dinucleotides [**64**,**65**]. In mammalian, CpG nucleotide pairs are generally chemically modified having a methyl group covalently attached to the C5 position of the cytosine ring and function to repress transcription epigenetically [**66**]. Whereas the non-methylated CpG islands, which are generally about 1 kb in length, are found in the promoter regions of approximately 60 – 70% of all human genes [**67**,**68**,**69**,**70**]. As mentioned previously, dispersed promoters, which can initiate transcription from multiple positions, are typically found in CpG islands. These motifs are present in the 5’ flanking region of most housekeeping genes, which are genes constitutively expressed in all tissues and cell types, as well as in many tissue-specific genes [**65**]. However, the majority of tissue-specific core promoters lack CpG islands as well as TATA boxes [**71**]. TATA boxes, DPE, or Inr motifs are usually not found in CpG islands, but multiple binding sites for the ubiquitous transcription factor Sp1 (GC boxes) are generally present [**65**,**10**]. Whereas GC boxes comprise the sequence GGGCGG or its reverse complement CCGCCC and are located upstream and downstream from the transcription start site. On the one hand contributes the binding of Sp1 to the protection of CpG islands regarding *de novo* methylation [**72**], on the other hand it is likely that Sp1 interacts with general transcription factors in order to initiate transcription [**9**]. Beside GC boxes, it is expected that binding sites for many ubiquitous transcription factors are present in CpG islands [**67**]. Another interesting feature that can be observed with several CpG island promoters is the bidirectional transcription, whereas the CpG island is located between two genes, which are arranged head-to-head [**73**].

1.3.4 Proximal promoter elements

The proximal promoter is defined as the regulatory region located directly upstream (about -250 to -30 relative to the +1 transcription start site) of the core promoter and typically contains multiple binding sites for activators and repressors [**74**,**75**,**36**]. GC boxes and CCAAT boxes, which are binding sites for sequence specific activator proteins, are commonly found in proximal promoter regions. GC boxes are binding sites for the transcription factor Sp1 as already described above. The CCAAT motif allows a specific interaction with several proteins like CCAAT-box-binding factor (CBF; also called nuclear factor Y, or NF-Y) which requires a highly conserved CCAAT sequence, CCAAT-binding transcription factor (CTF; also called nuclear factor I, or NF-I), or CCAAT-enhancer-binding protein (C/EBP) [**76**,**77**]. Some of these proximal promoter factors interact directly with the core transcriptional machinery and thus function most effectively in close proximity to the core promoter. However, these proximal proteins might also function as tethering elements that recruit distal enhancer complexes to the core transcription complex [**78**]. Sp1 bound to both distal enhancer and proximal promoter binding site can self-associate which enables the formation of DNA loops for transcriptional activation [**79**,**80**].

1.3.5 Distal regulatory elements

The *cis*-regulatory DNA of higher metazoans is highly structured and has a modular organization which comprises beside proximal and core promoter also more distant regulatory elements like enhancers, silencers, insulators, and locus control regions (Figure 1-3). The activity of a single transcription unit is exactly controlled through the cooperation of multiple enhancers, silencers, insulator, and promoters. Long-range regulation has not been observed in yeast and so it seems to be a common feature of genes that are involved in morphogenesis and are therefore subject to a stringent regulation [**74**].

ENHANCERS were initially identified as regions of the SV40 genome which could significantly increase the transcription of a heterologous expressed gene in cultured mammalian cells [**81**]. However, metazoan genes typically contain several enhancers themselves as well, which can be located in regulatory regions upstream and downstream of the gene and even within introns. They are generally involved in the spatial and temporal regulation of transcription and can function independently of their distance and orientation relative to the promoter. A typical enhancer is approximately 500 bp in length and is generally composed of a relative closely grouped cluster of about ten binding sites for at least three different specific transcription factors that work cooperatively to enhance transcription [**74**,**36**].

Enhancers are functionally quite similar to proximal promoter elements. Binding sites for some activator such as Sp1 are found in enhancers as well as in proximal promoter regions. However, unlike proximal promoter elements, enhancers typically act over long distances (Figure 1-3a) and can be located several hundred kilobase pairs of the promoter [**36**]. Recent studies indicate that DNA looping is the process by which enhancers function. This model suggests that distal enhancer regions can be brought in close proximity to the core promoter region by looping the intervening DNA [**82**].

As enhancers generally reside far from their target promoters, there need to be mechanisms that ensure a correct enhancer-promoter interaction, especially when an enhancer must activate only one of multiple promoters in its immediate vicinity. There are at least three mechanisms by which enhancer-promoter selectivity can be achieved. First, Insulators (also known as boundary elements) block undesired enhancer-promoter interaction. Second, there might be specific interactions between enhancer-binding proteins and transcription factors that interact with the core promoter and third, tethering elements that bind at the proximal promoter site can recruit distal enhancer [**83**,**74**].

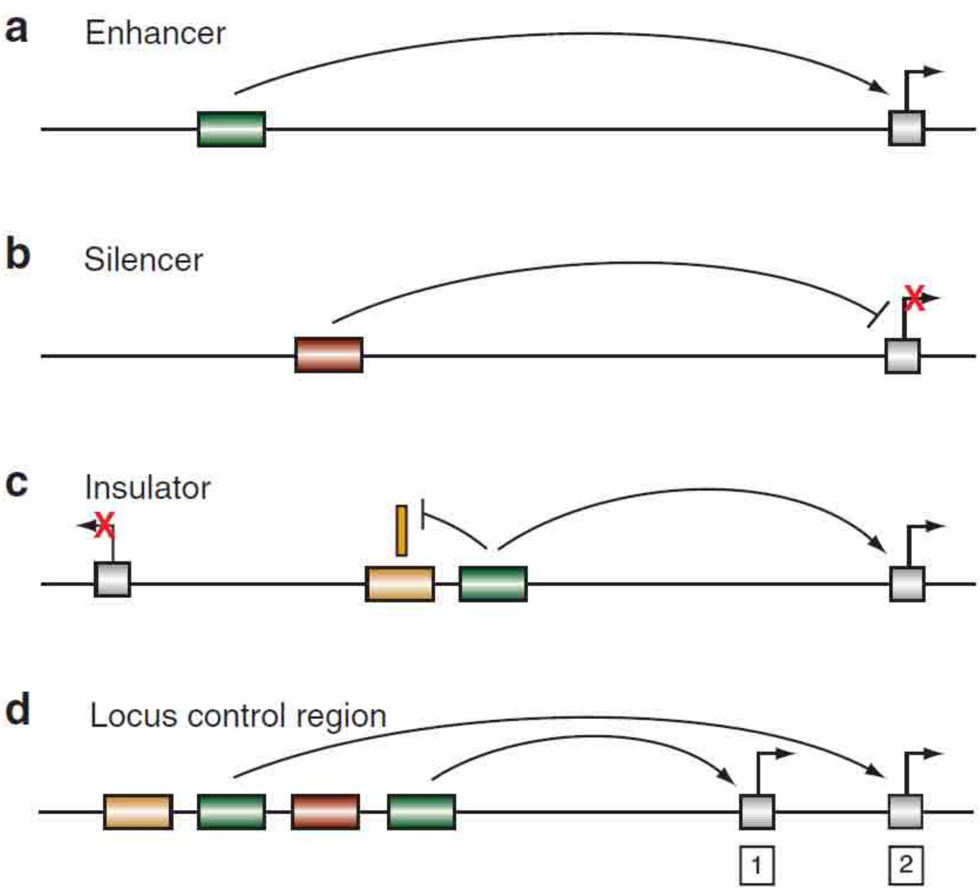


Figure 1-3: Distal transcriptional regulatory elements
(a, b): Enhancer and silencer typically act from a long distance to activate or repress transcription, respectively. (c) Insulators ensure a correct interaction of transcriptional regulatory elements by blocking genes from being affected. (d) Locus control regions are generally composed of multiple regulatory elements that work cooperatively to enable proper temporal- and/or spatial-specific gene expression to a cluster of nearby genes [**36**].

SILENCERS are genetic elements that can repress (or silence) transcription of a target gene (Figure 1-3b). Beside the negative regulatory effect, their properties are quite similar to enhancers [**84**]. Classical silencers are capable of repressing promoter activity in orientation- and position-independent manner. However, some position-dependent silencers have been identified as well. Silencers can be located in proximal promoters or in distal regulatory regions, whereas they can be part of an enhancer or function independently. Like enhancers, they can reside far from their target genes and even in introns, exons, and 3’ untranslated regions (3’ UTRs).

Silencers are binding sites for repressors which are negative transcription factors. Repressor function may rely on negative coregulators also named corepressors, which need to be recruited [**85**]. Many transcription factors show a dual functionality, having the potential to act as repressors or enhancers depending on promoter element [**84**].

Repressors might function by blocking the binding of an activator or by directly competing for the same binding site [**36**]. Furthermore, repressors may establish a repressive chromatin structure which sterically hinders the access of activators or GTFs to the promoter. However, another study suggested that silencing acts primarily by inhibiting the formation of the PIC [**86**].

INSULATORS (also known as boundary elements) are DNA sequence elements that block genes from being affected by the transcriptional activity of other genes (Figure 1-3c). Some insulators can block the interaction of promoters with distal enhancers associated with neighboring genes, others can function as barriers that can prevent the spread of inactive condensed chromatin (heterochromatin) [**87**]. Insulators are typically about 0.5 to 3 kb in length and act in a position-dependent and orientation-independent fashion [**36**]. They are often composed of clustered binding sites for large zinc finger proteins like Su(Hw) and CTCF [**88**,**87**]. Many enhancer blocking proteins have already been identified in *Drosophila* [**89**], however CTCF as the only known protein showing enhancer-blocking activity in vertebrates so far seems to play a crucial role in many different loci [**90**].

Insulator elements with enhancer blocking activity function by interference of the enhancer-promoter communication when located between the two. However, they have no or just little effect when positioned at either side [**90**]. The insulator might bind the enhancer associated activator and so prevent an interaction with its target promoter. The barrier activity against heterochromatin spreading may be explained by the recruitment of gene-activating factors or histone-modifying activities [**36**].

LOCUS CONTROL REGIONS (LCRS) consist of various regulatory elements that are involved in the regulation of an entire locus or gene cluster (Figure 1-3d). They are operationally defined as elements that tissue-specifically enhance gene expression to physiological levels from a distant in a position-independent and copy number-dependent fashion [**91**]. LCRs are generally composed of several gene regulatory elements such as enhancers, silencers, insulators, and matrix attachment regions (MARs) or scaffold attachment regions (SARs), which can be bound by transcription factors, co-regulators, and chromatin modifiers [**36**]. The cooperation of these factors functionally defines the proper spatial- and temporal-specific gene expression.

The LCR was initially identified in the human β-globin locus [**92**], but further LCRs have already been discovered in various mammalian loci [**91**]. They are typically positioned upstream of their target gene(s). However, they can also be located within an intron, downstream of the gene, or even in introns of a neighboring gene [**36**].

1.4 Strategies for identifying mammalian regulatory sequences

Beside its potential biotechnological value (see chapter 1.1), identifying of gene regulatory elements is of tremendous importance to study and understand transcriptional regulation as well as for improving genome annotation. Unlike coding sequences which can be identified and characterized quite easily by studying cDNAs and proteins, the recognition of *cis*-regulatory elements poses a great challenge. As experimental methods for identifying regulatory sequences are very labor-intensive, modern approaches are mainly focused on the computational analysis of large genomic data sets. However, although more than 56 complete sequences of eukaryotic genomes are already publicly available, the Chinese Hamster is not among them [**93**].

1.4.1 Experimental approaches

FUNCTIONAL ASSAYS (REPORTER GENE ASSAYS) are quite versatile methods for identifying and analyzing the activity of transcriptional regulatory elements. Today, they are mainly used for direct promoter studies or to verify transcriptional activity of already identified putative gene regulatory elements. However, functional assays are promising to be adapted for the use in genome-wide screens [**36**]. In these assays, the DNA element to be tested is cloned upstream of a reporter gene that allows an easy quantification, such as the chloramphenicol acetyltransferase (CAT), β-galactosidase, green fluorescent protein (GFP), or luciferase gene. The resulting construct is then transfected into cultured mammalian cells and the activity of the reporter is measured. Detected genomic sequences showing regulatory activity can then be truncated by serial deletions in order to more accurately locate the functional element(s). Depending on the configuration of the reporter vector construct, all types of gene regulatory elements like core promoter, proximal promoters, enhancers, silencers, insulators, or LCR can be tested [**36**].

Several drawbacks are related to functional assays for identifying gene regulatory elements and need to be considered when using these approaches. First, regulatory elements can be widely dispersed and so it can be possible that only a portion of a promoter element will be captured in the reporter construct that may not be sufficient to drive transcription. Second, the identified regulatory element may fail to show transcriptional activity due to differences in the chromatin structure between the reporter gene and the endogenous counterpart. Third, the regulatory element may only be used in a specific tissue, development stage, or physiological response pathway and thus may not be active under the culture conditions used in the reporter assay [**36**].

GENOMIC ANALYSIS OF TRANSCRIPTION FACTOR BINDING SITES (TFBS) is an approach on which several methods for identifying *cis*-regulatory sequences are based.

DNASE I HYPERSENSITIVE SITE MAPPING is a technique which enables the recognition of the precise location of many different regulatory elements. This method is based on the detection of genomic DNA regions in a relaxed chromatin structure which is more sensitive to DNase I digestion and can occur due to transcription factor binding. As this technique is very labor-intensive its application was limited to only a small number of genes. However, a more recently developed method allows the identification of DNase I hypersensitive sites on a genome-wide scale [**94**].

The **DNA FOOTPRINTING ASSAY** is an approach that allows the identification of DNA regions that are protected from digestion by DNase I because of the binding of proteins such as transcription factors [**95**].

The **GEL SHIFT ASSAY** is another technique to determine sequences that bind various transcription factors. This assay is gel-based and allows the detection of proteins that bind to a DNA fragment because of the reduced migration of the DNA [**95**].

In recent years, **CHROMATIN IMMUNOPRECIPITATION (CHIP)** has become a very popular method for the detection and identification of DNA sequences bound to a given protein. DNA-bound proteins are captured by chemical crosslinking and the genomic DNA is fragmented. An antibody that recognizes a specific transcription factor is used to isolate specific complexes. The enriched DNA population is then amplified, labeled, and hybridized to a DNA microarray. This combined technique known as **CHIP-CHIP** can provide a genome-wide view of protein-DNA interactions [**96**]. Alternatively, ChIP material can be used to construct a tag library which enables the analysis of the ChIP products by sequencing [**97**]. ChIP-chip experiments are currently limited by the coverage and availability of the microarray of the genome of interest. On the contrary, ChIP cloning is more labor-intensive.

TRANSCRIPT-BASED METHODS FOR IDENTIFYING PROMOTER ELEMENTS are targeted on locating the 5’ boundaries of a transcript. Promoters contain the TSS and so they always overlap with the first exon of a gene. This allows the determination of the promoter region by looking upstream of the first exon in genomic sequences [**98**]. These kinds of mRNA analyses require a reliable isolation of full-length cDNA. However, cDNAs are traditionally amplified starting at the 3’ end which results in truncated 5’ end in the majority of cases and so the promoter region may be several kb out of reach [**98**].

RAPID AMPLIFICATION OF CDNA ENDS (RACE) can be used to identify the 5’ ends of individual mRNAs [**99**]. In the first step a phosphatase treatment removes the phosphate groups from truncated and uncapped RNA molecules, whereas full-length mRNAs retain their 5’ cap structure. Then the cap is removed via tobacco acid pyrophosphatase, leaving a 5’ end phosphate group that allows the addition of an oligonucleotide adapter to the 5’ end. Transcript-specific primers are then used for a reverse transcription PCR (polymerase chain reaction). The products can be cloned and sequenced. RACE is useful for targeting a particular transcript of interest such as from a highly expressed gene. However, high-throughput screens are not feasible.

One genome-wide, high-throughput approach for TSSs discovery uses the 5’ ends of cloned full-length cDNA libraries (5’ ESTs) for sequencing [**100**].

CAP ANALYSIS OF GENE EXPRESSION (CAGE) is a further high-throughput technique that allows the identification of TSSs. This method is based on the preparation and sequencing of concatamers of short DNA tags derived from 5’ ends of capped mRNAs [**101**].

5’ END SERIAL ANALYSIS OF GENE EXPRESSION (5’ SAGE) is a similar technology for genome-wide, high-throughput analysis of TSSs [**102**]. This method combines 5’ RACE and the original SAGE [**103**], which generates concatamers of short tags derived from the 3’ end of transcripts, in order to locate their 5’ boundaries.

GENE IDENTIFICATION SIGNATURE (GIS) ANALYSIS is another strategy for the identification of TSSs in a high-throughput manner, in which 5’ and 3’ short ends are extracted to generate so-called paired-end ditags (PETs). These PETs are concatenated for efficient sequencing and mapping to the genome [**104**].

A drawback of all these DNA tag and sequencing strategies is the requirement of matching the 5’ ends to genomic DNA sequences in order to identify upstream located gene regulatory elements, as genome sequence data are currently not available for all organisms of interest. However, some studies indicated that promoter activity can also be found in 5’ UTRs of transcripts [**105**].

Other transcript-based methods rely on cDNA sequence information. Transcript-specific primers binding to exon 1 are used for PCR amplification of the 5’ flanking regions of respective gene from a library containing genomic DNA fragments of various length. PCR products can then be cloned and sequenced (see master thesis at hand).

A variant of this approach, which uses self-ligations of genomic DNA as template and Inverse PCR for the identification of 5’ flanking regions, is suggested in this master thesis (see chapter 5.7). Even though these methods are not suitable for genome-wide, high-throughput screens, they allow directly targeting of specific genes of interest such as highly expressed ones.

1.4.2 Computational approaches

PROMOTER PREDICTION PROGRAMS (PPPS) aim to identify promoter regions in genomic DNA sequences using computational models. Promoter regions can be distinguished from other parts of the genome because their properties are considerably different. Some features that has turned out to be useful for identifying promoters include CpG islands [**106**], typical transcription factor binding sites [**107**], and statistical properties of the core and proximal promoters [**108**]. Today, most successful PPPs search for these promoter-specific features by using machine learning techniques such as discriminant analysis, hidden markov models, and artificial neural networks in order to predict promoters [**109**]. However, these tools require large amounts of high-quality training data, preferably from experimentally verified core promoters. Furthermore, they are limited to find core promoters which are similar to already known ones. Recently, Abeel et al. presented the Easy Promoter Prediction Program (EP3), which uses GC content and large-scale structural features of DNA for promoter identification and requires no training [**109**]. Some selected, publicly available PPPs are listed in Table 1-1.

Table 1-1: Selected, publicly available promoter prediction programs (PPPs)

Name	URL	Reference
ARTS	http://www.fml.tuebingen.mpg.de/raetsch/projects/arts	[**110**]
CoreBoost	http://rulai.cshl.edu/tools/CoreBoost/	[**111**]
EP3	http://bioinformatics.psb.ugent.be/webtools/ep3/	[**109**]
Eponine	http://www.sanger.ac.uk/resources/software/eponine/	[**108**]
FirstEF	http://rulai.cshl.org/tools/FirstEF/	[**112**]
McPromoter	http://tools.igsp.duke.edu/generegulation/McPromoter/	[**113**]
NNPP	http://www.fruitfly.org/seq_tools/promoter.html	[**114**]
Promoter 2.0	http://www.cbs.dtu.dk/services/Promoter/	[**115**]
ProSOM	http://bioinformatics.psb.ugent.be/software/details/ProSOM	[**116**]

Transcription factor binding site (TFBS) prediction programs are based on the comparison with known TFBSs, which have been experimentally identified from other gene regulatory sites. Experimental data of most well-characterized TFBSs have been used to develop databases such as TRANSFAC® [**117**] or more recent JASPAR [**118**]. From a collection of binding sites, position weight matrices (PWM) were then derived for each factor. Web-based software tools like MatInspector [**119**] or Match™ [**120**] screen DNA sequences with all the matrices in the database and return a list of potential TFBSs based on a statistical algorithm. However, a significant number of predicted sites are likely false positive using these methods. Another drawback is that the underlying databases are not complete as most likely not all TBPSs have been identified and implemented.

An alternative popular analysis technique involves the use of phylogenetic footprinting for the discovery of regulatory elements. The idea behind phylogenetic footprinting is that due to selective pressure, functional regulatory elements evolve slower than other DNA sequences. So the most highly conserved motifs in a collection of orthologous regulation regions should be the best candidates as regulatory elements [**121**]. However, studies have shown that not nearly all TFBSs are conserved among species [**122**]. An example for a publicly available web-based tool performing such analyses is FootPrinter2 [**123**].

Today, the most commonly used programs for TFBSs identification such as rVista [**124**], ConSite [**125**], and FootPrinter3 [**126**] combine matrix-based site prediction with phylogenetic footprinting. Some selected, publicly available TFBSs prediction programs are listed in Table 1-2.

Table 1-2: Selected, publicly available TFBS prediction programs

Name	**URL**	**Reference**
ConSite	http://www.phylofoot.org/consite	[**125**]
FootPrinter3	http://bio.cs.washington.edu/software.html	[**126**]
Match™	http://www.gene-regulation.com/pub/programs.html#match	[**120**]
rVista	http://rvista.dcode.org/	[**124**]

Ab initio identification of gene regulatory elements by computational approaches requires whole genome sequence data, which are currently not available for all organisms of interest including the Chinese hamster. However, these methods can be useful for analyzes of experimentally identified potential gene regulatory regions of those organisms. The web-based software tools NNPP (neural network promoter prediction program) and ConSite were used in this study for recognizing TSSs and TFBSs of experimentally discovered putative transcriptional regulatory regions, respectively.

NNPP is a PPP based on an artificial neural network model using a time-delayed network architecture which has one feature layer for the TATA box and another for the Inr. This neural network can detect the TATA box and the Inr and is insensitive to their relative spacing [**114**]. However, it has to be considered that NNPP is limited to these types of promoters. A recent promoter prediction evaluation study even suggested that NNPP is not suited to identify promoters [**127**].

ConSite is an online tool for the *in silico* prediction of *cis*-regulatory elements in a genomic DNA sequence. It is based on the integration of TFBSs prediction via binding profile models and phylogenetic footprinting. ConSite uses the ORCA alignment program and the JASPAR database for this purpose [**125**].

2 Objectives

The ultimate ambition of the master thesis at hand was the identification of CHO endogenous gene regulatory elements, which are able to regulate the transcription of foreign genes in CHO cells without leading to undesired side effects.

In this thesis, I sought to continue the research which was previously conducted by Martina Baumann [**128**]. In the first part, a genomic CHO library constructed by Martina Baumann was used to identify the 5' flanking region of known CHO genes via nested PCR amplification. The availability of cDNA sequence information of highly abundant CHO genes enabled this approach. The major aims were:

- Identifying genomic DNA fragments located upstream of highly expressed CHO genes.
- Testing obtained DNA sequences for potential promoter activity.

The second part comprised the discovery of the 5' and 3' flanking regions of known CHO genomic DNA sequences using Inverse PCR. The challenging goals were:

- Establishing and optimizing the Inverse PCR approach for mammalian genomic DNA.
- Applying the Inverse PCR approach to discover the 5' and 3' flanking regions of obtained promoter candidates in order to potentially increase their regulatory activity.

The aim of the third part was the characterization of obtained fragments showing gene regulatory activity. The main purposes were:

- Analyzing the experimentally derived promoter candidates using *in silico* tools in order to predict putative transcription start sites and transcription factor binding sites.
- Analyzing the functional activity of discovered gene regulatory elements by generating and testing of truncation mutants.

3 Materials and methods

3.1 Equipment

All equipment used in this study is listed in Table 3-1.

Table 3-1: Equipment

Device	Description	Producer
Balance	PM460	Mettler-Toledo, USA
Balance	SM1220	Mettler-Toledo, USA
Balance	MC1 Laboratory LC6200	Sartorius, Germany
Centrifuge	Centrifuge 5414 D	Eppendorf, Germany
Centrifuge	Centrifuge 5415 R	Eppendorf, Germany
Centrifuge	Heraeus FRESCO 17 Centrifuge	Thermo Scientific, USA
Centrifuge	Jouan C312	Jouan, France
Centrifuge	Avanti™ J-20 XP	Beckmann Coulter, USA
Centrifuge rotor	JLA-10.500	Beckmann Coulter, USA
Concentrator	Savant ISS110 SpeedVac® Concentrator	Thermo Scientific, USA
Coulter counter	Multisizer™ 3 COULTER COUNTER®	Beckmann Coulter, USA
Electroporator	MicroPulser™	Bio-Rad, USA
Electroporator	Nucleofector® Device	Lonza, Switzerland
Hand dispenser	Eppendorf hand dispenser	Eppendorf, Germany
Incubation shaker	Infors	Infors, Switzerland
Incubator	BNA-311	Espec, Japan
Laminar flow hood	Herasafe®	Heraeus, Germany
Laminar flow hood	HBB 2448	Holten LaminAir, Denmark
Microplate reader	Infinite® M1000	Tecan, Switzerland
Microplate reader	Synergy™ 2 Multi-Mode Microplate Reader	BioTek, USA
Mixer	Mixer 5432	Eppendorf, Germany
Molecular Imager	Gel Doc™ XR System	Bio-Rad, USA
Pipettes	Pipetman® 2µl, 10µl, 20µl, 100µl, 200µl, 1000µl	Gilson, USA
Pipette	CellMateII®	Matrix, USA
Power supply	PS 250	Hybaid, USA
Power supply	EBS-300 II	CBS Scientific, USA
Power supply	E 802	Consort, Belgium

Table 3-1: Equipment (continued)

Device	Description	Producer
Spectrophotometer	NanoDrop 1000 Spectrophotometer	Thermo Scientific, USA
Spectrophotometer	NanoPhotometer™	Implen, Germany
Thermoblock	Thermomixer compact	Eppendorf, Germany
Thermoblock	ThermoStat plus	Eppendorf, Germany
Thermocycler	T3 Thermocycler	Biometra, Germany
Thermocycler	TProfesional Thermocycler	Biometra, Germany
Thermocycler	C1000™ Thermal Cycler	Bio-Rad, USA
Transilluminator	TPB-M/WL	Vilber Lourmat, France
Vortex mixer	Vortex-Genie 2	Scientific Industries, USA

3.2 Laboratory consumables

All consumables used in this study are listed in Table 3-2.

Table 3-2: Laboratory consumables

Device	Description	Producer
6 well plates	Cell Culture Multiwell Plate, 6 well	Greiner Bio-One, Austria
96 well plates	Costar® Assay Plate 96 Well Flat Bottom	Corning, USA
Centrifugation tubes	15 ml, 50 ml	VWR, Austria
Cryotubes	CryoTube™ vials	Nalge Nunc, USA
Dispensers	Combitips 0.5 ml, 1.25 ml	Eppendorf, Germany
Steril filters	Filters 0.22 µl	Millipore, USA
Microtubes	Screw Cap Micro Tubes 1.5 ml, 2 ml	Sarstedt, Germany
Microtubes	Plastibrand® microcentrifuge tubes	Brand, Germany
Petri dishes	Petri dish 94 × 16 mm	Greiner Bio-One, Austria
Roux flasks	Cell Culture Flask, 50 ml, 250 ml, 550 ml	Greiner Bio-One, Austria
PCR tubes	Micro Tube Strips, 0.2ml	Biotix, USA
Pipettes	Costar® 5 ml, 10 ml and 25 ml Stripette® Serological Pipets	Corning, USA
Pipette tips	0.1 – 10 µl, 2 – 200 µl, 100 – 1000 µl	VWR, Austria
Cuvettes	PLASTIBRAND® UV-Cuvettes micro 2 mm gap	Brand, Germany

3.3 Molecular biology reagents and kits

All molecular biology reagents and kits used in this study are listed in Table 3-3.

Table 3-3: Molecular biology reagents and kits

Item	Description	Producer
DNA markers	2-log DNA Ladder, 1 kb DNA Ladder	New England Biolabs, USA
DNA markers	Lambda DNA/EcoRI+HindIII Marker, 3 FastRuler™ DNA Ladder, Low Range	Fermentas, Canada
DNA polymerase	Biotools DNA Polymerase	Biotools B&M Labs, Spain
DNA polymerase	Phusion® High-Fidelity DNA Polymerase	Finnzymes, Finland
dNTPs	Deoxynucleotide Solution Mix	New England Biolabs, USA
Gel-extraction and PCR clean-up kit	NucleoSpin® Extrakt II	Macherey-Nagel, Germany
Gel-extraction and PCR clean-up kit	illustra™ GFX™ PCR DNA and Gel Band Purification Kit	GE Healthcare, USA
Gel-extraction and PCR clean-up kit	Wizard® SV Gel and PCR Clean-Up System	Promega, USA
Genomic DNA purification kit	DNeasy® Blood & Tissue Kit	QIAGEN, Germany
Kinase	T4 Polynucleotide Kinase	New England Biolabs, USA
Ligase	T4 DNA Ligase	New England Biolabs, USA
Ligase buffer	T4 DNA Ligase Reaction Buffer	New England Biolabs, USA
Loading dye	6× DNA Loading Dye	Fermentas, Canada
Luciferase assay system	Dual-Glo® Luciferase Assay System	Promega, USA
Midi-prep kit	PureYield™ Plasmid MidiPrep System	Promega, USA
Midi-prep kit	Nucleobond® Xtra Midi Plus EF	Macherey-Nagel, Germany
Mini-prep kit	Wizard®Plus SV Minipreps DNA Purification System	Promega, USA
Phosphatase	Antarctic Phosphatase	New England Biolabs, USA
Phosphatase buffer	Antarctic Phosphatase Reaction Buffer	New England Biolabs, USA
Primer and oligonucleotides	Custom DNA Oligos	Sigma-Aldrich, USA
Restriction buffers and supplements	NEBuffer 1, NEBuffer 2, NEBuffer 3, NEBuffer 4, NEBuffer EcoRI, BSA	New England Biolabs, USA
Restriction endonucleases	Various types	New England Biolabs, USA
Transfection kit	Amaxa® Cell Line Nucleofector® Kit V	Lonza, Switzerland

3.4 Chemical reagents

All chemicals used in this study are listed in Table 3-4.

Table 3-4: Chemicals

Item	**Producer**
Agarose peqGOLD Universal	PEQLAB, Germany
Cell Culture Water EP-Grade	PAA, Austria
Ethanol ≥ 99.5% pro analysis	Merck, Germany
Ethidium bromide 1%	Carl Roth, Germany
Glycerol ROTIPURAN®, ≥99.5 %, p.a., anhydrous	Carl Roth, Germany
Isopropanol 99,5+%	Sigma-Aldrich, USA

3.5 Growth media

All salts, chemicals, and media ingredients were purchased either from AppliChem (Germany), Merck (Germany), Carl Roth (Germany), or Sigma-Aldrich (USA).

3.5.1 Bacterial growth media

LURIA-BERTANI (LB) MEDIUM

Table 3-5: Composition of LB medium

Concentration	Components
10 g l^{-1}	Peptone from casein
5 g l^{-1}	Yeast extract
10 g l^{-1}	NaCl

The components listed in Table 3-5 were dissolved in double distilled water (ddH_2O), portioned into 500 ml flasks and autoclaved at 121°C for 20 min. The sterile media was stored at 16°C and required antibiotics were added before usage. Stock solutions of antibiotics were diluted thousand fold to obtain working concentrations. Ampicillin (100 mg ml^{-1}) and kanamycin (50 mg ml^{-1}) stock solutions were prepared by dissolving the antibiotics in ddH_2O, followed by sterile filtration using 0.22 µm filters and stored in 500 µl aliquots at -20°C.

Luria-Bertani (LB) agar (1.5% w/v)

Table 3-6: Composition of LB agar (1.5% w/v)

Concentration	Components
10 g l^{-1}	Peptone from casein
5 g l^{-1}	Yeast extract
10 g l^{-1}	NaCl
7.5 g l^{-1}	Agar agar

LB agar was prepared same as LB media except for dissolving 7.5 g agar-agar in 500 ml LB medium before sterilization. The sterile agar was stored at 16°C and melted in a microwave oven before preparing agar plates. Required antibiotics were added to the cooled down but still melted agar (55°C) just before pouring the agar plates using 20 ml agar for each 90 mm petri dish. Agar plates were stored at 4°C.

Super Optimal Catabolite (SOC) Medium

Table 3-7: Composition of SOC medium

Concentration	Components
10 g l^{-1}	Peptone from casein
5 g l^{-1}	Yeast extract
0.5844 g l^{-1}	NaCl
0.2237 g l^{-1}	KCl
2.0330 g l^{-1}	$MgCl_2 \cdot 6 H_2O$
2.4648 g l^{-1}	$MgSO_4 \cdot 7 H_2O$
3.6032 g l^{-1}	Glucose

All component listed in Table 3-7 except glucose were dissolved in ddH_2O and autoclaved at 121°C for 20 min. Glucose was sterilized separately in order to prevent undesired Maillard reactions and added to the rest of the medium before portioning to 10 ml aliquots. SOC medium was stored at 4°C.

3.5.2 Cell culture media

CHO DHFR$^-$ GROWTH MEDIUM

CHO dihydrofolat reductase deficient (dhfr$^-$) suspension cells were cultivated in Dulbecco's Modified Eagle Medium (DMEM) and Ham's F-12 medium mixed in a 1:1 ratio (Invitrogen, USA). The medium was supplemented with 4 mM L-glutamine (Sigma Aldrich, USA), 0.25% (w/v) soya peptone (HyPep 1510, Sheffield Pharma, UK), 0.1% (w/v) Pluronic F68, and a protein free supplement (by courtesy of Polymun, Austria) and added with 1× HT supplement (Biochrom AG, Germany) leading to a final concentration of 100 µM hypoxanthine and 16 µM thymidine.

3.6 Solutions

3.6.1 Cell culture solutions

PHOSPHATE BUFFERED SALINE (PBS) BUFFER

Table 3-8: Composition of PBS 10×

Concentration	Components
10 g l^{-1}	KH_2PO_4
12 g l^{-1}	$Na_2HPO_4 \cdot 2\ H_2O$
2 g l^{-1}	KCl
80 g l^{-1}	NaCl

PBS (pH 7.4) was used as washing solution for CHO dhfr$^-$ cells.

3.6.2 Solutions for agarose gel electrophoresis

50× TRIS-ACETATE-EDTA (TAE) BUFFER

Table 3-9: Composition of 50× TAE buffer

Concentration	Components
242 g l^{-1}	Tris(hydroxymethyl)-aminomethan (TRIS)
57.1 ml l^{-1}	Glacial acetic acid
100 ml l^{-1}	0.5 M ethylenediaminetetraacetic acid (EDTA) pH 8

50× TAE buffer was used for the preparation of agarose gels and as running buffer for agarose gel electrophoresis.

TAE RUNNING BUFFER

Table 3-10: Composition of TAE running buffer

Concentration	Components
20 ml l^{-1}	50× TAE buffer
30 µl l^{-1}	Ethidium bromide

6× GEL LOADING BUFFER (BX BUFFER)

Table 3-11: Composition of 6× gel loading buffer

Quantity (w/v)	Components
0.25%	Bromophenol blue
0.25%	Xylene cyanol FF
30%	Glycerol in water

3.7 Strains and cell lines

3.7.1 Bacterial strains

- *Escherichia coli* (*E. coli*) strains DH5α, DH10B, JM109, and NEB10β (lab stocks) were used for cloning purpose.
- Electrocompetent *E. coli* MegaX DH10B (Invitrogen, USA) were used for library construction.

3.7.2 Mammalian cell line

- CHO dhfr$^-$ cell line (American Type Culture Collection, ATCC, USA)

3.8 Plasmids

3.8.1 pGL3 luciferase reporter vectors

The pGL3 luciferase reporter vectors (Promega, USA) were used for the quantitative analysis of fragments that potentially regulate CHO gene expression. The backbone of these vectors contains a modified coding region for firefly (*Photinus pyralis*) luciferase that has been optimized for monitoring transcriptional activity in transfected eukaryotic cells. The vectors contain the ampicillin resistance gene for selection in *E. coli*.

pGL3-Basic vector

The pGL3-Basic vector lacks eukaryotic promoter and enhancer elements. Any expression of luciferase activity in transfected cells depends on the inserted DNA fragment into the MCS (multiple cloning site) upstream from the firefly luciferase gene. Beside for testing transcriptional activity for obtained promoter candidates cloned into the vector, the initial pGL3-Basic vector was used as negative control.

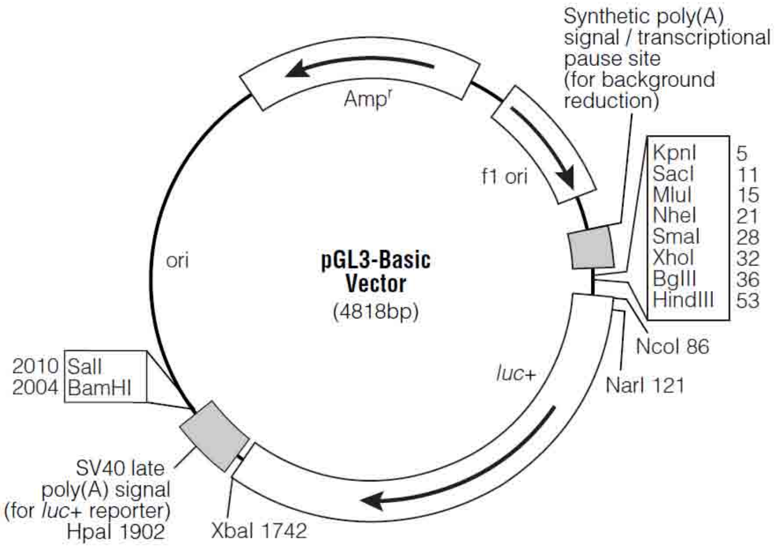


Figure 3-1: pGL3-Basic vector circle map
luc+: cDNA encoding the modified firefly luciferase; Amp^r: ampicillin resistance gene; f1 ori: origin of replication derived from filamentous phage; ori: origin of replication in *E. coli*; Expression of firefly luciferase depends on insertion of a functional promoter upstream from *luc+* [**129**].

pGL3-Promoter vector

The pGL3-Promoter vector contains the SV40 promoter upstream of the luciferase gene. The vector was used as a positive control.

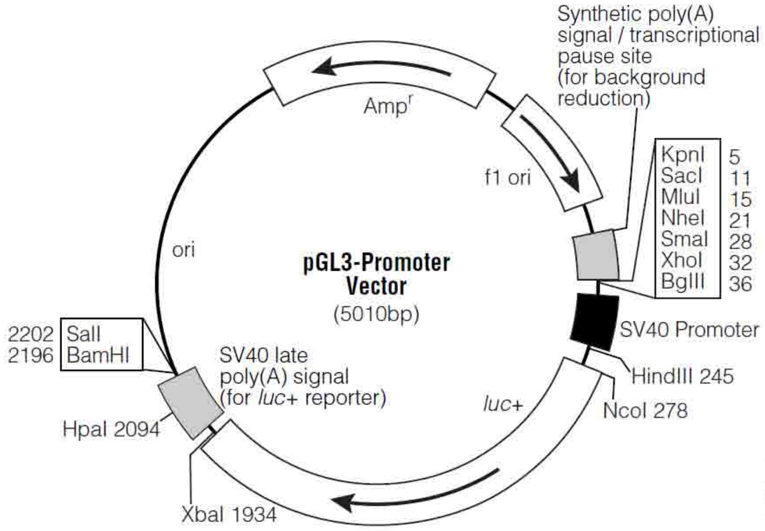


Figure 3-2: pGL3-Promoter vector circle map
luc+: cDNA encoding the modified firefly luciferase; Amp^r: ampicillin resistance gene; f1 ori: origin of replication derived from filamentous phage; ori: origin of replication in *E. coli*; The expression of the firefly luciferase is driven by a SV40 promoter [**129**].

3.8.2 pRL-SV40 vector

The pRL-SV40 vector (Promega, USA) is an internal control reporter and was used in combination with the pGL3 luciferase reporter vectors to co-transfect CHO cells. The pRL-SV40 vector contains a cDNA encoding the *Renilla* luciferase, which was originally cloned from the marine organism *Renilla reniformis*. Furthermore, the vector contains the SV40 enhancer and early promoter elements driving the *Renilla* luciferase gene as well as the ampicillin resistance gene for selection in *E. coli*.

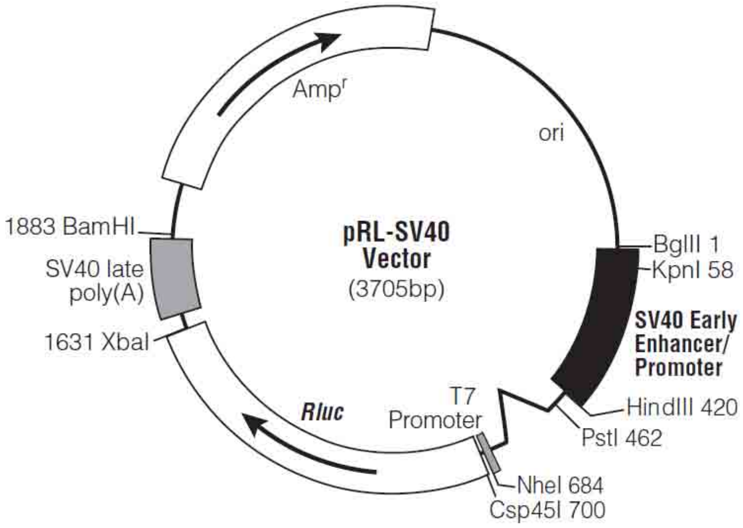


Figure 3-3: pRL-SV40 vector circle map
Rluc: cDNA encoding the *Renilla* luciferase; Amp^r: ampicillin resistance gene; ori: origin of replication in *E. coli*; Expression of *Renilla* is driven by a SV40 early enhancer/promoter [**130**].

3.9 Molecular biology methods

All cloning procedures were performed as described by Sambrook and Russel [**131**] or as recommended by the manufacturer of used kits. Sterile double distilled water (ddH_2O) was used for all reactions and preparations of buffers and solutions.

3.9.1 Polymerase chain reaction (PCR)

Currently, polymerase chain reaction (PCR) is the most important method used in molecular biology. This technique facilitates the enzymatic *in vitro* amplification of specific DNA regions. The principle of PCR is based on the cyclic repetition of three steps.

- **DENATURATION:** The double-stranded DNA template is heated to 92 – 98°C depending on the DNA polymerase used. This causes separation of DNA template by disrupting the hydrogen bonds between complementary bases.
- **PRIMER ANNEALING:** The temperature is reduced allowing the annealing of the two specific primers to the single-stranded DNA template (50 – 69°C). The annealing temperature primarily depends on the melting temperature (T_m) of the primer and should be similar for the pair of primers used.
- **ELONGATION (EXTENSION):** The temperature is increased to the optimum activity temperature of the heat-stable DNA polymerase used (commonly 72°C). The polymerase synthesizes a new DNA strand that is complementary to the DNA template strand by adding deoxynucleoside triphosphates (dNTPs) in 5’ to 3’ direction.

All PCR experiments were conducted using a Biometra® T3 Thermocycler, a Biometra® TProfessional Thermocycler, or a BIO-RAD C1000™ Thermal Cycler and 0.2 ml reaction tubes.

COMMON PCR

The Phusion® High-Fidelity DNA polymerase was used for cloning purposes requiring blunt end fragments. All PCR runs were performed in 50 µl reaction volumes. The composition of one PCR reaction is listed in Table 3-12. Phusion® DNA polymerase was always the last item added.

Table 3-12: Pipetting instruction for Phusion® DNA polymerase

Component	Volume / 50 µl	Final concentration
ddH_2O	add to 50 µl	
5× Phusion® HF buffer	10 µl	1×
dNTPs (10mM)	1 µl	200 µM each
Primer sense (10 µM)	2.5 µl	0.5 µM
Primer antisense (10 µM)	2.5 µl	0.5 µM
Template DNA*	x µl	
Phusion® DNA polymerase (2 U $µl^{-1}$)	0.5 µl	0.02 U $µl^{-1}$

* For low complexity DNA (e.g. plasmid) 1 pg – 10 ng per 50 µl reaction volume was used and for high complexity genomic DNA the amount was 50 – 250 ng per 50 µl reaction volume.

The cycling conditions for the generally performed 3-step protocol are listed in Table 3-13. A 2-step protocol was used when primer T_m values were at least 69°C. In the 2-step protocol the combined annealing/extension step was performed at 72°C.

The primer T_m values were calculated using Finnzymes` web-based T_m calculator (http://www.finnzymes.com/tm_determination.html).

Table 3-13: Cycling instruction for Phusion® DNA polymerase (3-step protocol)

Cycle step	Temperature	Time	Number of cycles
Initial denaturation	98°C	30 s	1
Denaturation	98°C	7 s	30 – 35
Annealing	T_m + 3°C	20 s	
Extension	72°C	15 – 30 s/kb*	
Final extension	72°C	5 min	1

* For low complexity DNA (e.g. plasmid) an extension time of 15 seconds per 1 kb was used and for high complexity genomic DNA 30 seconds per 1 kb.

Colony PCR

Colony PCR was used to screen for positive transformants by directly amplifying specific DNA regions out of bacterial cells. This technique enables the verification of the presence of the desired plasmid sequence in bacterial clones. For PCR screening the Biotools DNA polymerase was used. The composition of a master mix for eight PCR reactions is listed in Table 3-14.

For PCR screening a single colony from a transformation plate was picked with a sterile pipette tip and whisked in 30 µl of the master mix. The same tip was used to streak the remaining bacterial cells on an LB agar master plate, which was necessary for subsequent amplification of positive clones.

Table 3-14: Pipetting instruction for Biotools DNA polymerase (500 µl master mix)

Component	**Volume / 500µl**	**Final concentration**
ddH$_2$O	412.5 µl	
10× MgCl$_2$ free buffer	50 µl	1×
MgCl$_2$ solution (50 mM)	20 µl	2 mM
dNTPs (10 mM)	5 µl	100 µM each
Primer sense (10 µM)	5 µl	0.1 µM
Primer antisense (10 µM)	5 µl	0.1 µM
Biotools DNA polymerase (5 U µl^{-1})	2.5 µl	0.025 U µl^{-1}

The cycling program is listed in Table 3-15.

Table 3-15: Cycling instruction for Biotools DNA polymerase

Cycle step	**Temperature**	**Time**	**Number of cycles**
Initial denaturation	95°C	2 min	1
Denaturation	94°C	30 s	30
Annealing	T_m - 5°C	30 s	
Extension	72°C	1 min/kb	
Final extension	72°C	5 min	1

NESTED PCR

Nested PCR was used to increase the specificity of PCR amplification by reducing background due to unspecific amplification products. This allows the amplification of least amounts of template DNA. For this purpose, two sets of primers were used in two successive PCR runs. The products of the first PCR were used to conduct a second PCR with a pair of primers located between the pair of primers used in the first run. For amplification, the Phusion® High-Fidelity DNA polymerase was used as already described above.

PCR PRIMERS

Primers were generally designed using the online program Primer3-web 0.4.0 (http://frodo.wi.mit.edu/primer3/input.htm) [**132**]. All primers used in this study are listed in Table 3-17 to Table 3-21.

Table 3-16: Primers used for PCR screening (Colony PCR)

Designation	**Sequence 5’ → 3’**
pGL3_Luc_screen_antisense	AGGAACCAGGGCGTATCTCT
pGL3_screening_sense	CAAAATAGGCTGTCCCCAGT

Table 3-17: Primers used for identifying 5’ regions of known CHO genes

Designation	**Sequence 5’ → 3’**
Jund_C_sense	TGTTTTGGCTTTTGAGGGTCTTGACTTTCTCCTCC
Jund_D_sense	CTTCCAAAAGGCAAAAAGGAAAAAGAAAAAGGCAGAGC
Rpl6 anti_NEW	CTTGGTTTTTGCAGCTGAGTA
Rpl6 nes_NEW	TTTTTAACCTTAGGGTCACCC
Rpl27 anti_NEW	CTTCACGATGACGGCTTTGC
Rpl27 nes_NEW	CTTTCCCGGGTTTCATGAACTT
Rpl35 anti_NEW	TCACCGCTGAAGCCTCCT
Rpl35 nes_NEW	CAATAGCGTCCCGGCTTG
Rplp1 anti_NEW	GACTGTCACCTCGTCGTCGT
Rplp1 nes_NEW	TTAGCTCCCTCGGAAGAACC
Rps6 anti_NEW	CTTCCACTCGTCACCCAGAG
Rps6 nes_NEW	TCCACTTCGATGAGTTTCTGG
Rps8 anti_NEW	GGTAGGGCTTTCTCTTACCCC
Rps8 nes_NEW	GTGCCAGTTGTCCCGAGAGAT
Tpt1 anti_NEW	TCCCGGTAGATGATCATGGTG
Tpt1 nes_NEW	GGAAAAGGCCGACTCGGG
pMACS left aussen	ACATTTCCCCGAAAAGTGC
pMACS left nested	TCGTCTTCAAGAATTGGTCG
hCD4 aussen	GCACCACTTTCTTTCCCTGA
hCD4 nested	ACAGAAATGGCAGGGCTCT

Table 3-17: Primers used for identifying 5’ regions of known CHO genes (continued)

Designation	Sequence 5’ → 3’
pMACS_Library_long_sense	AGAAACCATTATTATCATGACATTAACCTATAAAAATAGGCGTATCACG
pMACS_Library_long_AS	AGTAGGGACCTGAGCCCACAGAAATGGCAG
pVITRO_upstream_S	GCATCAGAGCAGATTGTACTGAGAG
pVITRO_upstream_S_nested	GCAAGGCGATTAAGTTGGGTA
hVITRO_Lib_AS	ATGGACAGTGGCATTGTTTTTC
hVITRO_Lib_AS_nested	GGTAGTAAGAGCAGAGCTCGTCAC
pVITRO_Library_long_sense	GATTAAGTTGGGTAACGCCAGGGTTTTCCCAGTC
pVITRO_Library_long_AS	TAGTAAGAGCAGAGCTCGTCACACTGGCACTTCTTGTCC

Table 3-18: Primers used for Inverse PCR

Designation	Sequence 5’ → 3’
CHO_DNA_1_sense	CAGGAAAGAGTAATTCCCAGAACAGT
CHO_DNA_1_sense_NESTED	GTAGACACTCAGAGAGACAGATGAACCT
CHO_DNA_1_ANTIsense	TAGTGGCAATCCTCTGACAAGATAAAG
CHO_DNA_1_ANTIsense_NESTED	TACAGGGACCAGAACAAATACAAAGAG
CHO_DNA_2_sense	CAGGCTGACCTCAAGCTTACTATTTTA
CHO_DNA_2_sense_NESTED	TAGCGTATACCTTTCATCCTAGCACTC
CHO_DNA_2_ANTIsense	CTTCTCAACCTTTCTAATGCTATGACC
CHO_DNA_2_ANTIsense_NESTED	CAAGGCTAGACTGGATACCTCATTAAG
Rpl6_A6_sense_2	CTGTGATAGGTACAGATGTGGGTGTT
Rpl6_A6_ANTIsense_2	ACCACAGTGACTCTCACTTCTAGCAT
Rpl6_A6_sense_NESTED_2	AAGGAAGCAGATGGCTCACTTGTA
Rpl6_A6_ANTIsense_NESTED_2	CTCAAACTGCCCTTATCCAGTGTC
Rps6_D1_S	CTGTTTTTATTGACAGGCTTGGACT
Rps6_D1_AS	ATTCTTGAGCTGTGTGCTTCCTTAG
Rps6_D1_S_Nested	TTGTTTACTGTGCATGTCATTTCCT
Rps6_D1_AS_Nested	TGTTTCTGATTAAAATCCCTTGCAT
Rps8_G15_S	TCTGTCTCTGGACCTAGGAGCTTTA
Rps8_G15_AS	AAAGCCTAAACTCCATTCCCTCTC
Rps8_G15_S_Nested	ATCTGTGGGAGTAGCTTAAGTGTGC
Rps8_G15_AS_Nested	ACTATCTCAGCCAGCCCACTACAC

Table 3-19: Primers used for CHO genomic PCR

Designation	**Sequence 5’ → 3’**
Rpl6_A6_3kb_4	TGCTGGAGACCAACTGTAAGG
Rpl6_3kb_genomic_AS	CAGAGGCTGACCACCATCTCTTC
Rps6_D1_gemomic_S1_NheI	agtagtagtGCTAGCTCACTGGATCAGCACAATCTTACAT
Rps6_D1_gemomic_S2_NheI	agtagtagtGCTAGCCACCCAGAAGTACACAAGAGTGAATC
Rps6_D1_gemomic_S3_NheI	agtagtagtGCTAGCGCTACCAGGGTATGTTCGATAAGAAG
Rps6_D1_gemomic_S4_NheI	agtagtagtGCTAGCTCTGGCTGGTTTTCACTGTG
Rps6_D1_genomic_AS3_XmaI	actactactCCCGGGCTCATCGTCCACTTCAATGAGTTT
Rps6_D1_genomic_AS4_XmaI	actactactCCCGGGCTTCACACAGCCAACCGC
Rps8_G15_1.4kb_genomic_S	TCCCTAATCCTGCTAATCTTGCTG
Rps8_G15_1.4kb_genomic_AS	CAGATGAAAGGCAAATTCAAACAT
Rps8_A6_gemomic_S1_NheI	agtagtagtGCTAGCCTGAGCAAAAGATATTTGTGAGCCT
Rps8_A6_gemomic_S2_NheI	agtagtagtGCTAGCCATCAATTTCCCAGGCAGACT
Rps8_A6_gemomic_S3_NheI	agtagtagtGCTAGCGTCTCTCATTGAATTATACTGGAAGCA
Rps8_A6_gemomic_S4_NheI	agtagtagtGCTAGCTTCAACTATCCCTTTCTCTGTCCTC
Rps8_A6_gemomic_S5_NheI	agtagtagtGCTAGCCAGGAAATTGTCAACAACAGTGTTT
Rps8_A6_genomic_AS1_XmaI	actactactCCCGGGTGCTCGGTGCTGGCTG

Table 3-20: Primers used for sequencing

Designation	**Sequence 5’ → 3’**
Rpl6_A6_3kb_1	GACATGGTGACAAACAAGAGGAC
Rpl6_A6_3kb_2	GAGACTTCCTAAGGTGAAGGG
Rpl6_A6_3kb_3	TAACCTCTGAGCCATCTCTC
Rpl6_A6_2kb_seq_1	TAGCGTAGGAATCAACTCTCTCG
Rpl6_A6_2kb_seq_2	ACTCGTGTACAAGTGAGCCATCT
Rps6_D1_EcoRI_seq1_S	GGACATTCCTGGACTGACAGAT
Rps6_D1_EcoRI_seq1_AS	GGGTTGGAATTAAAGGTGTGAG
Rps6_D1_EcoRI_seq2_S	GTGGACGGATTCATTGTCCT
Rps6_D1_EcoRI_seq2_AS1	ACCAGCCAGAGACCTGAGAA
Rps6_D1_EcoRI_seq2_AS2	GGCTGTGGGAATGTGTACCT
Rps6_D1_EcoRI_seq2_AS3	CTGATCCCAAACGAGGTCTT
Rps6_D1_EcoRI_seq3_AS	GCTGGCCTAGAACTTGGAAA

Table 3-21: Primers used for generating truncation mutants

Designation	**Sequence 5’ → 3’**
pGL3 back	AGATCTGCGATCTAAGTAAGCTTGG
pGL3 for	ACGCGTAAGAGCTCGGTAC
Rpl6_A6_2kb_opt_1	CCACAGACACTGGATAAGGG
Rpl6_A6_2kb_opt_2_NheI	agtagtagtGCTAGCCCTTGGGAGAAACACAGAGC
Rpl6_A6_2kb_opt_3	CAGGGCTCTGAGACTCGTGTA
Rpl6_A6_2kb_opt_3_XmaI	actactactCCCGGGCAGGGCTCTGAGACTCGTGTA
Rpl6_A6_2kb_opt_4_NheI	agtagtagtGCTAGCGGACAGGCTAGGGCTCTCTC
Rpl6_A6_2kb_opt_5	ATCCTGCCATGCCTTCCT

3.9.2 Agarose gel electrophoresis

Agarose gel electrophoresis is a common technique used in molecular biology to separate DNA molecules by size. Due to the negative charge of nucleic acid, the molecules move to the anode within an electric field. Shorter DNA strands move faster than large DNA strands through the agarose gel matrix which leads to separation. The most common method to make DNA visible is using the dye ethidium bromide which fluoresces under UV light when intercalated into DNA. Size and concentration of the DNA fragments can be estimated by comparison to an appropriate DNA marker.

Agarose gel electrophoresis was used for analytical purposes such as evaluation of PCRs or restriction digests as well as for preparation of a specific DNA fragment or plasmid. For all gel electrophoreses performed, 1% agarose gels were used. For this purpose, agarose was melted in the appropriate volume of 1× TAE buffer using a microwave oven.

Table 3-22: Composition of 1% agarose gel (360 g for 3 gels)

Amount	Components
3.6 g	Agarose
7.2 g	50× TAE Buffer
349.2 g	ddH_2O

After cooling down to about 60°C, ethidium bromide (18 µl for 360 g gel) was added. Then the liquid gel was poured into gel preparation trays assembled with the appropriate combs and allowed to solidify at room temperature. The gel was put into an electrophoresis chamber and covered with 1× TAE running buffer.

DNA samples were mixed with a suitable amount of loading buffer and applied to the gel. As loading dye a 6× BX buffer was used containing the two tracking dyes bromophenol blue and xylene cyanol FF. In 1% agarose gels bromophenol blue co-migrates with ~300 bp DNA, while xylene cyanol FF co-migrates with ~4000 bp DNA. One volume of loading buffer was added to five volumes of DNA sample. Slots of an analytical gel were loaded with up to 30 µl and slots for preparative gels with up to 60 µl. An analytical gel electrophoresis was performed at 130 V whereas 90 V was used for a preparative gel. For DNA detection a molecular imager (Gel Doc™ XR System) containing an UV-transilluminator and a digital camera was used. For preparation purpose, bands of the desired size were cut out using a bench UV-transilluminator (TPB-M/WL).

3.9.3 DNA extraction and purification from agarose gel

For DNA extraction and purification from excised gel slices, NucleoSpin® Extrakt II, the illustra™ GFX™ PCR DNA and Gel Band Purification Kit or the Wizard® SV Gel and PCR Clean-Up System were used as recommend by the manufacturer. All these kits are based on the same basic principle. The gel slice are completely dissolved and loaded to a column containing a silica membrane which binds DNA in presents of chaotropic salts added by the binding buffer. Residual contaminations like salts and soluble macromolecular components are removed by washing with an ethanolic buffer. Pure DNA is finally eluted with nuclease-free ddH_2O.

3.9.4 DNA quantification

The quantity of DNA in a solution was determined using a spectrophotometer (NanoPhotometer™ or NanoDrop 1000) for measuring the absorbance at 260 nm and calculating the corresponded concentration. Additionally, these spectrophotometers determine the ratio of the absorbance at 260 nm to the absorbance at 280 nm which gives an indication of protein contamination in the solution. A protein free DNA solution has a ratio of 1.8 – 2.0.

3.9.5 Restriction digest

Restriction enzymes are prokaryotic endonucleases which recognize and cleave specific double stranded DNA sequences generating sticky ends (3’ or 5’ overhangs) or blunt ends. Restriction digests were used for the purpose of preparing inserts and plasmids for cloning, for fragmentation of genomic DNA, or for checking a recombinant plasmid map.

All restriction endonucleases and appropriate buffers were purchased from New England Biolabs (NEB).

For complete digestions, a ratio of 2 – 10 U of restriction endonuclease to 1 µg DNA, the appropriate NEBuffer, bovine serum albumin (BSA) if required, and ddH_2O were used and incubated over night at the optimal temperature as recommended by the manufacturer. For other purposes the reaction was carried out for at least 1 h.

Inserts and plasmids for cloning were isolated by preparative agarose gel electrophoresis following DNA extraction and purification.

Double digests were performed according to NEB’s double digest finder on the corporate website.

3.9.6 Dephosphorylation

Phosphatases catalyze the hydrolysis of 5’ phosphate groups of DNA and are often used for the dephosphorylation of restriction digested cloning vectors in order to prevent self-ligation.

Dephosphorylation was performed after the digest of a plasmid vector and thermal inactivation of the restriction endonuclease using the Antarctic Phosphatase. For this purpose 1 U of the enzyme per 1 µg DNA and the appropriate amount of Antarctic Phosphatase Reaction Buffer were directly added to the digest mixture and incubated for 1 h at 37°C following inactivation at 65°C for 5 min and immediate purification of the plasmid DNA by preparative gel electrophoresis and DNA extraction and purification.

3.9.7 Phosphorylation

For the preparation of inserts generated by PCR for cloning or self-ligation, the addition of 5’ phosphates to the ends of the PCR product is required.

The phosphorylation reaction was performed using 10 U of T4 Polynucleotide Kinase, the appropriate amount of T4 DNA Ligase Buffer, ddH_2O, and the purified PCR product resulting in a 20 µl reaction volume. After incubation at 37°C for 45 min, T4 Polynucleotide Kinase was temperature inactivated at 70°C for 1 h. Phosporylated DNA was directly used for subsequent ligations.

3.9.8 Ligation

For DNA ligation, the T4 DNA Ligase was used which catalyzes the formation of a phosphodiester bond between juxtaposed 5' phosphate and 3' hydroxyl termini in double stranded DNA. For a 15 µl ligation reaction, 100 ng of linearized vector DNA were used. The amount of insert DNA was calculated according to following formula:

$$m_{insert}\ [ng] = \frac{k\ \times m_{vector}\ [ng] \times length_{insert}\ [bp]}{length_{vector}\ [bp]}$$

For sticky-end ligations a 3-fold excess of insert DNA (k = 3) was used, whereas a 5-fold excess (k = 5) was used for blunt-end ligations. The assays for sticky-end ligation containing digested vector and insert with compatible cohesive ends, the appropriate amount of T4 Ligase Buffer, 1 µl T4 DNA Ligase (400 U), and ddH_2O were incubated at room temperature for 1 h or at 16°C over night. However, blunt-end DNA ligation reactions were always incubated at 16°C over night. Isopropanol or ethanol precipitation was used to purify the ligated DNA in order to remove interfering salts for subsequent transformation into *E. coli* via electroporation.

3.9.9 DNA precipitation

Alcohol precipitation was used to concentrate DNA solutions and to remove unwanted salts (e.g. after a DNA ligation).

ISOPROPANOL PRECIPITATION

Isopropanol precipitation was carried out by adding an equal volume of isopropanol and a 0.1-fold volume of 3 M sodium acetate (pH 5.2) to the DNA solution. After incubation at room temperature for 10 min and subsequent centrifugation at 15°C and 16,000 g for 30 min, the obtained DNA pellet was washed with 70% ethanol. The pellet was completely air dried to remove residual ethanol and dissolved in at least 10 µl ddH$_2$O.

ETHANOL PRECIPITATION

For ethanol precipitation a 2.5-fold volume of 96% ethanol and a 0.1-fold volume of 3 M sodium acetate (pH 5.2) was added to the DNA solution. After incubation for 30 min at -20°C, centrifugation was performed at 4°C and 16,000 g for 30 min. The pellet was washed with 70% ethanol, air dried, and resuspended in at least 10 µl ddH$_2$O.

3.9.10 Transformation of *E. coli* by electroporation

Electroporation is the most common method for introducing plasmid DNA into bacterial cells. Applying an electric field increases the permeability of the cell membrane significantly which leads to the formation of pores and enables the transfer of the plasmid into the cells.

PREPARATION OF ELECTROCOMPETENT *E. COLI* CELLS

For the preparation of electrocompetent *E. coli* cells the strains DH10B, DH5α, JM109, or NEB10β were used and incubated in 20 ml LB medium without antibiotics at 37°C and 200 rpm using an incubation shaker over night. 400 ml LB medium were inoculated 1:100 with the overnight culture. After subsequent incubation at 37°C and 200 rpm until reaching OD$_{600}$ of 0.6 – 0.8, the cell suspension was centrifuged at 4°C and 4000 rpm for 8 min. The supernatant was discarded and the pellet was resuspended in 500 ml of 1 mM HEPES (4°C). The washing step was repeated three more times whereas the pellet was resuspended in 250 ml of 1 mM HEPES (4°C) the first time, 100 ml of 1 mM HEPES (4°C) the second time and 60 ml 10% glycerol (4°C) the third time. Then the cell suspension was centrifuged at 4°C and 5000 rpm for 8 min, the supernatant discarded and the pellet resuspended in 6 ml 10 % glycerol (4°C). Finally, the cell suspension was aliquoted (50 µl) into iced microtubes and shock frozen in liquid nitrogen and stored at -80°C. All centrifugation steps were carried out using the Avanti J-20 XP centrifuge with a JLA 10.500 rotor.

For evaluation of the suitability of the electrocompetent *E. coli* cells, the transformation efficiency was tested by electroporation using 10 pg of the pUC19 plasmid and plating on LB agar plates containing ampicillin.

ELECTROPORATION OF *E. COLI* CELLS

Cuvettes (2 mm gap) had to be prepared by lining opposed sides of the cuvette with stripes of self-adhesive, conductive aluminum foil prior electroporation.

50 µl electrocompetent *E. coli* were thawed on ice, mixed with plasmid DNA and transferred to a pre-chilled electroporation cuvette. Bacteria were electroporated using program Ec2 for bacteria (2.5 kV, 1000 Ω, and 25 µF) of the electroporator MicroPulser™, whereas the time constant should be about 5.7 ms. After applying a single pulse, the cells were immediately transferred into 900 µl SOC medium and incubated at 37°C for 30 min with shaking. The cells were then plated onto pre-dried LB agar plates containing the appropriate selection marker and incubated at 37°C over night.

3.9.11 Plasmid preparation (Mini- and midi-prep)

All kits used for the isolation of plasmid DNA from *E. coli* cells are based on alkaline lysis. The cells were cultured in LB medium containing the appropriate antibiotic, then pelleted and lysed by a solution containing the anionic detergent sodium dodecyl sulfate (SDS) following neutralization. Then the solution was clarified by pelleting the bacterial debris and the supernatant was loaded on a column containing a silica membrane that binds plasmid DNA. Contaminations and salt residues were then washed away using an ethanolic buffer. After drying of the membrane, plasmid DNA was eluted with nuclease-free water.

Plasmid mini-preparations were carried out using the Wizard®Plus SV Minipreps DNA Purification System according to the manufacturer’s instructions. Plasmid midi-preparations were performed with the PureYield™ Plasmid MidiPrep System as recommended by the producer. For the preparation of endotoxin-free plasmid DNA for transfection into CHO cells, the Nucleobond® Xtra Midi Plus EF kit was used according to the supplier’s manual.

3.9.12 DNA concentration

For concentration of DNA solutions, the vacuum concentrator centrifuge (speed vac) Savant ISS110 SpeedVac® Concentrator was used. A speed vac is made of a heated table centrifuge in which a vacuum can be generated. This decreases the boiling point of the solvent and leads to a quick evaporation of the liquid.

3.9.13 Purification of PCR products or restriction digests

For the purification of PCR products or restriction digests NucleoSpin® Extrakt II, the illustra™ GFX™ PCR DNA and Gel Band Purification Kit, or the Wizard® SV Gel and PCR Clean-Up System were used as recommend by the manufacturer. All these kits are based on the same basic principle. The DNA solutions are loaded to a column containing a silica membrane which binds DNA in presents of chaotropic salts added by the binding buffer. Residual contaminations like salts and soluble macromolecular components are removed by washing with an ethanolic buffer. Pure DNA is finally eluted with nuclease-free ddH$_2$O.

3.9.14 Sequencing of PCR products and plasmids

Sequencing of PCR products and plasmids was performed either by Agowa (Germany) or Eurofins MWG Operon (Germany).

3.9.15 Genomic DNA isolation

Genomic DNA was isolated from 5 × 10^6 CHO dhfr$^-$ cells using the DNeasy® Blood & Tissue Kit according to the manufacturer's recommendation for cultured cells.

In the first step of the protocol, the cells were directly lysed with proteinase K and then the lysat was loaded onto a spin column. During centrifugation, DNA was selectively bound to a silica-based membrane while contaminants passed through. Remaining contaminants and enzyme inhibitors were removed in two wash steps and purified DNA was finally eluted in nuclease-free ddH$_2$O.

3.9.16 Inverse PCR

Inverse PCR was used to amplify the unknown 5' and 3' flanking regions of a known genomic DNA sequence.

RESTRICTION DIGEST OF GENOMIC DNA AND SELF-LIGATION

To generate templates for Inverse PCR CHO, genomic DNA was digested using different restriction endonucleases at the appropriate working temperature for at least 16 hours (over night). 10 µg genomic DNA were used per digest.

Table 3-23: Restriction endonucleases used to digest CHO genomic DNA

Name	Sequence	Overhang	Name	Sequence	Overhand
Apa LI	GTGCAC	5´ - TGCA	*Hae* II	RGCGCY	GCGC - 3´
Apo I	RAATTY	5´ - AATT	*Hind* III	AAGCTT	5´ - AGCT
Bam HI	GGATCC	5´ - GATC	*Kpn* I	GGTACC	GTAC - 3´
Bgl II	AGATCT	5´ - GATC	*Nhe* I	GCTAGC	5´ - CTAG
Bsa HI	GRCGYC	5´ - CG	*Pci* I	ACATGT	5´ - CATG
Bsa WI	WCCGGW	5´ - CCGG	*Sac* I	GAGCTC	AGCT - 3´
Bsp HI	TCATGA	5´ - CATG	*Sph* I	GCATGC	CATG - 3´
Bsr FI	RCCGGY	5´ - CCGG	*Xba* I	TCTAGA	5´ - CTAG
Eae I	YGGCCR	5´ - GGCC	*Xho* I	CTCGAG	5´ - TCGA
Eco RI	GAATTC	5´ - AATT			

Table 3-24: Pipetting instruction for restriction digest of CHO genomic DNA

Component	Volume / 80µl	Final amount
gDNA template	x µl	10 µg
NEBuffer 10×	8 µl	
BSA 10× (if required)	8 µl	
ddH$_2$O	Add to 80 µl	
Restriction endonuclease	x µl	50 U*

* 35 U for *Bsa* WI and 21 U for *Eae* I were used.

The digested DNA was purified using a PCR clean-up kit (either NucleoSpin® Extrakt II or illustra™ GFX™ PCR DNA and Gel Band Purification Kit) and eluted with 200 µl sterile ddH$_2$O. In order to generate circular DNA fragments, the digested DNA was ligated over night at 16°C using T4 DNA ligase. The self-ligated, circular DNA was purified using isopropanol precipitation and dissolved in 30 µl sterile ddH$_2$O.

PCR USING THE SELF-LIGATED GENOMIC DNA AS TEMPLATE

For Inverse PCR, Phusion® High-Fidelity DNA Polymerase was used as described in chapter 3.9.1. The PCR was performed as nested PCR. Primers used for Inverse PCR are listed in the Table 3-18.

3.10 Cell culture methods

3.10.1 Cultivation of CHO dhfr$^-$ cells

CHO dhfr- cells were cultivated in a spinner flask and routinely subcultured twice a week at a starting density of 1.5 × 10^5 cells ml^{-1}.

3.10.2 Determination of cell number

HEMOCYTOMETER

To determine the cell number and viability, a hemocytometer (Neubauer counting chamber) was used. For this purpose, 1 ml cell suspension was mixed with 200 µl of 0.5% trypan blue, which can penetrate the cell membrane of dead cells, whereas viable cells remain unstained. The hemocytometer consists of several large squares of 1 mm² and the depth of 0.1 mm gives each large square a volume of 100 nl. The cell concentration and the viability were calculated according to following formulas:

$$cell\ concentration\ [cells\ ml^{-1}] = \frac{number\ of\ cells}{large\ square} \times 10000\ \times 1.2$$

$$viability\ [\%] = \frac{viable\ cell\ concentration}{total\ cell\ concetration} \times 100$$

COULTER COUNTER

For the determination of a more accurate total cell number by nuclei counting of lysed cells, the Multisizer™ 3 COULTER COUNTER® was used according to the manufacturer’s manual. For this purpose, 4 ml of cell culture suspension were centrifuged at 440 g for 10 min and the pellet resuspended in 1 ml Triton citric acid buffer. After incubation for 1 h at room temperature, 100 µl of the nuclei suspension (volume depends on the cell number; measured number of counts should be between 10,000 and 20,000) was mixed with 9 ml Coulter Isoton and the number of counts measured. The total cell concentration was calculated according to following formula:

$$\boldsymbol{cell\ concentration}\ [\boldsymbol{cells\ ml^{-1}}] = \boldsymbol{number\ of\ counts}\ \times \mathbf{2}\ (500\ \mu l\ measured)$$
$$\times\ \mathbf{9.1} \div \mathbf{0.1}\ (100\ \mu l\ sample\ deluted\ to\ 9.1\ ml)\ \div\ \mathbf{4}\ (4\ ml\ cell\ culture\ suspension)$$

3.10.3 Transfection of CHO dhfr$^-$ cells

4 × 10^6 CHO dhfr$^-$ cells were transfected with 10 µg of the cloned reporter vectors using the electroporation system Amaxa® Cell Line Nucleofector® Kit V. Transfection was carried out as described in the manufacturer's instruction using program H-14.

First of all, the 4 × 10^6 CHO dhfr$^-$ cells were centrifuged at 200 g for 10 min. Then the supernatant was removed completely and the cell pellet was resuspended with the nucleofection mixture containing the 10 µg plasmid DNA and 100 µl of the transfection medium composed of 82 µl Nucleofector Solution V and 18 µl Supplement. Immediately after transfection, the cell suspension was transferred into 6-well plates containing 4 ml preheated culture medium and incubated at 37°C and 7% CO_2.

The pGL3 Luciferase reporter vectors containing the firefly luciferase (*Photinus pyralis*) gene were used as transfection plasmids. The pGL3-Basic vector containing no promoter served as negative control whereas the pGL3-Promoter vector containing the SV40 promoter was used as positive control.

To generate consistent results, normalization to account for transfection efficiency and cell number variability was required. For this purpose, 1 µg of pRL-SV40 containing the *Renilla* luciferase were co-transfected.

3.10.4 Measurement of Bioluminescence

Bioluminescence was measured 24 h and/or 48 h post transfection using the Bright-Glo™ Luciferase Assay System or the Dual-Glo™ Luciferase Assay System.

For general testing of luciferase activity, the Bright-Glo™ Luciferase Assay System was used. For that, 100 µl of the cell suspension was transferred into black 96 well plates and 100 µl of Bright-Glo™ Reagent was added. After incubation for 10 min, bioluminescence was measured using either the Synergy 2 multi-mode microplate reader (Gen5™ reader control and data analysis software) or the infinite® M1000 microplate reader (i-control™ microplate reader software).

For transfections normalized by co-transfection with pRL-SV40, detection of bioluminescence was conducted using the Dual-Glo™ Luciferase Assay System. To measure firefly luciferase activity, 50 µl of the cell suspension was transferred into black 96 well plates and 50 µl of Dual-Glo® Luciferase Reagent was added. After incubation for 10 min, bioluminescence was measured using either the Synergy 2 multi-mode microplate reader (Gen5™ reader control and data analysis software) or the infinite® M1000 microplate reader (i-control™ microplate reader software). For subsequent measurement of *Renilla* luciferase activity, 50 µl of Dual-Glo® Stop & Glo® Reagent was added and bioluminescence was measured after incubating for another 10 min. The normalized promoter activity was determined from the ratio of firefly luciferase activity to *Renilla* luciferase activity.

The bioluminescence reaction catalyzed by firefly and *Renilla* luciferases is illustrated in Figure 3-4.

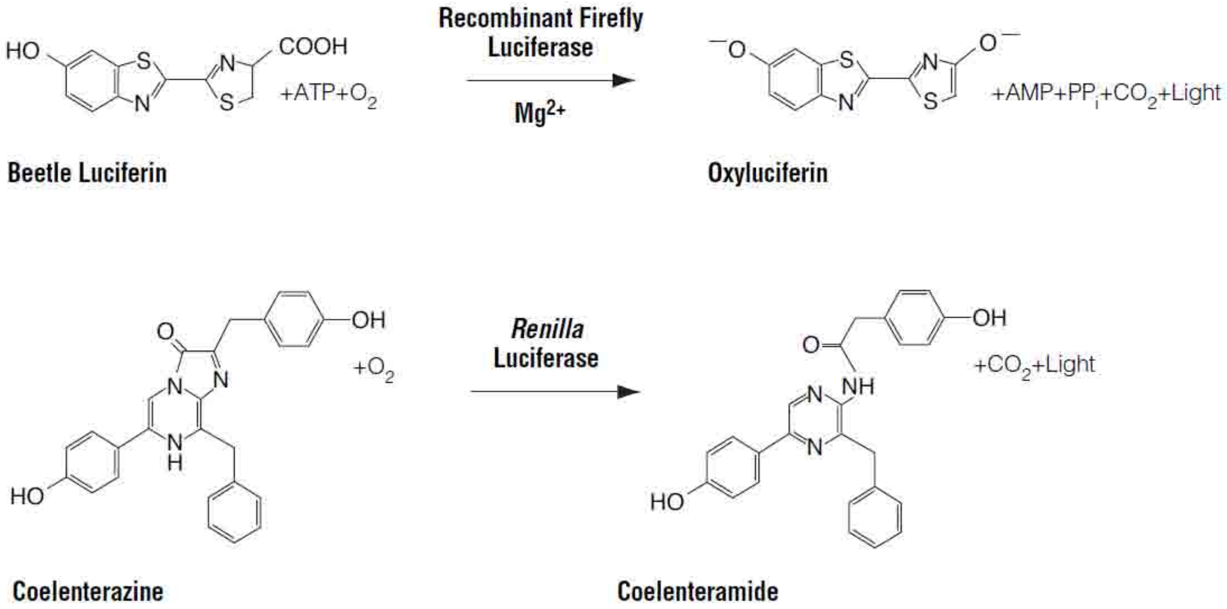


Figure 3-4: Bioluminescent reactions catalyzed by firefly and *Renilla* luciferases
Mono-oxygenation of beetle luciferin is catalyzed by firefly luciferase in the presence of Mg^{2+}, ATP, and molecular oxygen. Unlike beetle luciferin, coelenterazine undergoes mono-oxygenation catalyzed by *Renilla* luciferase but requires only molecular oxygen [**133**].

4 Experiments

4.1 Part I: Identifying the 5’ flanking regions of known CHO genes

Part I continued the research which was previously conducted by Martina Baumann for her master thesis [**128**].

Transcription regulatory sequences are generally located upstream of the transcription start site [**74**,**40**]. The aim of this approach was the identification of the 5’ flanking region of highly expressed CHO genes. For this purpose, known cDNA sequences of eight highly abundant genes derived from the Consortium for Chinese Hamster Ovary Cell Genomics [**30**] were chosen. These were the genes coding for the ribosomal proteins S6 (*Rps6*), S8 (*Rps8*), L6 (*Rpl6*), L27 (*Rpl27*), L35 (*Rpl37*), the ribosomal protein large P1 (*Rplp1*), the tumor protein translationally-controlled 1 (*Tpt1*), and the jun proto-oncogene related gene d1 (*Jund1*).

The described method relied on the quality of the available cDNA sequence data, the construction of a genomic library with high sequence coverage, and the design of PCR primer with great binding specificity.

4.1.1 Library construction

The construction of a genomic library requires the generation of a random pool of genomic DNA fragments. For this purpose, many different methods are described in literature including mechanical techniques like passage through a large gauge needle of a syringe [**134**,**135**,**136**], nebulization [**137**], sonication [**138**], stirring in a blender [**139**], or enzymatical treatments like digestion by restriction endonucleases [**131**] or DNase I [**140**]. Library construction was carried out by Martina Baumann and is quoted here to complete the whole experiment [**128**].

Genomic DNA was isolated from 2 × 10^7 CHO cells and fragmented using two different methods to ensure maximum heterogeneity of the library. The chromosomal CHO DNA was sheared by nebulization as well as cut by single digestion using the blunt-end generating restriction endonucleases *Msc* I, *Sca* I, *Ssp* I, and *Stu* I. These endonucleases were chosen from a panel of different restriction endonucleases tested with regard to high efficiency and absence of methylation sensitivity. Restriction sites are located in AT rich regions (*Ssp* I), in GC rich regions (*Stu* I and *Msc* I) as well as in purine and pyrimidine rich regions (*Sca* I) in order to create a library revealing considerable diversity. After removal of small DNA pieces by preparative agarose gel electrophoresis or isopropanol precipitation, the obtained fragments were inserted into the two vector systems pMACS 4.1 (Miltenyi Biotec, Germany) and pVitro [**128**] containing the ampicillin resistance gene via blunt-end cloning (Figure 4-1), whereas the type of vector had no relevance for this experiment.

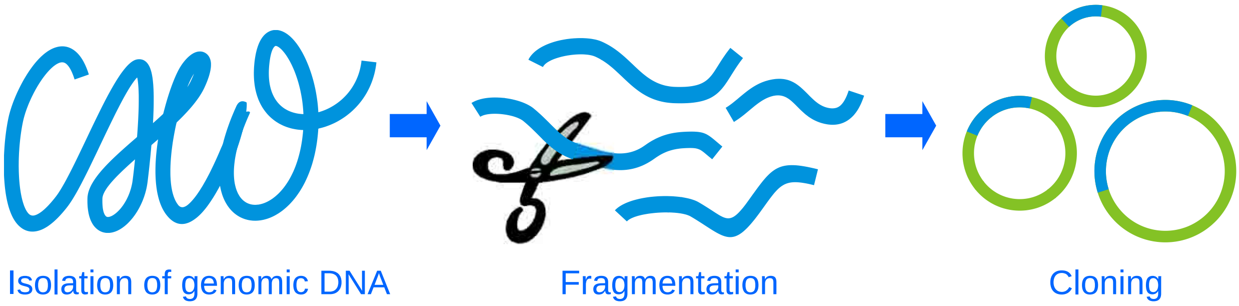


Figure 4-1: Library construction
CHO genomic DNA was isolated, fragmented using different methods to ensure maximum heterogeneity, and inserted into a vector via blunt-end cloning.

For library construction, PCR amplified vector copies were used in order to avoid re-ligation without an insert which drastically reduced background. After transformation of the ligation mixture into the electrocompetent *E. coli* MegaX DH10B and plating on LB/ampicillin agar plates, the resulting colonies were counted and the coverage of the libraries calculated. For the pMACS system a library with 4 × 10^5 colonies could be generated, whereas the library gained from the pVitro system had less diversity with 6 × 10^4 colonies. Colonies on LB/ampicillin agar plates were then rinsed with 10 ml of LB/ampicillin medium and cultured in 50 ml LB/ampicillin medium over night. Subsequent midi-prep resulted in a genomic DNA library.

PCR amplification of the pooled genomic DNA libraries showed an even size distribution over a wide range (Figure 4-2).

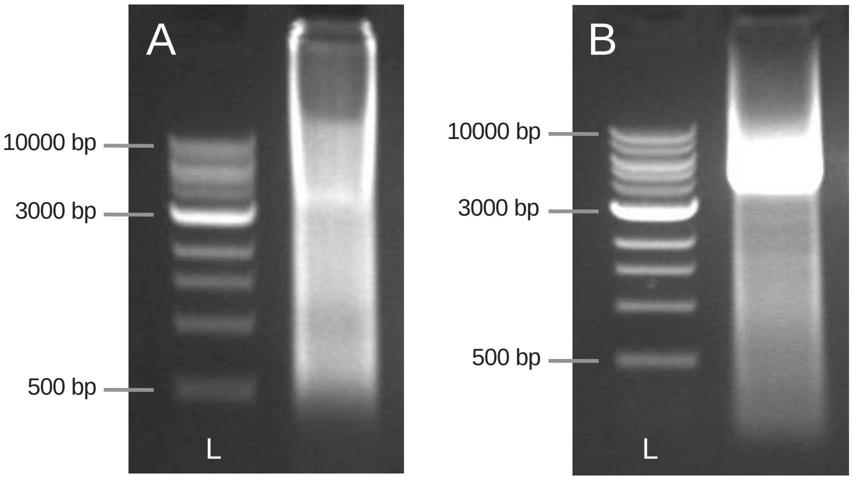


Figure 4-2: Size distribution of the CHO genomic DNA libraries
Agarose gel electrophoresis; A: Library in pMACS vector; B: Library in pVitro vector; Lane L: 1 kb DNA Ladder; Size distribution was determined by PCR amplification of the library pool [**128**].

4.1.2 Target gene identification

The known cDNA sequences of the eight highly expressed CHO genes *Rps6*, *Rps8*, *Rpl6*, *Rpl27*, *Rpl37*, *Rplp1*, *Tpt1*, and *Jund1* were aligned against the fully annotated genome of the house mouse (*Mus musculus)* in order to identify exon 1 of the respective genes. Two primers specifically binding to exon 1 were designed to perform a nested PCR. Additionally, two other sets of primers were constructed annealing to the backbone of the pMACS vector or to the pVitro vector, respectively, upstream and downstream of the inserted CHO DNA sequence in order to achieve amplification of the specific genomic fragment which was inserted in random orientation into the plasmid via blunt-end cloning (Figure 4-3).

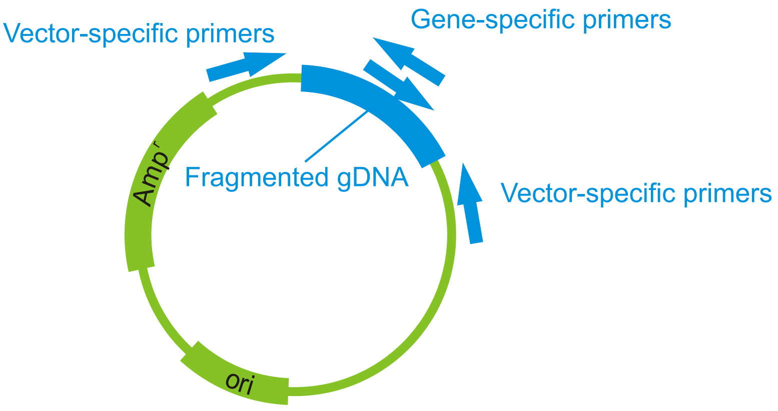


Figure 4-3: Primer design for Library PCR
Amp^r: ampicillin resistance gene; ori: origin of replication for propagation in *E. coli*

4.1.3 Library PCR

Using 6 ng of the constructed library as template, nested PCR was performed with the designed primers, whereas 2 µl of the first PCR run where directly used without purification as template for the second one. The Phusion® High-Fidelity DNA polymerase was used for amplification to achieve high accuracy of the products.

The resulting PCR products were separated via agarose gel electrophoresis and cut out bands purified using a gel-extraction kit to remove impurities like enzymes and primers. The obtained fragments were cloned into the multiple cloning region of the pGL3-Basic reporter vector upstream of the firefly (*Photinus pyralis*) luciferase gene in order to analyze putative promoter activity. Derived plasmid constructs were transfected into CHO dhfr$^-$ cells and promoter activity was evaluated by measuring the bioluminescence 48 h post transfection.

4.2 Part II: Discovery of the flanking regions of known genomic CHO sequences

In order to further characterize identified sequences and to potentially increase promoter or enhancer activity, the flanking regions of in Part I discovered transcriptionally active fragments have been further investigated by Inverse PCR.

4.2.1 The Inverse PCR approach

Inverse PCR is a method that enables the rapid *in vitro* amplification of unknown DNA sequences that flank a region of a known sequence. This technique uses the common PCR, but has the primers oriented in the reverse direction of the usual orientation. The template for Inverse PCR is a restriction fragment that has been self-ligated in order to form a circular DNA molecule [**141**].

To generate templates for Inverse PCR, CHO genomic DNA was isolated and digested using a single restriction endonuclease for each preparation that generates sticky ends and does not cut the sequence of which the flanking regions should be identified. Several preparations have been conducted in parallel using various restriction nucleases (see Table 3-23) in order to increase the likelihood of generating a suitable template. Subsequently, derived fragmented genomic DNA was purified and self-ligated in order to generate circular DNA molecules followed by an isopropanol precipitation. The obtained pool of circular DNA fragments served as template for Inverse PCR (Figure 4-4).

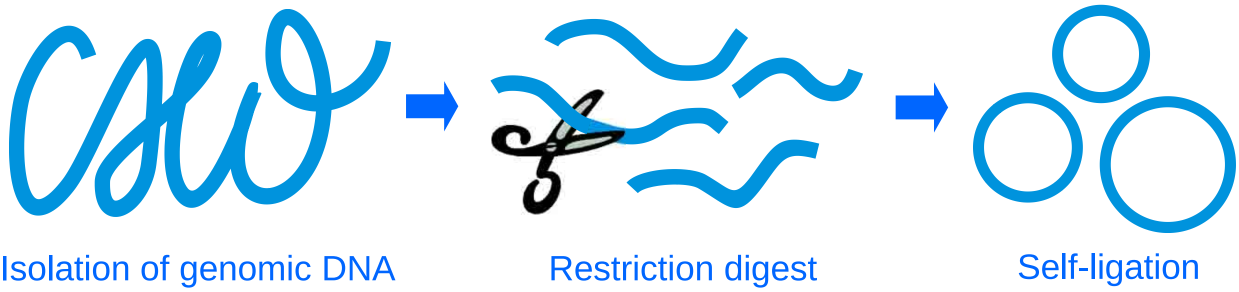


Figure 4-4: Generation of templates for Inverse PCR
Isolated CHO genomic DNA is digested using a single restriction endonuclease for each preparation following self-ligation of derived fragments.

Two sets of Inverse PCR primers oriented in reverse direction of the usual orientation were designed on the obtained DNA sequence of the fragments derived from Library PCR in order to perform a nested PCR. The first PCR reaction was conducted using 200 ng of the self-ligated genomic DNA fragments per 50 µl reaction volume as template and the primers SP1 (sense primer 1) and AP1 (antisense primer 1). In order to increase specificity, a second round of PCR amplification was performed using 2 µl of the PCR product as template and the primers SP2 (sense primer 2) and AP2 (antisense primer 2). The resulting PCR product were separated via agarose gel electrophoresis and cut out bands purified. Obtained DNA fragments contained the flanking region of the initial DNA fragment. As the ligation site was known, the 3' and 5' regions could easily be identified be sequencing the PCR product (Figure 4-5).

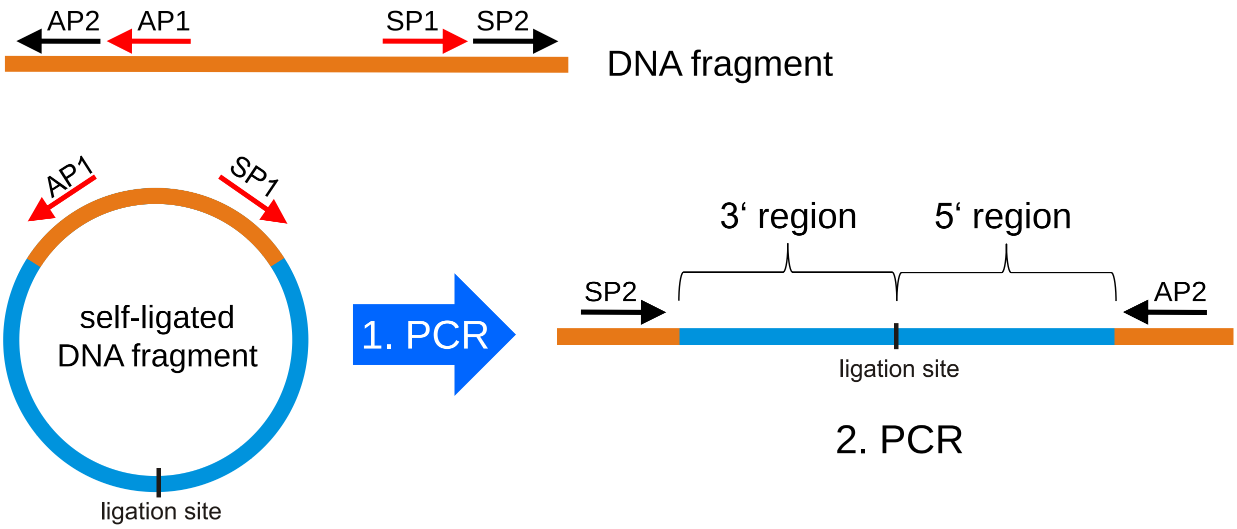


Figure 4-5: Principle of Inverse PCR
SP1: sense primer 1; AP1: antisense primer 1; SP2: sense primer 2; AP2: antisense primer 2; Inverse PCR primers are designed pointing of each other. First PCR is performed using the self-ligated CHO DNA fragments as template and the primers AP1 and SP1. The second PCR using the product of the first round of PCR amplification as template and SP2 and AP2 as primers leads to a fragment containing the 3' and 5' flanking region of the initial fragment.

4.2.2 Verification and optimization of Inverse PCR for CHO genomic DNA

Although Inverse PCR is a well-known and well established method for the *in vitro* amplification of the unknown flanking regions of a known genomic sequence, the application of this technique for CHO genomic DNA has not been published yet.

In order to evaluate and optimize the Inverse PCR technique for CHO genomic DNA sequences, the method was applied to retrieve known CHO genomic DNA sequences. For that, a restriction endonuclease which is suitable for Inverse PCR and cuts the known CHO genomic DNA sequence twice was identified (e.g. *Kpn* I; Figure 4-6). Primers for Inverse PCR were designed in the center between the two restriction sites in order to get a large enough PCR product for proper identification. Using this restriction endonuclease, CHO genomic DNA was digested and resulting DNA fragments self-ligated to generate templates for Inverse PCR. Inverse PCR was conducted as described and resulting PCR products were sequenced and compared to the initial known CHO genomic DNA sequence in order to verify the functionality of this method for CHO genomic DNA fragments (Figure 4-6).

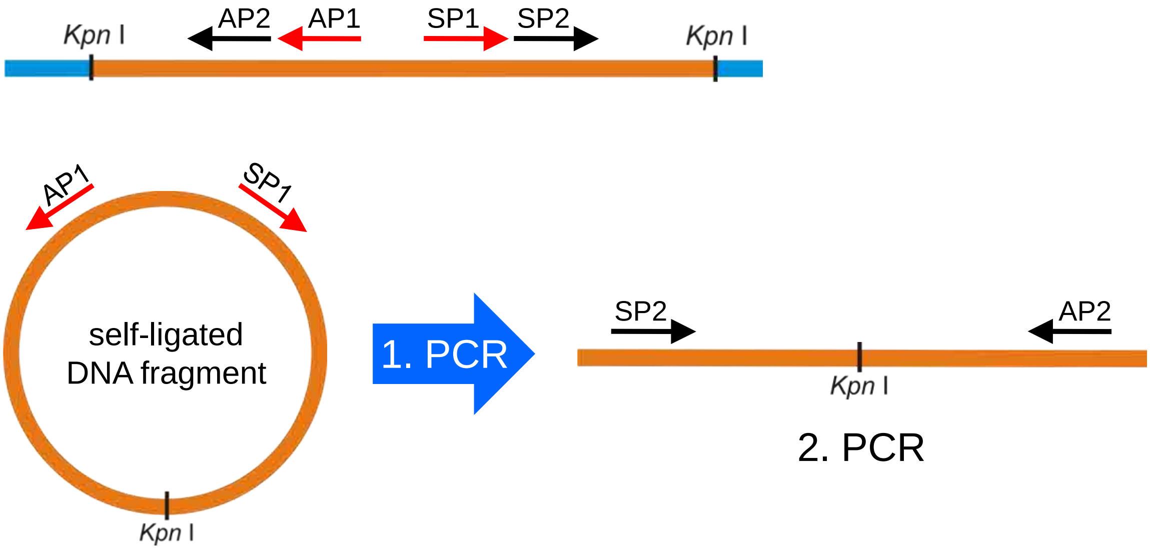


Figure 4-6: Verification of Inverse PCR for CHO genomic DNA
SP1: sense primer 1; AP1: antisense primer 1; SP2: sense primer 2; AP2: antisense primer 2; Inverse PCR primers were designed that specifically bind in the center between two equal restriction sites (e.g. *Kpn* I) of the known genomic CHO sequence. CHO genomic DNA was digested with the respective restriction enzyme and self-ligated in order to generate the template for subsequent PCR amplification.

Besides evaluating the functional capability of the Inverse PCR approach for genomic CHO DNA, the optimal PCR conditions were determined by varying the template concentration (10 ng – 400 ng) as well as the PCR cycle conditions.

4.2.3 Genomic PCR

Fragments obtained by Inverse PCR were fully sequenced in order to rediscover the self-ligation sites. This enabled the exact identification of the 5' and 3' flanking region of the initial sequence. In order to generate the complete section of the genome in the correct order, primers for PCR amplification specific to the newly-discovered flanking regions were designed pointing of each other as shown in Figure 4-7. PCR reaction was performed directly from genomic DNA using primers containing the restriction sites *Nhe* I and *Xma* I for direct cloning into the multiple cloning region of the pGL3-Basic reporter vector. 200 ng of CHO genomic DNA per 50 µl reaction volume were used as template for PCR amplification.

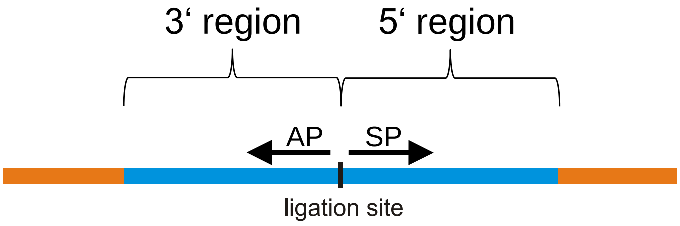


Figure 4-7: Primers for genomic PCR
SP: sense primer; AP: antisense primer

4.3 Part III: Characterization of putative *cis*-regulatory elements

4.3.1 *In silico* sequence analyses

Sequences obtained by Inverse PCR were fully sequenced in order to verify the results and for further analyses. The sequences were aligned against the initially used CHO cDNA sequences as well as against the fully annotated genome of the house mouse (*Mus musculus*) using the nucleotide blast of the Basic Local Alignment Search Tool (BLAST; http://blast.ncbi.nlm.nih.gov/Blast.cgi) in order to assign the identified promoter regions to a specific gene. Furthermore, potential transcription factor binding sites like TATA boxes, Sp1 binding sites, or NF-κB binding sites were identified using the web-based transcription factor binding sites (TFBSs) prediction tool ConSite (http://www.phylofoot.org/consite) [**125**] or the promoter prediction program NNPP 2.2 (neural network promoter prediction 2.2; http://www.fruitfly.org/seq_tools/promoter.html) [**114**].

4.3.2 Preparation of promoter constructs

Based on the data from *in silico* sequence analyses, fragments of different length were generated in order to identify motifs essential for promoter activity. For this purpose, several PCR primers were designed based on the sequenced fragments derived from Inverse PCR to generate different constructs of various length by PCR amplification. PCR was performed either directly from chromosomal DNA using 200 ng of CHO genomic DNA per 50 µl reaction volume as template or from the plasmid constructs obtained by Inverse PCR approach using 10 ng of template DNA per 50 µl reaction volume. The used primers contained the restriction sites *Nhe* I and *Xma* I for direct cloning into the multiple cloning region of the pGL3-Basic reporter vector. All plasmids were transfected into CHO cells and after 48 h bioluminescence was measured.

5 Results and discussion

5.1 Identified 5’ flanking regions of known CHO genes

5.1.1 Fragments obtained by Library PCR

Employing the Library PCR technique, several 5’ flanking regions of known genes could be obtained by amplification via nested PCR. Figure 5-1 shows the PCR products of the potential 5’ flanking regions of the genes *Rps6*, *Rps8*, *Rpl6*, *Rpl35*, *Rplp1*, and *Tpt1* which were identified by Martina Baumann from the pMACS library loaded onto a 1% agarose gel [**128**].

Library PCR was repeated with modified PCR conditions and using both libraries. The amount of template per 50 µl reaction volume was increased from initially used 1.2 ng to 6 ng. Furthermore, melting temperature T_m for the primers was calculated using the Finnzymes’ T_m calculator. The optimal annealing temperature for the Phusion® High-Fidelity DNA polymerase was calculated by adding 3°C to the lower calculated T_m value of the two primers used for the PCR reaction. The changed conditions led to several more PCR fragments. PCR products of further potential 5’ flanking regions of the genes *Rps6*, *Rps8*, *Rpl35*, *Tpt1*, and *Jund1* loaded onto a 1% agarose gel are shown in Figure 5-2.

For gene *Rpl27* no suitable PCR product could be obtained. However, Library PCR for the other seven genes yielded into at least one usable fragment ranging from 250 bp up to 1800 bp. Fragments could be obtained from the two libraries, pMACS as well as pVitro. Although most of the derived fragments are just 500 bp or shorter, they likely contain gene regulatory sequences as promoter regions are enriched within the 500 bp segment upstream of the transcription start site [**142**,**52**].

All DNA fragments obtained by Library PCR were blunt-end cloned into the multiple cloning site (*Sma* I site) of the reporter vector pGL3-Basic directly upstream of the firefly luciferase reporter gene resulting in constructs covering both directions of the inserts.

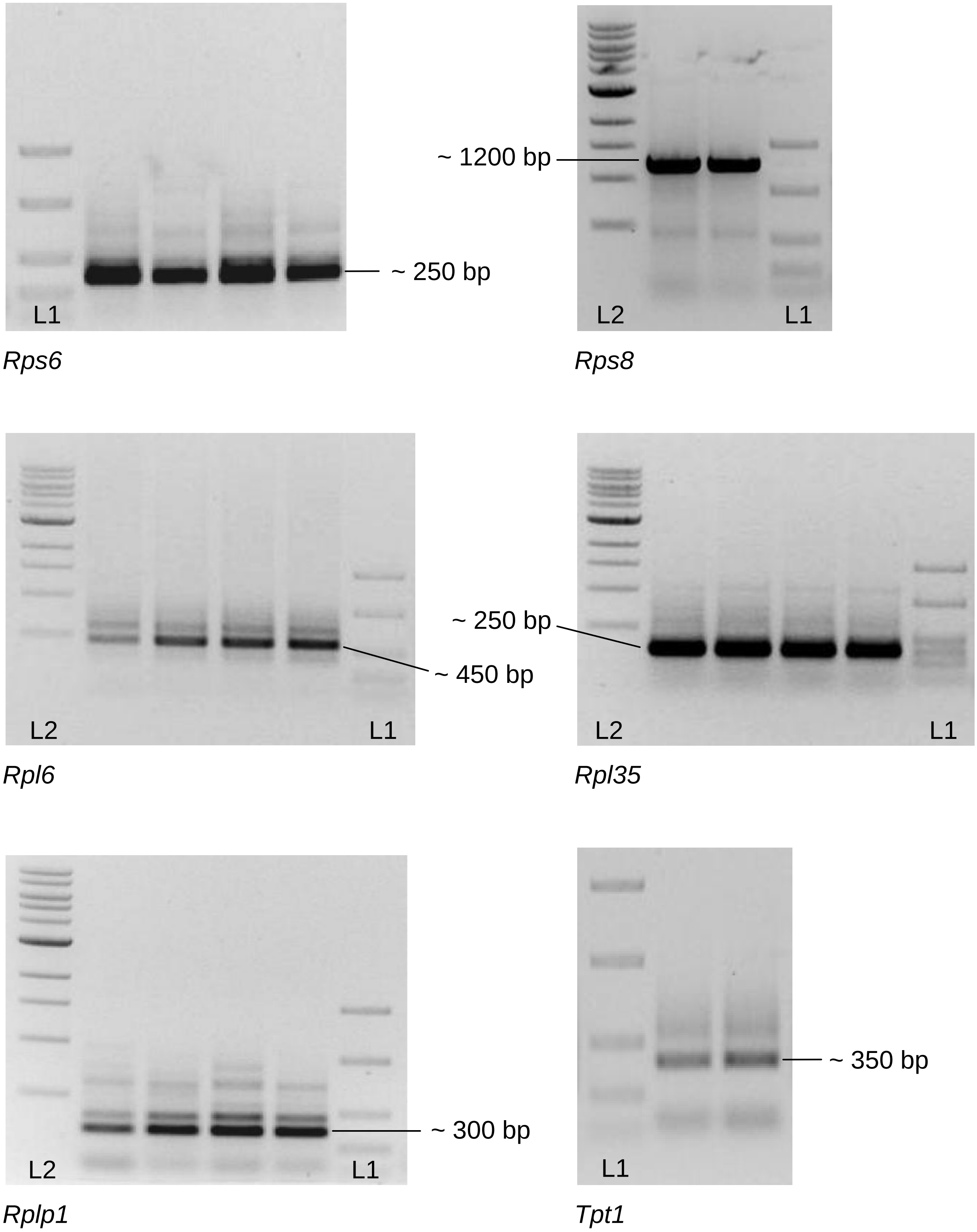
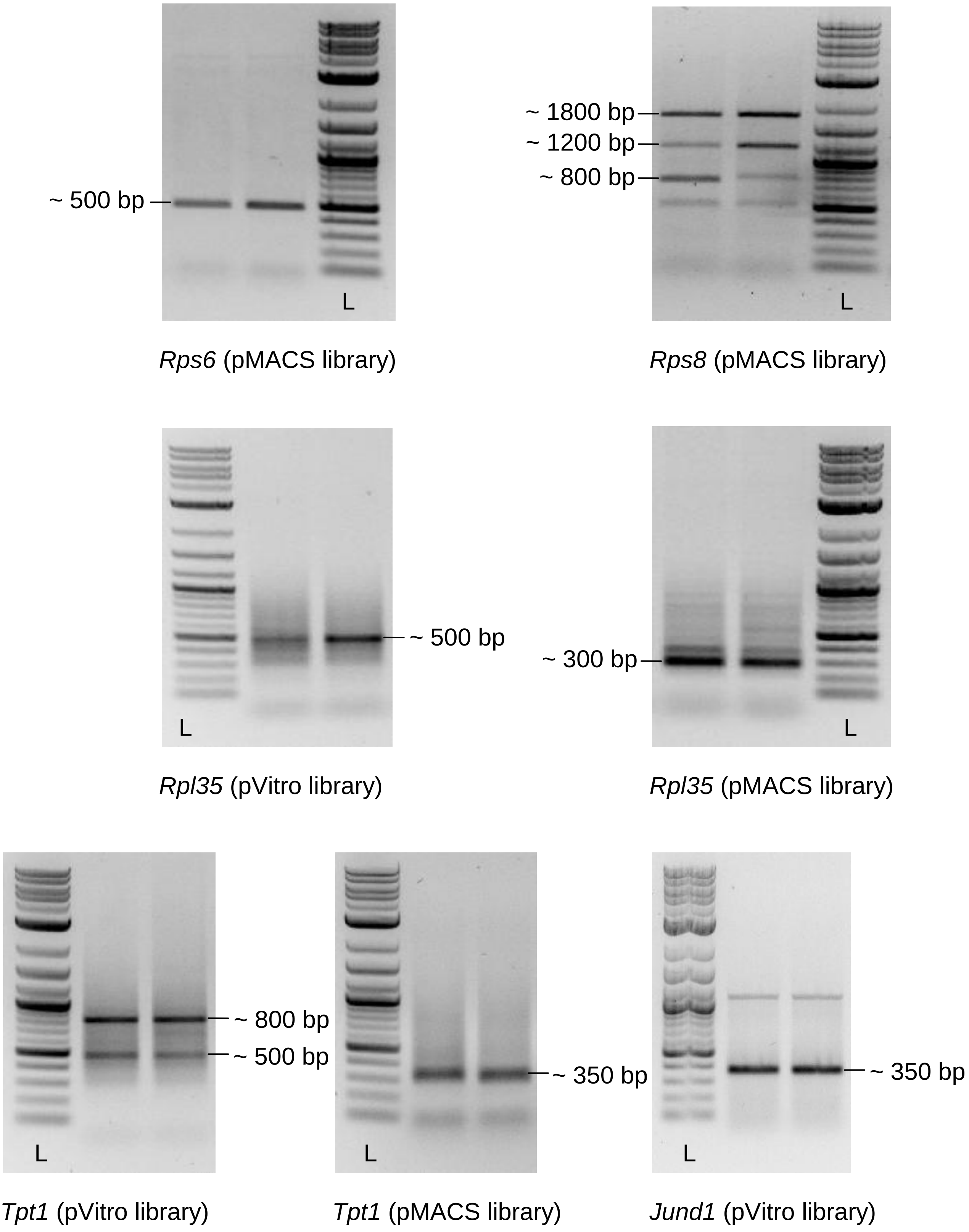


Figure 5-1: Fragments obtained via Library PCR by Martina Baumann
Agarose gel electrophoresis; All fragments derived from pMACS library; L1: FastRuler™ DNA Ladder Low Range (1500 bp, 850 bp, 400 bp, 200 bp, and 50 bp); L2: 1 kb DNA Ladder (10 kb, 8 kb, 6 kb, 5 kb, 4 kb, 3 kb, 2 kb, 1.5 kb, 1 kb, and 0.5 kb) [**128**]



Rps6 (pMACS library)

Rps8 (pMACS library)

Rpl35 (pVitro library)

Rpl35 (pMACS library)

Tpt1 (pVitro library)

Tpt1 (pMACS library)

Jund1 (pVitro library)

Figure 5-2: Additional fragments obtained by Library PCR
Agarose gel electrophoresis; L1: 2-log DNA Ladder (10 kb, 8 kb, 6 kb, 5 kb, 4 kb, 3 kb, 2 kb, 1.5 kb, 1.2 kb, 1 kb, 0.9 kb, 0.8 kb, 0.7 kb, 0.6 kb, 0.5 kb, 0.4 kb, 0.3 kb, 0.2 kb, and 0.1 kb)

5.1.2 *In silico* sequence analysis – alignment against cDNA

All sequences obtained by Library PCR were fully sequenced for further analyses. To verify the functionality of the Library PCR approach, obtained sequences were aligned against the CHO cDNA sequences initially used for primer design, in order to rediscover the gene-specific primer binding sites using the nucleotide blast (blastn algorithm) of the Basic Local Alignment Search Tool (BLAST). Many of the analyzed sequences turned out to be unspecific as they were PCR products generated just by the vector-specific primer or products of PCR amplification with the first nested PCR primer that did not contain the binding sequence of the second primer. However, for some sequences the binding site of the second gene-specific PCR primer could be rediscovered as well as the residual upstream region of exon 1. Figure 5-3 shows the alignment results for the promoter candidates Rps6 D1 and Rps8 A6[*].

Rps6 D1:

```
Score =  136 bits (150),  Expect = 3e-36
Identities = 90/99 (90%), Gaps = 2/99 (2%)
Strand=Plus/Plus

                                     5' UTR
Rps6 D1       389  GCTCTTTTTC--GTGGCACCTCCTAGGCGGTTGGCTGTGTGAAGATGAAGCTGAATTTCT  446
                   ||| ||||||  ||||| ||||||||| ||| |||||||||||||||||||||||| |||
Rps6 cDNA     173  GCTTTTTTTCCCGTGGCGCCTCCTAGGTGGTCGGCTGTGTGAAGATGAAGCTGAATATCT  232


                                         Primer binding site
Rps6 D1       447  CCTTCCTGGCCACCAGCTGCCAGAAACTCATCGAAGTGG  485
                   |||||| ||||||| |||||||||||||||||||||||
Rps6 cDNA     233  CCTTCCAGGCCACCGGCTGCCAGAAACTCATCGAAGTGG  27
```

Rps8 A6:

```
Score = 53.6 bits (58),  Expect = 6e-11
Identities = 32/34 (94%), Gaps = 0/34 (0%)
Strand=Plus/Plus

                    5' UTR          Primer binding site
Rps8 A6      1155  CCGAGCAATGAGCATCTCTCGGGACAACTGGCAC  1188
                   |||||| ||| |||||||||||||||||||||||
Rps8 cDNA    2     CCGAGCGATGGGCATCTCTCGGGACAACTGGCAC  35
```

Figure 5-3: Alignments of Rps6 D1 and Rps8 A6 against the respective cDNA
Sequences were aligned using the blastn algorithm of the BLAST. Primer binding sites could be identified as well as the upstream part of exon 1 including the 5' UTR. The ATG start codon is marked in bold.

[*] Nomenclature of promoter candidates: The first part refers to the genes which was used to design PCR primers specifically binding to exon 1 of respective genes for Library PCR (e.g. Rps6 and Rps8). The second part refers to the clone that was picked for PCR screening (e.g. D1 and A6).

The significant homologies of the primer binding sites' upstream region indicate that the correct regions of the genome have been amplified by Library PCR in case of Rps6 D1 and Rps8 A6. For two other promoter candidates (Rpl35 F3 and Tpt1 D6), the primer binding sites could by re-discovered as well but there were no significant homologies in the 5' flanking regions of the primer binding sites. The numerousness of unspecific PCR products highlights the importance of a proper primer design. The length of the primers used for Library PCR was between 18 and 21 nucleotides. Although the statistical probability of 18 nucleotide sequence occurring in a genome of approximately 3 × 10^9 base pairs is already far below 1[*], the practical experiment showed that 18 – 21 nucleotide sequences are maybe not necessarily unique in the CHO genome. But also similar sequences that allow primers annealing could have led to unspecific PCR amplifications. However, using longer primers and an optimized annealing temperature might decrease the probability of unspecific PCR products making Library PCR a more reliable tool for the identification of further 5' flanking regions of known genes.

5.1.3 Promoter activity

All promoter candidates were transfected into CHO dhfr$^-$ cells and bioluminescence was measured 48 h post transfection. The promoterless pGL3-Basic reporter vector was used as negative control (-) to detect the background expression level of the luciferase reporter assay. The pGL3-Promoter reporter vector containing the Simian virus 40 (SV40) promoter served as positive control (+).

The three constructs Rps6 D1 (485 bp), Rpl6 A6 (499 bp), and Rps8 G15 (761 bp)[†] showed considerable promoter activity. In order to get significant quantitative data, bioluminescence values were normalized to *Renilla* luciferase measurements to account for transfection efficiency and cell number variability. Therefore, the plasmid pRL-SV40 encoding the *Renilla* luciferase gene under control of the SV40 promoter was co-transfected. Promoter activity was calculated as percentage relative to the bioluminescence value measured for the SV40 promoter. Determined promoter activities of the three constructs Rps6 D1, Rpl6 A6, and Rps8 G15 are illustrated in Figure 5-4. The values shown are the average of two independent experiments with triplicate samples.

[*] $3 \times 10^9 \div 4^{18} = 0.04$

[†] The values put in parentheses show the length (number of base pairs) of the promoter candidates.

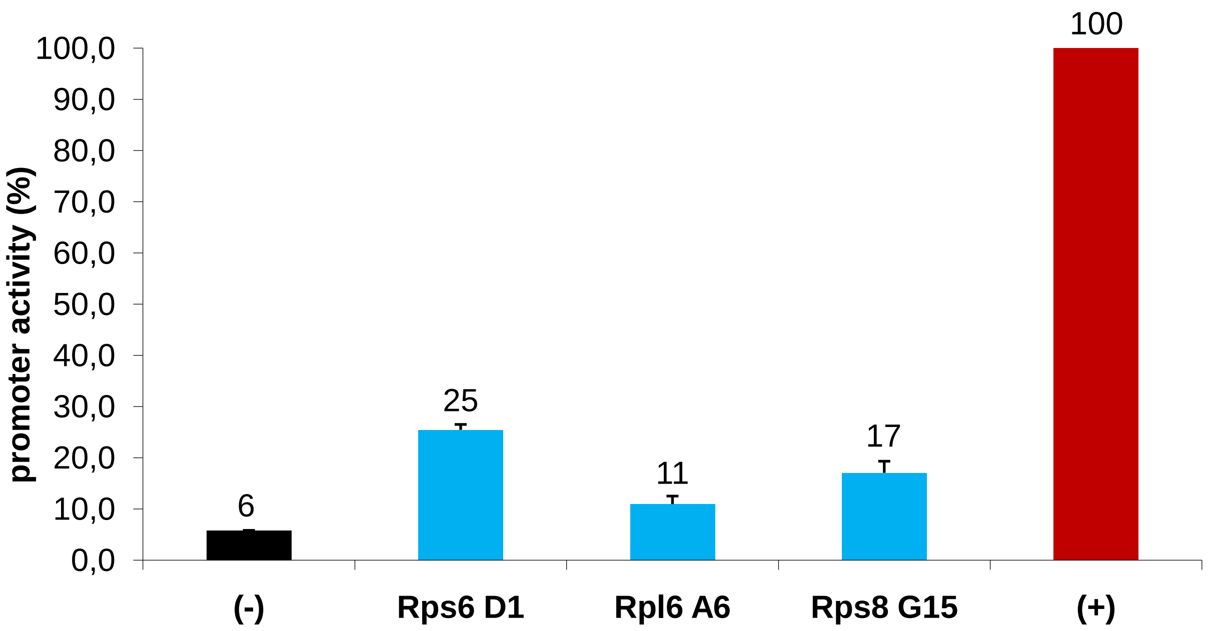


Figure 5-4: Reporter activity assay of fragments derived from Library PCR
pGL3 reporter vector constructs were transfected into CHO cells and the promoter activities were determined as percentage relative to measured bioluminescence value for the SV40 promoter; (-): negative control, promoterless luciferase reporter vector pGL3-Basic; (+): positive control, luciferase reporter vector pGL3-Promoter containing the SV40 promoter

Promoter candidate Rps6 D1 showed the highest promoter activity with about 25% of the SV40 promoter. Furthermore, the sequence could be linked to the *Rps6* gene by *in silico* analysis (see chapter 5.1.2). The constructs Rpl6 A6 and Rps8 G15 showed some promoter activity as well, 11% and 17% of SV40 respectively. However, alignment against the cDNA revealed them as unspecific PCR products.

Surprisingly, the putative promoter candidate Rps8 A6 showed no promoter activity. But maybe the fragment is just too long having 1188 bp and so might contain regulatory elements which mediate the repression of gene expression. Hence, truncated fragments of Rps8 A6 could show gene regulatory activity.

It might be possible that obtained fragments showing promoter activity do not contain the complete sequence necessary for maximum activity. So the identification of the 5’ flanking region could potentially increase promoter or enhancer activity. One possible approach would be a second round of Library PCR using primers annealing to the 5’ end of the identified regions. However, a more convenient method is the Inverse PCR, which allows the amplification of unknown genomic DNA sequences that flank a region of a known sequence. Here, this technique was used to further investigate identified promoter candidates.

5.2 Identified flanking regions by Inverse PCR

5.2.1 Rediscovery of a known CHO sequence by Inverse PCR

In order to verify the suitability of the Inverse PCR approach for genomic CHO DNA, the method has been applied to rediscover two known sequences (here referred to as CHO_DNA_1 and CHO_DNA_2). Inverse PCR was performed using different amounts of template for the first round of PCR amplification in order to determine the optimum. The experiment for CHO_DNA_1 was performed with two different self-ligated templates generated by using the restriction endonucleases *Kpn* I and *Pci* I, respectively. For CHO_DNA_1 just the *Kpn* I self-ligated template was used. Amplification was performed as nested PCR. The resulting PCR products were loaded onto a 1% agarose gel. Figure 5-5A illustrates the fragments obtained after the first round of PCR amplification and Figure 5-5B shows the fragments derived after the second round.

Fragments of the correct length could already be generated after the first PCR amplification for both CHO_DNA_1 and CHO_DNA_2 using the *Kpn* I self-ligated template. However, the preparation using the *Pci* I self-ligations as template generated no DNA fragments. Although the length of the primers was 27 nucleotides and more, PCR products obtained from a single primer were generated underlining again the importance of proper primer design for PCR amplification from large genomes in order to get specific products.

The second round of PCR amplification generated more unspecific fragments, but also fragments of the correct size. To finally verify the functionality of the Inverse PCR approach, obtained bands of the correct size were cut out, purified, and sequenced. The sequence date confirmed the rediscovery of the correct sequences.

The experiment verified the suitability of Inverse PCR for CHO genomic DNA. However, it is necessary to perform several preparations in parallel using different restriction enzymes for the generation of self-ligated templates. This increases the chance of generating circular DNA fragments of adequate size for PCR amplification.

Furthermore, the experiment showed that 10 ng of template per 50 µl reaction volume were insufficient to generate an adequate amount of PCR products. On the contrary 200 ng and more of template DNA per 50 µl reaction volume yielded the most reliable and best results.

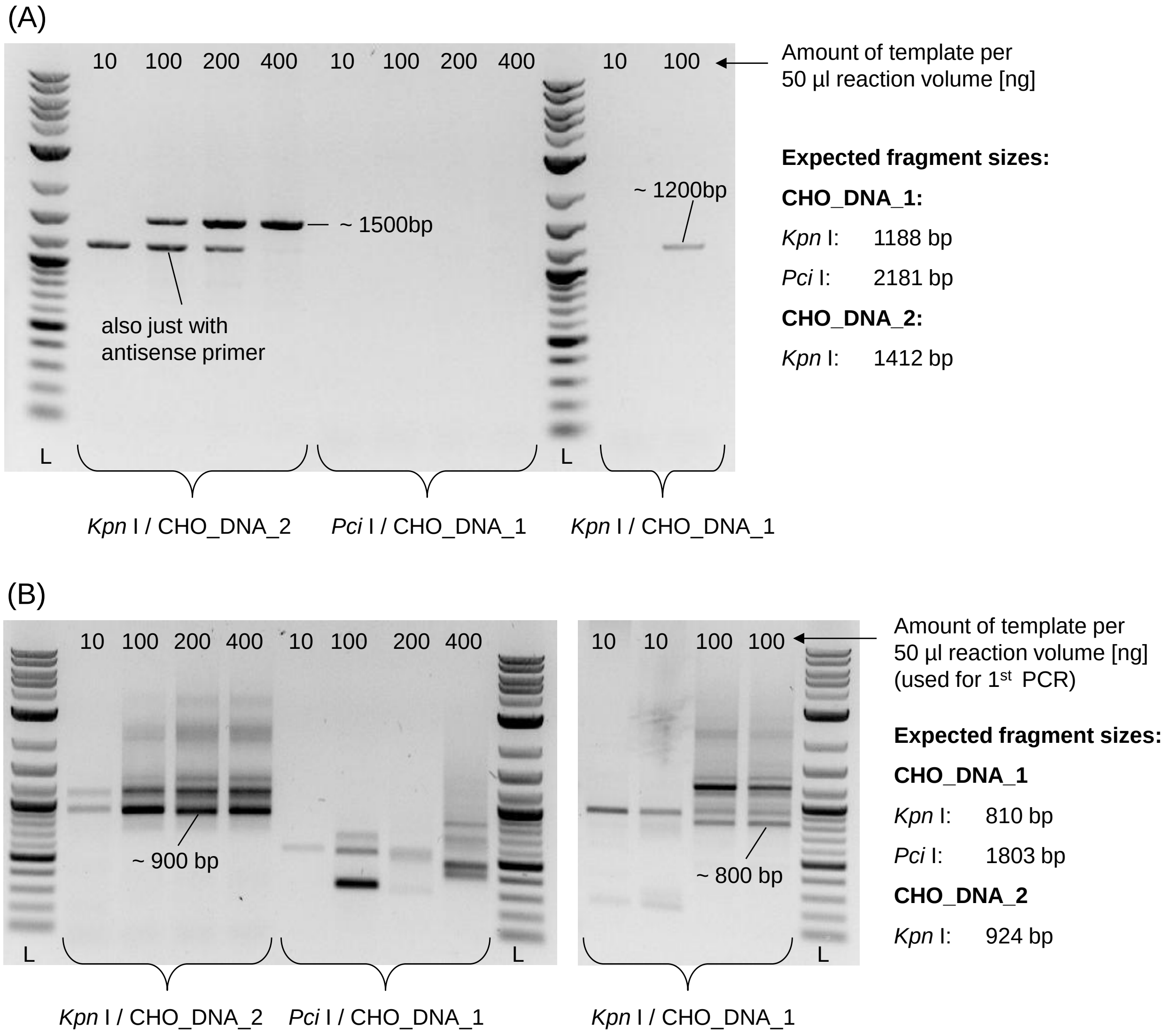


Figure 5-5: Identified known 5’ and 3’ regions of 2 CHO DNA fragments by Inverse PCR
Agarose gel electrophoresis; Lane L: 2-log DNA Ladder (10 kb, 8 kb, 6 kb, 5 kb, 4 kb, 3 kb, 2 kb, 1.5 kb, 1.2 kb, 1 kb, 0.9 kb, 0.8 kb, 0.7 kb, 0.6 kb, 0.5 kb, 0.4 kb, 0.3 kb, 0.2 kb, and 0.1 kb); Remaining lanes: Products of Inverse PCR using different templates, different amount of template, and different primers; *Kpn* I and *Pci* I refer to the restriction endonuclease used for generating self-ligated DNA templates; CHO_DNA_1 and CHO_DNA_2 refer to the used primers specific to the first CHO DNA sequence and the second CHO DNA sequence, respectively; Inverse PCR was performed as nested PCR; (A) PCR products obtained after the first PCR amplification run; (B) PCR products derived from the second PCR reaction

5.2.2 Identified flanking regions of Rps6 D1, Rps8 G15, and Rpl6 A6

Applying the Inverse PCR approach for the three fragments Rps6 D1, Rps8 G15, and Rpl6 A6 which showed promoter activity revealed 5' and 3' flanking regions up to almost 5000 bp (Figure 5-6). One fragment each was fully sequenced and the self-ligation sites (Rps8 G15: *Eae* I-site, Rpl6 A6: *Bgl* II-site, and Rps6 D1: *Eco* RI-site) could be rediscovered and thus enabled the exact identification of the 5' and 3' flanking region of the initial sequences.

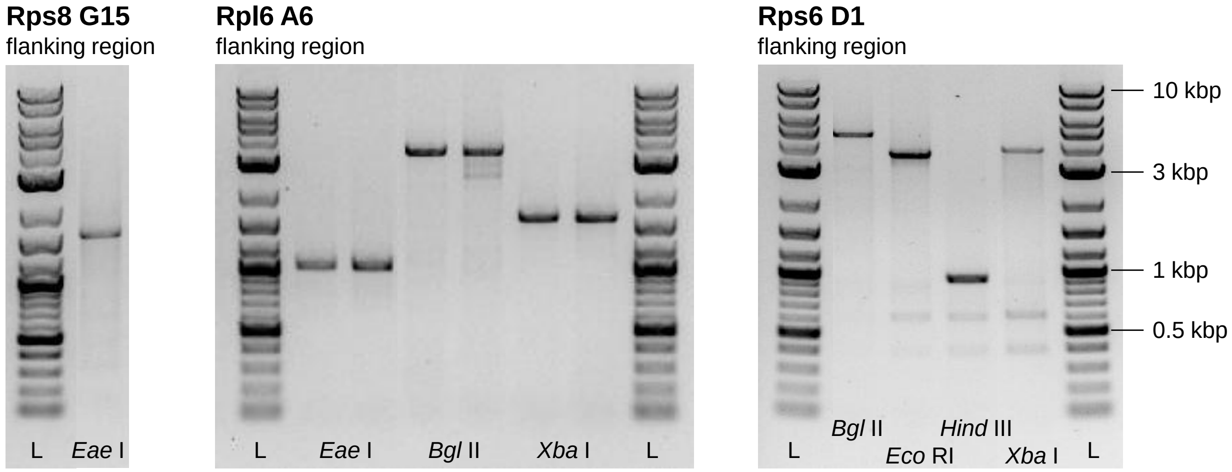


Figure 5-6: Identified 5' and 3' flanking regions of 3 promoter candidates by Inverse PCR
Agarose gel electrophoresis; Lane L: 2-log DNA Ladder (10 kb, 8 kb, 6 kb, 5 kb, 4 kb, 3 kb, 2 kb, 1.5 kb, 1.2 kb, 1 kb, 0.9 kb, 0.8 kb, 0.7 kb, 0.6 kb, 0.5 kb, 0.4 kb, 0.3 kb, 0.2 kb, and 0.1 kb); Remaining lanes: Various restriction endonucleases used to generate self-ligated templates for Inverse PCR led to fragments of different size.

Subsequently, PCR amplifications were performed directly from CHO genomic DNA using primers specific to the newly-discovered flanking regions in order to get the complete section of the genome in the correct order. Obtained PCR products were loaded on a 1% agarose gel and cut out bands were purified. Resulting PCR fragments were cloned into the multiple coning site of the pGL3-Basic reporter vector and the inserts were fully sequenced. All of the three new sequences contained the sequence of the corresponding initially used fragment confirming the correct functionality of the experiment.

5.3 *In silico* sequence analyses

5.3.1 Sequence comparison

The elongated sequences obtained from Inverse PCR were aligned against the genomes of the house mouse (*Mus musculus*), brown rat (*Rattus norvegicus*), and human (*Homo sapiens*) using the blastn algorithm of the nucleotide blast.

Inverse PCR applied to Rps6 D1 revealed a new sequence of 3689 bp covering 980 bp upstream and 2240 bp downstream of initial sequence. The newly identified downstream region includes the complete coding region of *Rps6* which enabled the clear assignment to the *Rps6* gene. Alignment against the genome of *Mus musculus* showed that the coding sequence inclusive 5’ and 3’ UTR of identified putative CHO *Rps6* is 85% identical to the corresponding mouse transcript. However, no significant homology could be determined for the uncoding 5’ flanking region and the introns. The same observation could be made by comparison of the putative CHO *Rps6* to *Rattus norvegicus* and human genome. Whereas here the sequence homology of the coding sequence inclusive 5’ and 3’ UTR of CHO *Rps6* is 87% identical to the rat transcript and 82% identical to the human transcript. Another very interesting discovery was made with regard to the gene structure of the identified putative Chinese hamster *Rps6* as it highly diverges from the mouse, human, or rat *Rps6* gene structures (Figure 5-7). The exon/intron structure is quite similar in the mouse, rat, and human, but in the Chinese hamster some introns are missing, are much shorter, or just exist there like the second intron which splits the coding region corresponding to exon 4 in the other three species.

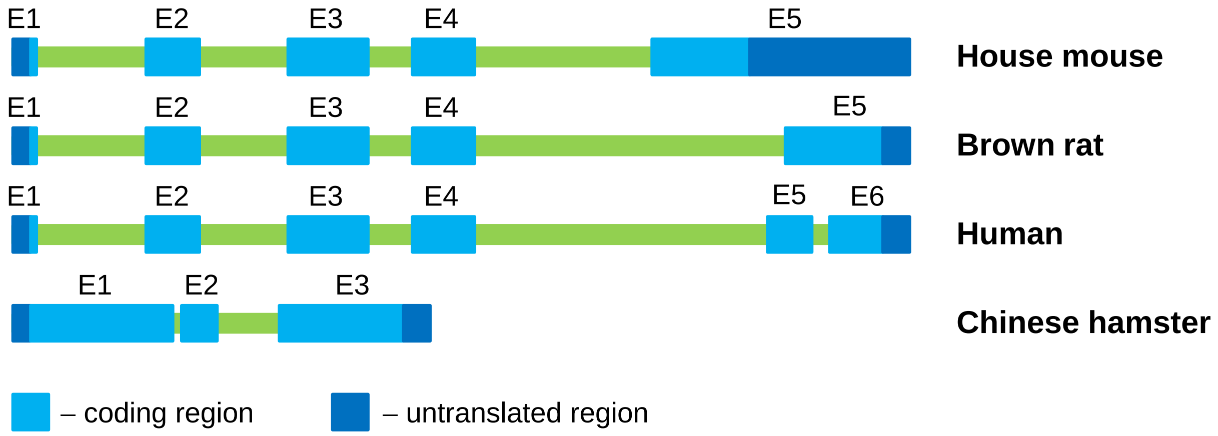


Figure 5-7: Structure of *Rps6* gene in house mouse, brown rat, human, and Chinese hamster
Schematic illustration; E1 – E6: exon 1 – exon 6

This considerable difference in gene structure indicates that the identified sequence might not be the functional *Rps6* gene but a pseudogene thereof.

Pseudogenes are generally defined as non-functional genomic sequences derived from functional genes by retrotransposition or via duplication of genomic DNA [**143**]. They are typically characterized by close similarities to one or more paralogous genes but lack function because of either failure of transcription or translation, or production of protein that has no or a different function. Pseudogenes originated by retrotransposition (or processed pseudogenes) typically lack both 5’ promoter sequence and introns.

Pseudogenes are very common in vertebrate genomes, for instant ribosomal RNA genes can have hundreds of paralogous pseudogenes [**144**].

Although most pseudogenes are considered to be transcriptionally inactive, transcribed ones could have been identified [**145**]. Processed pseudogenes cannot include all the gene regulatory elements of the paralogous functional gene, so they must use other transcriptional elements [**143**]. Promoters of such pseudogenes can be very different and much less efficient [**146**]. However, pseudogenes derived by duplication of genomic DNA may include the transcriptional control elements of the paralogous functional genes [**143**].

Since pseudogenes may show similar characteristics as the paralogous genes, the differentiation from the functional genes can be a difficult endeavor [**147**].

For Rpl6 A6, a new 2063 bp sequence covering 1082 bp upstream and 528 bp downstream of initial sequence was obtained by Inverse PCR. The sequence was aligned against house mouse, brown rat, and human but could not be linked to the *Rpl6* gene of these species. However, this finding was not surprising as Rpl6 A6 was an unspecific PCR product of the Library PCR. Nonetheless, the newly discovered 2 kb sequence was further investigated as it might harbor gene regulatory sequences of an unknown gene.

Applying Inverse PCR to the Rps8 G15 fragment led to a new 1546 bp sequence covering 2 bp upstream and 843 bp downstream of initial sequence. The sequence alignment against house mouse, brown rat, and human could not link the obtained sequence to the *Rps8* gene, but has a strong homology to the 3’ region of the *Stard3* gene. Like for Rpl6 A6, this finding was not unexpected as Rps8 G15 was an unspecific PCR product of the Library PCR. As the sequence seems to be part of the *Stard3* gene, no further investigations have been conducted.

5.3.2 Promoter and TFBSs prediction

Promoter and TFBSs prediction analyses were performed using the sequences Rps6 S1AS3, Rpl6 2kb, and Rps8 A6. The complete sequence data are listed in appendix 8.5.

Putative transcription factor (TF) binding sites were identified using the online program ConSite. The search was conducted for TBP (TATA-binding protein), Sp1, and NF-κB (nuclear factor kappa-light-chain-enhancer of activated B cells) using a TF score cutoff of 85% for Rps6 S1AS3 and Rpl6 2kb and 80% for Rps8 A8. Table 5-1 to Table 5-3 list the results of these analyses.

Table 5-1: Putative transcription factor binding sites of Rps6 S1AS3 predicted by ConSite

Transcription factor	**Sequence**	**Position**	**Score**
TBP	GTATATAAAACAGAA	560 – 574	10.885
Sp1	GGGGCTGGGA	1185 – 1194	9.737

Table 5-2: Putative transcription factor binding sites of Rpl6 2kb predicted by ConSite

Transcription factor	**Sequence**	**Position**	**Score**
Sp1	GGGGTGGGCT	807 – 816	8.129
Sp1	AGGGCTGGGT	965 – 974	9.583
Sp1	GAGGCTGGCT	1109 – 1118	8.274

Table 5-3: Putative transcription factor binding sites of Rps8 A6 predicted by ConSite

Transcription factor	**Sequence**	**Position**	**Score**
TBP	GTATTAATGCAGTGT	154 – 168	8.690
Sp1	AGGGCGGGGC	319 – 328	7.135
Sp1	CGGGCATTGT	993 – 1002	7.198

A second analysis tool used was the promoter prediction program NNPP 2.2 (neural network promoter prediction 2.2). No promoter sequence could be determined for Rpl6 2kb. Promoter predictions for the other two sequences with score cutoff 0.80 are shown in Table 5-4 and Table 5-5. Predicted TSSs are indicated in bold and underlined.

Table 5-4: Predicted promoter sequences of Rps6 S1AS3 by NNPP 2.2

Predicted promoter sequence	**Position**	**Score**
ATTCCAACACTAGGCAAGTATATAAAACAGAAGCAAGTTC**T**CAGA	554 – 604	0.99
AATAGTGGTTAATACCCATTATAAGAACCTAGGTACACAT**T**CCCA	788 – 838	0.82
GCATGTCATTTCCTATGAATAATAATAGGGTGCTTAGTAG**G**AGTT	1292 – 1342	0.85

Table 5-5: Predicted promoter sequences of Rps8 A6 by NNPP 2.2

Predicted promoter sequence	**Position**	**Score**
TATCCCTTTCTCTGTCCTCAAAAATCCTGCCCAAGAAAGG**C**CTTT	755 – 805	0.83

5.4 The putative CHO *Rps6* promoter

5.4.1 Mapping and fragmentation of *Rps6* 5’ flanking region

Based on the data of the *in silico* analyses, fragments of different length were generated in order to identify motifs essential for promoter activity. For this purpose, shorter DNA fragments were designed starting from the full length sequence (see Figure 8-1) by successive removal of 5’ nucleotides. Additionally, the coding region downstream from the ATG start codon was eliminated for all truncated fragments. The map of all different constructs is illustrated in Figure 5-8. Nucleotide positions are indicated relative to the ATG start codon.

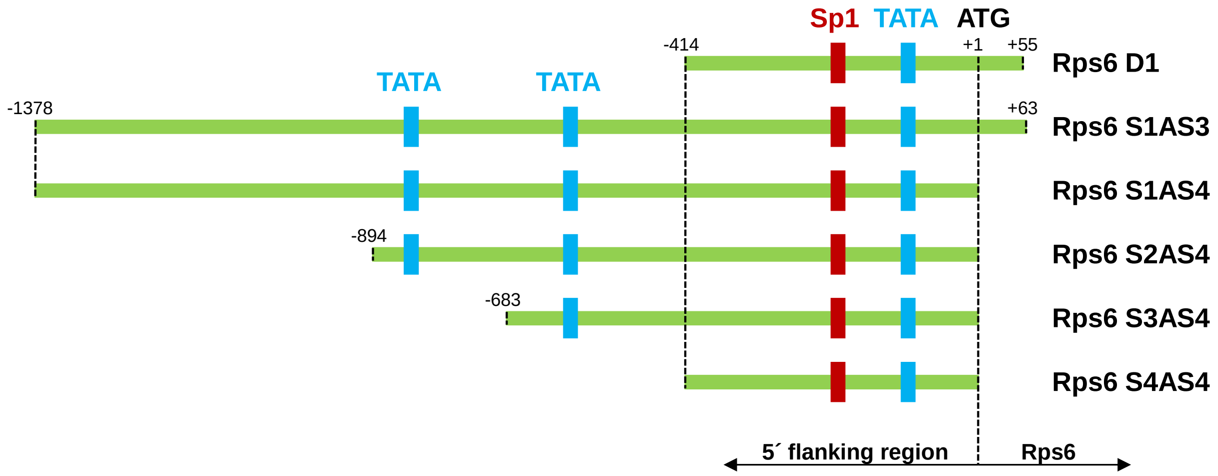


Figure 5-8: Schematic illustration of the 5’ non-coding flanking region of *Rps6*
Original fragment and fragments of various size used in promoter activity assay; TATA boxes were predicted by the online promoter prediction program NNPP 2.2; Sp1 binding site were predicted using the online TFBSs prediction program ConSite; Nucleotide positions are indicated relative to the ATG start codon.

All different promoter constructs were cloned into the promoterless pGL3-Basic reporter vector upstream of the firefly luciferase gene for promoter activity analysis.

5.4.2 Activity of the putative CHO *Rps6* promoter

All promoter constructs were transfected into CHO dhfr$^-$ cells and bioluminescence was measured 48 h post transfection. The promoterless pGL3-Basic reporter vector was used as negative control (-) to detect the background expression level of the luciferase reporter assay.

The pGL3-Promoter reporter vector containing the Simian virus 40 (SV40) promoter served as positive control (+). In order to get quantitatively comparable data for promoter activity, values were normalized to *Renilla* luciferase measurements to account for transfection efficiency and cell number variability. Therefore, the plasmid pRL-SV40 encoding the *Renilla* luciferase gene under control of the SV40 promoter was co-transfected. Promoter activity was calculated as percentage relative to the bioluminescence value measured for the SV40 promoter. Determined promoter activities of all constructs are illustrated in Figure 5-4. The values shown are the average of triplicate samples of a single experiment.

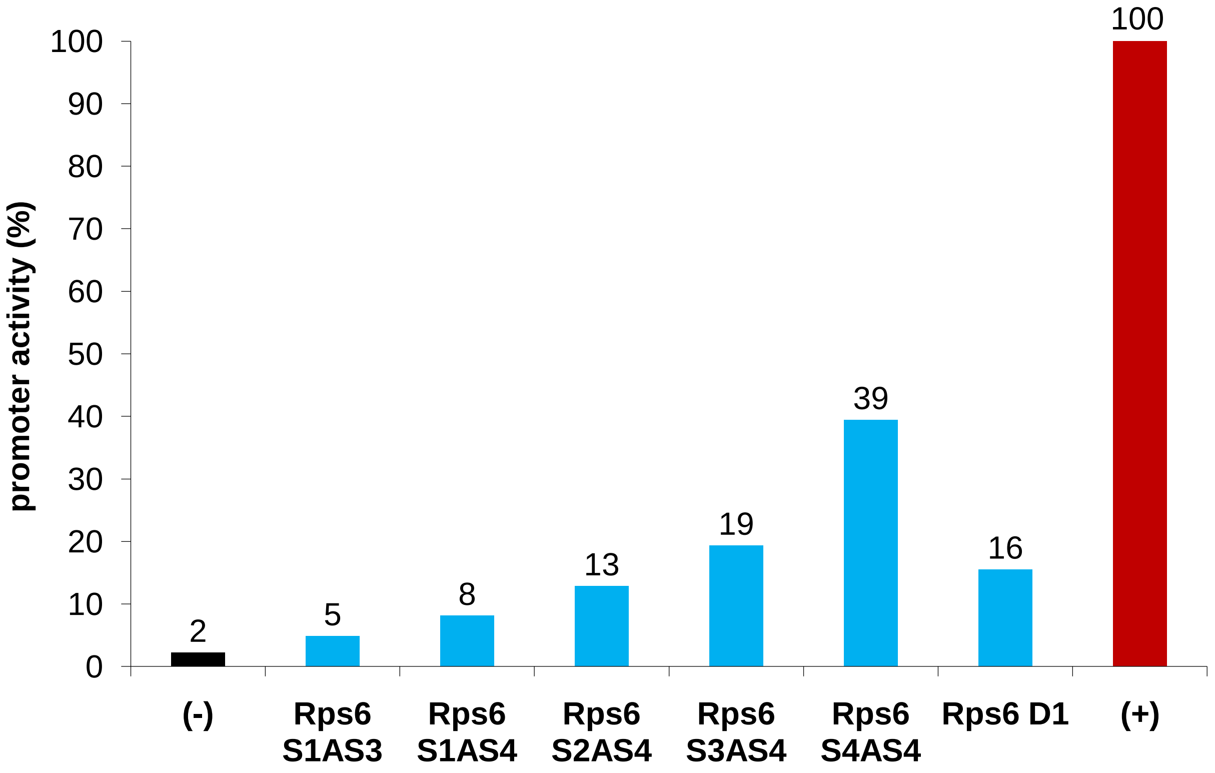


Figure 5-9: Reporter activity assay of *Rps6* promoter fragments
pGL3 reporter vector constructs were transfected into Chinese hamster ovary cells and the promoter activities were determined as percentage relative to the measured bioluminescence value for the SV40 promoter; (-): negative control, promoterless luciferase reporter vector pGL3-Basic; (+): positive control, luciferase reporter vector pGL3-Promoter containing the SV40 promoter

While the original construct Rps6 D1 derived via PCR amplification from the genomic DNA library (Library PCR) showed 16% of the activity of the SV40 promoter in this experiment, the transcriptional activity of the 3’ shortened construct Rps6 S4AS4 was 39%. This indicates that the region downstream of the ATG start codon might has a negative regulatory effect. However, it seems more likely that translation already starts at the additionally introduced ATG start codon resulting in a less active luciferase fusion protein. For an exact clarification of transcriptional activity, analyses on mRNA level would be necessary.

5’ extension could not increase promoter strength, though. Promoter activity decreased continuously from the shortest construct Rps6 S4AS4 (39% of SV40 promoter) to the longest fragment Rps6 S1AS4 which showed just 8% activity relative to the SV40 promoter. These results indicate that the predicted additional TATA boxes at least do not affect transcription positively. Furthermore, the additional 5’ region must comprise *cis*-acting negative regulatory elements (silencers).

The observation that the region downstream of the ATG start codon has a negative influence on determined promoter activity could also been noticed when comparing the constructs Rps6 S1AS4 (8% of SV40 promoter) with the construct Rps6 S1AS3 which had an extended 3’ end and just revealed 5% of the activity of the SV40 promoter.

Rps6 S4AS4 is the shortest fragment that has been tested by now, having a length of 414 bp. As it showed the highest promoter activity, it might be possible that an additional removal of 5’ nucleotides can further boost transcription.

5.5 The putative promoter of a unknown CHO gene

5.5.1 Mapping and fragmentation of the identified new 2 kb sequence

As the ATG start codon of supposed unidentified gene was unknown, truncated fragments were designed starting from the full length sequence Rpl6 A6 2kb (see Figure 8-20) by removing of 5' and 3' nucleotide sections in order to identify the DNA region showing the maximum promoter activity. The map of all different constructs is illustrated in Figure 5-10. Nucleotide positions are indicated relative to the 5' end of Rpl6 A6 2kb.

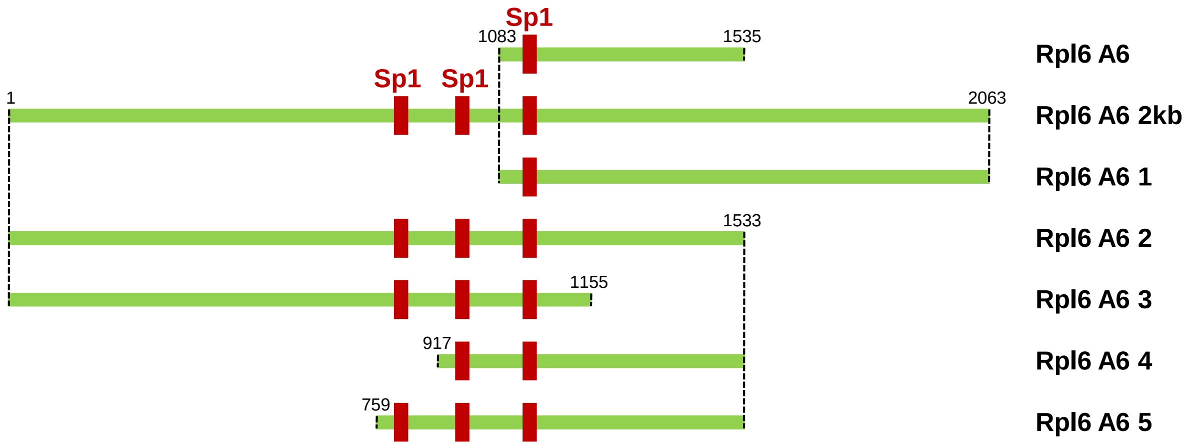


Figure 5-10: Schematic illustration of the new 2 kb CHO sequence
Original fragment and fragments of various size used in promoter activity assay; Sp1 binding sites were predicted using the online TFBSs prediction program ConSite; Nucleotide positions are indicated relative to the 5' end of Rpl6 A6 2kb.

All different promoter constructs were cloned into the promoterless pGL3-Basic reporter vector upstream of the firefly luciferase gene for promoter activity analysis.

5.5.2 Promoter activity of the new 2 kb CHO sequence

All constructs were transfected into CHO dhfr$^-$ cells and bioluminescence was measured 48 h post transfection. The promoterless pGL3-Basic reporter vector was used as negative control (-) to detect the background expression level of the luciferase reporter assay. The pGL3-Promoter reporter vector containing the Simian virus 40 (SV40) promoter served as positive control (+). However, the full length fragment Rpl6 A6 2kb as well as all truncated fragments did not show any significant transcriptional activity.

5.6 The putative CHO *Rps8* promoter

5.6.1 Mapping and fragmentation of *Rps8* 5' flanking region

Based on the identified transcription factor binding sites, fragments of different length were generated in order to identify motifs essential for promoter activity. For this purpose, shorter DNA fragments were designed starting from the full length sequence (see Figure 8-3) by successive removal of 5' nucleotides. Additionally, the coding region downstream from the ATG start codon was eliminated for all truncated fragments. The map of all different constructs is illustrated in Figure 5-11. Nucleotide positions are indicated relative to the ATG start codon.

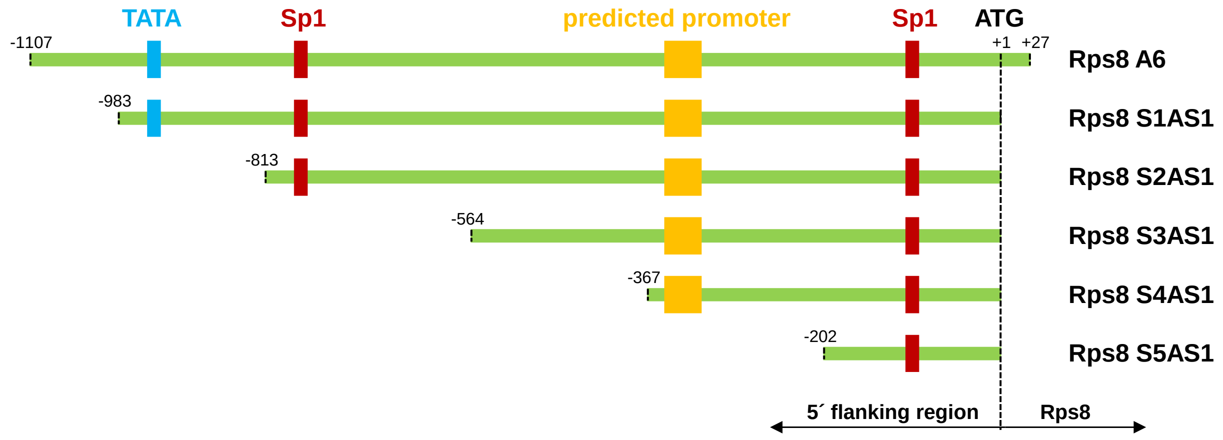


Figure 5-11: Schematic illustration of the 5' non-coding flanking region of *Rps8*
Original fragment and fragments of various size used in promoter activity assay; The predicted promoter was identified by the online promoter prediction program NNPP 2.2; TATA box and Sp1 binding sites were predicted using the online TFBSs prediction program ConSite; Nucleotide positions are indicated relative to the ATG start codon.

All different promoter constructs were cloned into the promoterless pGL3-Basic reporter vector upstream of the firefly luciferase gene for promoter activity analysis.

5.6.2 Activity of the putative *Rps8* promoter

All constructs were transfected into CHO dhfr^{-} cells and bioluminescence was measured 48 h post transfection. The promoterless pGL3-Basic reporter vector was used as negative control (-) to detect the background expression level of the luciferase reporter assay. The pGL3-Promoter reporter vector containing the Simian virus 40 (SV40) promoter served as positive control (+). In order to get significant quantitative data for promoter activity, values were normalized to *Renilla* luciferase measurements to account for transfection efficiency and cell number variability. Therefore, the plasmid pRL-SV40 encoding the *Renilla* luciferase gene under control of the SV40 promoter was co-transfected. Promoter activity was calculated as percentage relative to the bioluminescence value measured for the SV40 promoter. However, the full length fragment Rps8 A6 as well as all truncated fragments did not show any significant transcriptional activity.

The determined lack of promoter activity can be due to various reasons. First, the identified CHO genomic DNA sequence may not at all be related to the *Rps6* gene. This could easily be analyzed by identifying the 3' flanking coding region via Inverse PCR. Second, the identified DNA sequence might correspond to the 5' flanking region of a *Rps6* pseudogene that is not transcribed and so has no gene regulatory properties. Third, the core promoter might be located further upstream of the identified region. Although the majority of TFIID binding sites can be found within 500 bp of the TSSs, they can be located more distantly [**52**]. However, the precise location of the TSS is unknown. Even though the length of 5' UTRs typically ranges between 100 and 200 bp, 5' UTRs of more than 2 kb have been identified in vertebrates [**148**]. Furthermore, the genomic region corresponding to the 5' UTR of the transcript may additionally contain introns. The flanking 5' region could also be investigated by applying the Inverse PCR method as described in chapter 4.2.1.

5.7 Further considerations

Although the Library PCR method offers a quite convenient tool for the discovery of novel endogenous promoters of known genes, the efficiency of this approach is rather moderate. The putative 5’ flanking region for just two of the eight used genes could be identified. Though using longer PCR primers would probably be an improvement, this approach conceals another big drawback, as the libraries cover at most 10% of the CHO genome.

The conducted experiments showed that Inverse PCR is a very reliable method for the discovery of regions that flank a CHO genomic DNA sequence. So this technique could also be directly applied for the identification of 5’ flanking regions of known genes. Using the same principle for identifying the target gene as for the Library PCR approach, primers for Inverse PCR can be design annealing to exon 1 of the relevant gene. The only limitation could be the size of exon 1 since it might be too short to properly design four primers for a nested PCR. However, experiment showed that using just one run of PCR amplification is generally sufficient to generate an adequate amount of a specific PCR product.

6 Conclusion

Mammalian cells became the host system of choice for the production of recombinant proteins used for therapeutic applications mainly because of their excellent properties regarding product secretion and post-translational modification. One of the most widely used mammalian expression systems is the Chinese hamster ovary (CHO) cell line.

The maximization of therapeutic protein yield requires a viable cell biomass and a stable protein expression over an extended period of time. Efforts to optimize mammalian expression system mainly focus on process, media, and cell line improvements. However, the strength and efficiency of transcriptional regulatory sequences have a significant impact on the expression level of a heterologous gene. Hence, expression vectors are prevalently engineered to contain regulatory elements such as promoters, enhancers, introns, or chromatin modifiers. Today, strong viral promoters are most commonly used. They guarantee high-level production, but at the expense of premature activation of cellular apoptotic pathways. This and other undesired effects could be avoided by using cell endogenous transcription regulatory elements as they are under the control of host cell's regulatory network.

Experimental methods for identifying mammalian regulatory sequences are very labor-intensive and so the trend of modern approaches is towards computational analyses of large genomic data sets. However, genome-wide, high-throughput experimental methods as well as computational approaches always rely on whole genome sequence data which are currently not available for all organisms of interest including the Chinese hamster.

In this study, CHO endogenous promoters were directly identified from genomic DNA. For this purpose, CHO genomic plasmid libraries were constructed containing fragments of various lengths which were derived after enzymatic and mechanical fragmentation. This method referred to as Library PCR is based on the amplification of the 5' flanking region of a specific gene using primer pairs specifically binding onto exon 1 of corresponding gene and the vector sequence. The availability of CHO transcript (cDNA) sequence data from the Consortium for CHO Cell Genomics enabled this approach. Library PCR was performed for eight highly expressed CHO genes including *Rps6*, *Rps8*, *Rpl6*, *Rpl27*, *Rpl35*, *Rplp1*, *Tpt1*, and *Jund1*, yielding various fragments ranging from 250 bp up to 1800 bp. These fragments were analyzed concerning their gene regulatory capability using a luciferase reporter assay, and regarding their sequence specificity by sequencing and comparison to corresponding cDNA. Three fragments showed considerable transcriptional activity, however just one of them proved to be a specific product of Library PCR. For this 485 bp fragment which is supposed the cover the 5’ flanking region of the *Rps6* gene, the determined promoter activity was about 25% relative to the SV40 promoter. A second 1188 bp fragment which might comprise the 5’ flanking region of the *Rps8* gene according to sequence analyses showed no significant promoter activity.

In order to further characterize and to potentially boost transcriptional activity of identified genomic fragments, the flanking regions were elucidated via Inverse PCR. Although Inverse PCR is a well-known and widely used method for the *in vitro* amplification of the unknown region that flank a known genomic DNA sequence, the functionality of this technique for CHO genomic DNA was unknown. However, in this study the suitability of Inverse PCR for this purpose could be demonstrated by successfully applying this technique for identifying two known genomic sequences as well as for the correct identification of the 5’ and 3’ flanking regions of all three transcriptionally active fragments.

Applying Inverse PCR for the previously indentified putative promoter region of *Rps6* could reveal more than 2.2 kb of the 3’ flanking region which comprises the complete coding sequence. Alignment against the genomes of *Mus musculus, Rattus norvegicus,* and human showed 85%, 87%, and 82% identity to corresponding transcripts, respectively. However, the exon/intron structure of the putative CHO *Rps6* gene is highly diverging from the mouse, rat, and human *Rps6* gene structure. This indicates that the identified sequence might not be the functional *Rps6* gene but rather a pseudogene thereof. Pseudogenes are generally a tough problem for specific promoter identification as they may show similar characteristics as the paralogous functional gene and so differentiation can be very challenging.

Beside the 3’ flanking region, almost 1 kb upstream of the previously indentified putative promoter region of *Rps6* could be discovered via Inverse PCR. Based on the full-length sequence several truncated mutants were generated and the transcriptional activity analyzed using a luciferase reporter assay. Unfortunately, 5’ extension could not increase promoter strength. In fact, promoter activity continuously decreased from the shortest to the longest construct indicating the existence of silencer elements in this region. However, elimination of the 3’ region downstream of the ATG start codon showed a more than 2-fold enhancement of observed promoter strength.

Same analyses conducted for the previously identified putative *Rps8* promoter region as well as for a putative promoter region of an unknown gene did not lead to an enhanced promoter activity.

Overall, the Library PCR approach in combination with Inverse PCR offers a very convenient tool for the discovery of new endogenous promoters even though only regulatory elements of known gene sequences can be identified. However, this enables the direct targeting of highly expressed genes in order to increase the chance of finding strong *cis*-regulatory elements.

In addition, I suggest an approach that applies the same idea but uses Inverse PCR directly for identifying 5’ flanking regions of known genes. This should increase efficiency which was shown to be rather moderate in case of Library PCR mainly due to the poor sequence coverage of genomic plasmid libraries.

7 References

1 Werner, R. G., Noé, W., Kopp, K., and Schlüter, M. Appropriate mammalian expression systems for biopharmaceuticals. *Arzneimittelforschung*, 48 (1998), 870-880.

2 Coco-Martin, J. M. Mammalian Expression of Therapeutic Proteins. *BioProcess International*, 10 (2004), 32-40.

3 Wurm, F. M. Production of recombinant protein therapeutics in cultivated mammalian cells. *Nat. Biotechnol.*, 22 (2004), 1393-1398.

4 Barnes, L. M., Bentley, C. M., and Dickson, A. J. Advances in animal cell recombinant protein production: GS-NS0 expression system. *Cytotechnology*, 32 (2000), 109-123.

5 Wurm, F. M. and Bernard, A. R. Large-scale transient expression in mammalian cells for recombinant protein production. *Curr. Opin. Biotechnol.*, 10 (1999), 156-159.

6 Jones, D., Kroos, N., Anema, R. et al. High-level expression of recombinant IgG in the human cell line PER.C6. *Biotechnol. Prog.*, 19 (2003), 163-168.

7 Barnes, L. M. and Dickson, A. J. Mammalian cell factories for efficient and stable protein expression. *Curr. Opin. Biotechnol.*, 17 (2006), 381-386.

8 Ludwig, D. L. Mammalian Expression Cassette Engineering for High-Level Protein Production. *BioProcess International*, 4 (2006), 14-23.

9 Butler, J. E. F. and Kadonaga, J. T. The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes & Dev.*, 16 (2002), 2583-2592.

10 Smale, S. T. and Kadonaga, J. T. The RNA polymerase II core promoter. *Annu. Rev. Biochem.*, 72 (2003), 449-479.

11 Makrides, S. C. Components of Vectors for Gene Transfer and Expression in Mammalian Cells. *Protein Exp. Purif.*, 17 (1999), 183-202.

12 James, R. I., Elton, J. P., Todd, P., and Kompala, D. S. Engineering CHO Cells to Overexpress a Secreted Reporter Protein upon Induction from Mouse Mammary Tumor Virus Promoter. *Biotech. Bioeng.*, 67 (2000), 134-140.

13 Running Deer, J. and Allison, D. S. High-Level Expression of Proteins in Mammalian Cells Using Transcription Regulatory Sequences from Chinese Hamster EF-1α Gene. *Biotechnol. Prog.*, 20 (2004), 880-889.

14 Jayapal, K. P., Wlaschin, K. F., Yap, M. G. S., and Hu, W-S. Recombinant Protein Therapeutics from CHO Cells - 20 Years and Counting. *Chem. Eng. Prog.*, 103 (2007), 40-47.

15 Yerganian, G. Cytogenetic possibilities with the Chinese hamster, Cricetulus barabensis griseus. *Genetics*, 37 (1952), 638.

16 Tjio, J. H. and Puck, T. T. Genetic of somatic mammalian cells. II. Chromosomal constitution of cells in tissue culture. *J. Exp. Med.*, 108 (1958), 259-268.

17 Urlaub, G. and Chasin, L. A. Isolation of Chinese hamster cell mutants deficient in dihydrofolate reductase activity. *Proc. Natl. Acad. Sci.*, 77 (1980), 4216-4220.

18 Wiebe, M., Becker, F., Lazar, L. et al. A multifaceted approach to assure that recombinant tPA is free of adventitious virus. In *Advances in Animal Cell Biology and Technology for Bioprocesses*. Butterworth-Heinemann, London, 1989.

19 Seth, G., Hossler, P., Yee, J. C., and Hu, W.-S. Engineering Cells for Cell Culture Bioprocessing - Physiological Fundamentals. *Adv. Biochem. Eng. Biotechnol.*, 101 (2006), 119-164.

20 Reitzer, L. J., Wice, B. M., and Kennell, D. Evidence that glutamine, not sugar, is the major energy source for cultured HeLa cells. *J. Biol. Chem.*, 254 (1979), 2669-2676.

21 Jeong, D.-W., Cho, I. T., Kim, T. S., Bae, G. W., Kim, I.-H., and Kim, I. Y. Effects of lactate dehydrogenase suppression and glycerol-3-phosphate dehydrogenase overexpression on cellular metabolism. *Mol. Cell Biochem.*, 284 (2006), 1-8.

22 Wlaschin, K. F. and Hu, W.-S. Engineering cell metabolism for high-density cell culture via manipulation of sugar transport. *J. Biotechnol.*, 131 (2007), 168-176.

23 Cotter, T. G. and Al-Rubeai, M. Cell death (apoptosis) in cell culture systems. *Trends Biotechnol.*, 13 (1995), 150-155.

24 Goswami, J., Sinskey, A. J., Steller, H., Stephanopoulos, G. N., and Wang, D. I. C. Apoptosis in Batch Cultures of Chinese Hamster Ovary Cells. *Biotechnol Bioeng.*, 62 (1999), 632-640.

25 Meents, H., Enenkel, B., Eppenberger, H. M., Werner, R. G., and Fussenegger, M. Impact of Coexpression and Coamplification of sICAM and Antiapoptosis Determinants bcl-2/bcl-xL on Productivity, Cell Survival, and Mitochondria Number in CHO-DG44 Grown in Suspension and Serum-Free Media. *Biotechnol. Bioeng.*, 80 (2002), 706-716.

26 Wong, D. C. F., Wong, K. T. K., Lee, Y. Y., Morin, P. N., Heng, C. K., and Yap, M. G. S. Trannscriptional Profiling of Apoptotic Pathways in Batch and Fed-Batch CHO Cell Cultures. *Biotechnol. Bioeng.*, 94 (2006), 373-382.

27 Wu, S.-C. RNA interference technology to improve recombinant protein production in Chinese hamster ovary cells. *Biotechnol. Adv.*, 27 (2009), 417-422.

28 Wlaschin, K. F. and Hu, W.-S. A Scaffold for the Chinese Hamster Genome. *Biotechnol. Bioeng.*, 98 (2007), 429-439.

29 Wlaschin, K. F., Nissom, P. M., de Leon Gatti, M. et al. EST Sequencing for Gene Discovery in Chinese Hamster Ovary Cells. *Biotechnol. Bioeng.*, 91 (2005), 592-606.

30 CONSORTIUM FOR CHINESE HAMSTER OVARY CELL GENOMICS. Website: http://hugroup.cems.umn.edu/CHO/cho_index.html (2007).

31 Kantardjieff, A., Nissom, P. M., Chuab, S. H. et al. Developing genomic platforms for Chinese hamster ovary cells. *Biotechnol. Adv.*, 27 (2009), 1028-1035.

32 Wlaschin, K. F., Seth, G., and Hu, W.-S. Toward genomic cell culture engineering. *Cytotechnology*, 50 (2006), 121-140.

33 Omasa, T., Cao, Y., Park, J. Y. et al. Bacterial Artifical Chromosome Library for Genome-Wide Analysis of Chinese Hamster Ovary Cells. *Biotechnol. Bioeng.*, 104 (2009), 986-994.

34 Dolezel, J., Bartos, J., Voglmayr, H., and Greilhuber, J. Nuclear DNA content and genome size of trout and human. *Cytometry Part A*, 51 (2003), 127-128.

35 Fuda, N. J., Ardehali, M. B., and Lis, J. T. Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature*, 461 (2009), 186-192.

36 Maston, G. A., Evans, S. K., and Green, M. R. Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, 7 (2006), 29-59.

37 Juven-Gershon, T., Hsu, J.-Y., Theisen, J. W. M., and Kadonaga, J. T. The RNA polymerase II core promoter - the gateway to transcription. *Curr. Opin. Cell Biol.*, 20 (2008), 253-259.

38 Juven-Gershon, T. and Kadonaga, J. T. Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev. Biol.* (2009).

39 Carninci, P., Sandelin, A., Lenhard, B. et al. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.*, 38 (2006), 626-635.

40 Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y., and Hume, D. A. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat. Rev. Genet.*, 8 (2007), 424-436.

41 Thomas, M. C. and Chiang, C.-M. The General Transcription Machinery and General Cofactors. *Crit. Rev. Biochem. Mol. Biol.*, 41 (2006), 105-178.

42 Lewis, B. A., Sims III, R. J., Lane, W. S., and Reinberg, D. Functional characterization of core promoter elements: DPE-specific transcription requires the protein kinase CK2 and the PC4 coactivator. *Mol. Cell*, 18 (2005), 471-481.

43 Cler, E., Papai, G., Schultz, P., and Davidson, I. Recent advances in understanding the structure and function of general transcription factor TFIID. *Cell. Mol. Life Sci.*, 66 (2009), 2123-2134.

44 Deng, W. and Roberts, S. G. E. TFIIB and the regulation of transcription by RNA polymerase II. *Chromosoma*, 116 (2007), 417-429.

45 Conaway, J. W., Florens, L., Sato, S. et al. The mammalian Mediator complex. *FEBS Lett.*, 579 (2005), 904-908.

46 Smale, S. T. and Baltimore, D. The 'initiator' as a transcription control element. *Cell*, 57 (1989), 103-113.

47 FitzGerald, P. C., Sturgill, D., Shyakhtenko, A., Oliver, B., and Vinson, C. Comparative genomics of Drosophila and human core promoters. *Genome Biol.*, 7 (2006), R53.

48 Purnell, B. A., Emanuel, P. A., and Gilmour, D. S. TFIID sequence recognition of the initiator and sequences farther downstream in Drosophila class II genes. *Genes Dev.*, 8 (1994), 830-842.

49 Chalkley, G. E. and Verrijzer, C. P. DNA binding site selection by RNA polymerase II TAFs: A TAF(II)250-TAF(II)150 complex recognizes the initiator. *EMBO J.*, 18 (1999), 4835-4845.

50 Goldberg, M. L. *PhD Thesis: Sequence analysis of Drosophila histone genes*. Standford University, Stanford, California, USA, 1979.

51 Ponjavic, J., Lenhard, B., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., and Sandelin, A. Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters. *Genome Biol.*, 7 (2006), R78.

52 Kim, T. H., Barrera, L. O., Zheng, M. et al. A high-resolution map of active promoters in the human genome. *Nature*, 436 (2005), 876-880.

53 Lagrange, T., Kapanidis, A. N., Tang, H., Reinberg, D., and Ebright, R. H. New core promoter element in RNA polymerase II-dependent transcription: Sequence-specific DNA binding by transcription factor IIB. *Genes Dev.*, 12 (1998), 34-44.

54 Deng, W. and Roberts, S. G. E. A core promoter element downstream of the TATA box that is recognized by TFIIB. *Genes Dev.*, 19 (2005), 2418-2423.

55 Juven-Gershon, T., Hsu, J.-Y., and Kadonaga, J. T. Caudal, a key developmental regulator, is a DPE-specific transcriptional factor. *Genes Dev.*, 22 (2008), 2823-2830.

56 Burke, T. W. and Kadonaga, J. T. Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. *Genes Dev.*, 10 (1996), 711-724.

57 Kutach, A. K. and Kadonaga, J. T. The downstream promoter element DPE appears to be as widely used as the TATA box in Drosophila core promoters. *Mol. Cell. Biol.*, 20 (2000), 4754-4764.

58 Ohler, U., Liao, G. C., Niemann, H., and Rubin, G. M. Computational analysis of core promoters in the Drosophila genome. *Genome Biol.*, 3 (2002), RESEARCH0087.

59 Lim, C. Y., Santoso, B., Boulay, T., Dong, E., Ohler, U., and Kadonaga, J. T. The MTE, a new core promoter element for transcription by RNA poymerase II. *Genes Dev.*, 18 (2004), 1606-1617.

60 Lewis, B. A., Kim, T.-K., and Orkin, S. H. A downstream element in the human β-globin promoter: Evidence of extended sequence-specific transcription factor IID contacts. *Proc. Natl. Acad. Sci. USA*, 97 (2000), 7172-7177.

61 Lee, D.-H., Gershenzon, N., Gupta, M., Ioshikhes, I. P., Reinberg, D., and Lewis, B. A. Functional characterization of core promoter elements: The downstream core element is recognized by TAF1. *Mol. Cell. Biol.*, 25 (2005), 9674-9686.

62 Tokusumi, Y., Ma, Y., Song, X., Jacobson, R. H., and Takada, S. The new core promoter element XCPE1 (X core promoter element 1) directs activator-, mediator-, and TATA-binding protein-dependent but TFIID-independent RNA polymerase II transcription from TATA-less promoters. *Mol. Cell. Biol.*, 27 (2007), 1844-1858.

63 Anish, R., Hossain, M. B., Jacobson, R. H., and Takada, S. Chracterization of Transcription from TATA-Less Promoters: Identificatioin of a New Core Promoter Element XCPE2 and Analysis of Factor Requirements. *PLoS ONE*, 4 (2009), e5103.

64 Bird, A. P. CpG-rich islands and the function of DNA methylation. *Nature*, 321 (1986), 209-213.

65 Gardiner-Garden, M. and Frommer, M. CpG islands in vertebrate genomes. *J. Mol. Biol.*, 196 (1987), 261-282.

66 Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.*, 16 (2002), 6-21.

67 Antequera, F. Structure, function and evolution of CpG island promoters. *Cell. Mol. Life Sci.*, 60 (2003), 1647-1658.

68 Saxonov, S., Berg, P., and Brutlag, D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Nat. Acad. Sci. USA*, 103 (2006), 1412-1417.

69 Weber, M., Hellmann, I., Stadler, M. B., Ramos, L., Pääbo, S., Rebhan, M., and Schübeler, D. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.*, 39 (2007), 457-466.

70 Illingworth, R. S. and Bird, A. P. CpG islands - 'A rough guide'. *FEBS Letters*, 583 (2009), 1713-1720.

71 Zhu, J., He, F., Hu, S., and Yu, J. On the nature of human housekeeping genes. *Trends Genet.*, 24 (2008), 481-484.

72 Brandeis, M., Frank, D., Keshet, I., Siegfried, Z., Mendelsohn, M., Nemes, A., Temper, V., Razin, A., and Cedar, H. Sp1 elements protect a CpG island from de novo methylation. *Nature*, 371 (1994), 435-438.

73 Adachi, N. and Lieber, M.R. Bidirectional gene organization: A common architectural feature of the human genome. *Cell*, 109 (2002), 807-809.

74 Levine, M. and Tjian, R. Transcription regulation and animal diversity. *Nature*, 424 (2003), 147-151.

75 Kadonaga, J. T. Regulation of RNA Polymerase II Transcription by Sequence-Specific DNA Binding Factors. *Cell*, 116 (2004), 247-257.

76 Xiao, H., Lis, J. T., Greenblatt, J., and Friesen, J. D. The upstream activator CTF/NF1 and RNA polymerase II share a common element involved in transcriptional activation. *Nucl. Acid. Res.*, 22 (1994), 1966-1973.

77 Maity, S. N. and De Crombrugghe, B. Role of the CCAAT-binding protein CBF/NF-Y in transcription. *Trends Biochem. Sci.*, 23 (1998), 174-178.

78 Calhoun, V. C., Stathopoulos, A., and Levine, M. Promoter-proximal tethering elements regulate enhancer-promoter specificity in the Drosophila Antennapedia complex. *Proc. Natl. Acad. Sci. USA*, 99 (2002), 9243-9247.

79 Mastrangelo, I. A., Courey, A. J., Wall, J. S., Jackson, S. P., and Hough, P. V. C. DNA looping and Sp1 multimer links: A mechanism for transcriptional synergism and enhancement. *Proc. Natl. Acad. Sci. USA*, 88 (1991), 5670-5674.

80 Su, W., Jackson, S., Tjian, R., and Echols, H. DNA looping between sites for transcriptional activation: Self-association of DNA-bound Sp1. *Genes Dev.*, 5 (1991), 820-826.

81 Banerji, J., Rusconi, S., and Schaffner, W. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, 27 (1981), 299-308.

82 Vilar, J. M. G. and Saiz, L. DNA looping in gene regulation: From the assembly of macromolecular complexes to the control of transcriptional noise. *Curr. Opin. Genet. Dev.*, 15 (2005), 136-144.

83 Blackwood, E. M. and Kadonaga, J. T. Going the distance: A current view of enhancer action. *Science*, 281 (1998), 60-63.

84 Ogbourne, S. and Antalis, T. M. Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *Biochem. J.*, 331 (1998), 1-14.

85 Privalsky, M. L. The role of corepressors in transcriptional regulation by nuclear hormone receptors. *Annu. Rev. Physiol.*, 66 (2004), 315-360.

86 Chen, L., Widom, J. Mechanism of transcriptional silencing in yeast. *Cell*, 120 (2005), 37-48.

87 Burgess-Beusse, B., Farrell, C., Gaszner, M. et al. The insulation of genes from external enhancers and silencing chromatin. *Proc. Natl. Acad. Sci. USA*, 99 (2002), 16433-16437.

88 Bell, A. C., West, A. G., and Felsenfeld, G. Insulators and boundaries: Versatile regulatory elements in the eukaryotic genome. *Science*, 291 (2001), 447-450.

89 West, A. G., Gaszner, M., and Felsenfeld, G. Insulators: Many functions, many mechanisms. *Genes Dev.*, 16 (2002), 271-288.

90 West, A. G., Fraser, P. Remote control of gene transcription. *Hum. Mol. Genet.*, 14 (2005), R101-R111.

91 Li, Q., Peterson, K. R., Fang, X., and Stamatoyannopoulos, G. Locus control regions. *Blood*, 100 (2002), 3077-3086.

92 Grosveld, F., van Assendelft, G. B., Greaves, D. R., and Kollias, G. Position-independent, high-level expression of the human β-globin gene in transgenic mice. *Cell*, 51 (1987), 975-985.

93 Liolios, K., Mavromatis, K., Tavernarakis, N., and Kyrpides, N. C. The Genomes On Line Database (GOLD) in 2007: Status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.*, 36 (2008), D475-D479.

94 Crawford, G. E., Holt, I. E., Mullikin, J. C. et al. Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. *Proc. Natl. Acad. Sci. USA*, 101 (2004), 992-997.

95 Pennacchio, L. A. and Rubin, E. M. Genomic strategies to identify mammalian regulatory sequences. *Nat. Rev. Genet.*, 2 (2001), 100-109.

96 Ren, B., Robert, F., Wyrick, J. J. et al. Genome-wide location and function of DNA binding proteins. *Science*, 290 (2000), 2306-2309.

97 Loh, Y.-H., Wu, Q., Chew, J.-L. et al. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.*, 38 (2006), 431-440.

98 Werner, T. The state of the art of mammalian promoter recognition. *Breif. Bioinform.*, 4 (2003), 22-30.

99 Schaefer, B. C. Revolutions in rapid amplification of cDNA ends: New strategies for polymerase chain reaction cloning of full-length cDNA ends. *Anal. Biochem.*, 227 (1995), 255-273.

100 Carninci, P., Waki, K., Shiraki, T. et al. Targeting a complex transcriptome: The construction of the mouse full-length cDNA encyclopedia. *Genome Res.*, 13 (2003), 1273-1289.

101 Shiraki, T., Kondo, S., Katayama, S. et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. USA*, 100 (2003), 15776-15781.

102 Hashimoto, S.-I., Suzuki, Y., Kasai, Y., Morohoshi, K., Yamada, T., Sese, J., Morishita, S., Sugano, S., and Matsushima, K. 5′-end SAGE for the analysis of transcriptional start sites. *Nat. Biotechnol.*, 22 (2004), 1146-1149.

103 Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. Serial analysis of gene expression. *Science*, 270 (1995), 484-487.

104 Ng, P., Wei, C.-L., Sung, W.-K. et al. Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat. Methods*, 2 (2005), 105-111.

105 Han, B. and Zhang, J.-T. Regulation of gene expression by internal ribosome entry sites or cryptic promoters: The eIF4G story. *Mol. Cell. Biol.*, 22 (2002), 7372-7384.

106 Ioshikhes, I. P. and Zhang, M. Q. Large-scale human promoter mapping using CpG islands. *Nat. Genet.*, 26 (2000), 61-63.

107 Solovyev, V. V. and Shahmuradov, I. A. PromH: Promoters identification using orthologous genomic sequences. *Nucleic Acids Res.*, 31 (2003), 3540-3545.

108 Down, T. A. and Hubbard, T. J. P. Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res.*, 12 (2002), 458-461.

109 Abeel, T., Saeys, Y., Bonnet, E., Rouzé, P., and Van De Peer, Y. Generic eukaryotic core promoter prediction using structural features of DNA. *Genome Res.*, 18 (2008), 310-323.

110 Sonnenburg, S., Zien, A., and Rätsch, G. ARTS: Accurate recognition of transcription starts in human. *Bioinformatics*, 22 (2006), e472-e480.

111 Zhao, X., Xuan, Z., and Zhang, M. Q. Boosting with stumps for predicting transcription start sites. *Genome Biol.*, 8 (2007), R17.

112 Davuluri, R. V., Grosse, I., and Zhang, M. Q. Computational identification of promoters and first exons in the human genome. *Nat. Genet.*, 29 (2001), 412-417.

113 Ohler, U., Niemann, H., Liao, G.-C., and Rubin, G. M. Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. *Bioinformatics*, 17 (2001), S199-S206.

114 Reese, M. G. Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome. *Comput. Chem.*, 26 (2001), 51-56.

115 Knudsen, S. Promoter2.0: For the recognition of PolII promoter sequences. *Bioinformatics*, 15 (1999), 356-361.

116 Abeel, T., Saeys, Y., Rouzé, P., Van de Peer, Y. ProSOM: Core promoter prediction based on unsupervised clustering of DNA physical profiles. *Bioinformatics*, 24 (2008), i24-i31.

117 Matys, V., Fricke, E., Geffers, R. et al. TRANSFAC®: Transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, 31 (2003), 374-378.

118 Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W., and Lenhard, B. JASPAR: An open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, 32 (2004), D91-D94.

119 Cartharius, K., Frech, K., Grote, K. et al. MatInspector and beyond: Promoter analysis based on transcription factor binding sites. *Bioinformatics*, 21 (2005), 2933-2942.

120 Kel, A. E., Gößling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O. V., and Wingender, E. MATCH™: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.*, 31 (2003), 3576-3579.

121 Tagle, D. A., Koop, B. F., Goodman, M., Slightom, J. L., Hess, D. L., and Jones, R. T. Embryonic ε and γ genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.*, 203 (1988), 439-455.

122 Dermitzakis, E. T. and Clark, A. G. Evolution of transcription factor binding sites in mammalian gene regulatory regions: Conservation and turnover. *Mol. Biol. Evol.*, 19 (2002), 1114-1121.

123 Blanchette, M., Tompa, M. FootPrinter: A program designed for phylogenetic footprinting. *Nucleic Acids Res.*, 31 (2003), 3840-3842.

124 Loots, G. G and Ovcharenko, I. rVISTA 2.0: Evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res.*, 32 (2003), W217-W221.

125 Sandelin, A., Wasserman, W. W., and Lenhard, B. ConSite: Web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res.*, 32 (2004), W249-W252.

126 Fang, F. and Blanchette, M. FootPrinter3: Phylogenetic footprinting in partially alignable sequences. *Nucleic Acids Res.*, 34 (2006), W617-W620.

127 Abeel, T., Van de Peer, Y., and Saeys, Y. Towards a gold standard for promoter prediction evaluation. *Bioinformatics*, 25 (2009), i313-i320.

128 Baumann, M. *Master Thesis: Identification and Characterisation of potential CHO promoter sequences*. University of Natural Resources and Applied Life Sciences, Vienna, Austria, 2008.

129 PROMEGA. Technical Manual: pGL3 Luciferase Reporter Vector (2007).

130 PROMEGA. Technical Bulletin: pRL-SV40 Vector (2007).

131 Sambrook, J. and Russel, D. W. *Molecular cloning: a laboratory manual*. Cold Spring Harbor Laboratory Press, USA, 2001.

132 Rozen, S. and Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, 132 (2000), 365-386.

133 PROMEGA. Technical Manual: Dual-Glo Luciferase Assay System (2009).

134 Davison, P. F. The effect of hydrodynamic shers on the deoxyribonucleic acid from T2 and T4 bacteriophages. *Proc. Natl. Acad. Sci. USA*, 45 (1959), 1560-1568.

135 Davison, P. F. Sedimentation of deoxyribonucleic acid isolated under low hydrodynamic shear. *Nature*, 185 (1960), 918-920.

136 Schriefer, L. A., Gebauer, B. K., Qui, L. Q. Q., Waterston, R. H., and Wilson, R. K. Low pressure DNA shearing: a method for random DNA sequence analysis. *Nucleic Acids Res.*, 18 (1990), 7455-7456.

137 Okpodu, C. M., Robertson, D., Boss, W. F., Togasaki, R. K., and Surzycki, S. J. Rapid isolation of nuclei from carrot suspension culture cells using a BioNebulizer. *Biotechniques*, 16 (1994), 154-159.

138 Deininger, P. L. Random subcloning of sonicated DNA: application to shotgun DNA sequence analysis. *Anal Biochem.*, 129 (1983), 216-223.

139 Hershey, A. D. Molecular homogeneity of the deoxyribonucleic acid of phage T2. *BioProcess International*, 2 (1960), 143-152.

140 Anderson, S. Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Res.*, 9 (1981), 3015-3027.

141 Ochman, H., Gerber, A. S., and Hartl, D. L. Genetic Applications of an Inverse Polymerase Chain Reaction. *Genetics*, 120 (1988), 621 - 623.

142 Khambata-Ford, S., Liu, Y., Gleason, C., Dickson, M., Altman, R. B., Batzoglou, S., and Myers, R. M. Identification of Promoter Regions in the Human Genome by Using a Retroviral Plasmid Library-Based Functional Reporter Gene Assay. *Genome Res.*, 13 (2003), 1765-1774.

143 Mighell, A. J., Smith, N. R., Robinson, P. A., and Markham, A. F. Vertebrate pseudogenes. *FEBS Letters*, 468 (2000), 109-114.

144 Frederiksen, S., Cao, H., Lomholt, B., Levan, G., and Hallenberg, C. The rat 5S rRNA bona fide gene repeat maps to chromosome 19q12 → qter and the pseudogene repeat maps to 12q12. *Cytogenet. Cell Genet.*, 76 (1997), 101-106.

145 Berger, I. R., Buschbeck, M., Bange, J., and Ullrich, A. Identification of a transcriptionally active hVH-5 pseudogene on 10q22.2. *Cancer Genet. Cytogenet.*, 159 (2005), 155-159.

146 Nishimura, K., Liisananti, M., Muta, Y., Kashiwagi, K., Shirahata, A., Jänne, M., Kankare, K., Jänne, O. A., and Igarashi, K. Structure and activity of mouse S-adenosylmethionine decarboxylase gene promoters and properties of the encoded proteins. *Biochem. J.*, 332 (1998), 651-659.

147 Rouchka, E. C. and Cha, I. E. Current trends in pseudogene detection and characterization. *Current Bioinformatics*, 4 (2009), 112-119.

148 Mignone, F., Gissi, C., Liuni, S., and Pesole, G. Untranslated regions of mRNAs. *Genome Biol.*, 3 (2002), reviews0004.1-0004.10.

8 Appendices

8.1 List of figures

8.2 List of tables

8.3 Abbreviations

AMP	adenosine 5'-monophosphate
Amp^r	ampicillin resistance gene
AP	antisense primer
ATP	adenosine 5'-triphosphate
BAC	bacterial artificial chromosome
bcl-2	B-cell lymphoma 2 gene
BHK	baby hamster kidney
BLAST	Basic Local Alignment Search Tool
bp	base pair(s)
BRE	TFIIB recognition element
BRE^d	downstream BRE
BRE^u	upstream BRE
BSA	bovine serum albumin
C/EBP	CCAAT-enhancer-binding protein
CAGE	cap analysis of gene expression
CAT	chloramphenicol acetyltransferase
CBF	CCAAT-box-binding factor
cDNA	complementary DNA
CGI	CpG island
CHEF1	Chinese hamster EF-1α
ChIP	chromatin immunoprecipitation
CHO	Chinese hamster ovary
CMV	cytomegalovirus
CTF	CCAAT-binding transcription factor
DCE	downstream core element
ddH$_2$O	double distilled water
DHFR	dihydrofolate reductase
dhfr$^-$	dihydrofolate reductase deficient
DMEM	Dulbecco’s Modified Eagle Medium
DNA	deoxyribonucleic acid

dNTP	deoxynucleoside triphosphates
DPE	downstream promoter element
E. coli	*Escherichia coli*
EDTA	ethylenediaminetetraacetic
EF-1α	elongation factor-1α
ER	endoplasmatic reticulum
EST	expressed sequence tag
g	gravity
Gb	giga base pairs
gDNA	genomic DNA
GFP	green fluorescent protein
GIS	gene identification signature
GPDH	glycerol-3-phosphate dehydrogenase
GTF	general transcription factor
HBV	hepatitis B virus
HEK-293	human embryo kidney 293
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HIV	human immunodeficiency virus
Inr	initiator
Jund1	jun proto-oncogene related gene d1
kb	kilo base pairs
K_m	Michaelis-Menten constant
LB medium	Luria-Bertani medium
LCR	locus control region
LDH-A	lactate dehydrogenase A
LTR	long terminal repeat
luc	luciferase gene
M	molar [mol l^{-1}]
MAR	matrix attachment region
MCS	multiple cloning site
MMTV	mouse mammary tumor virus
mRNA	messenger RNA

MTE	motif ten element
MTX	methotrexate
NF-I	nuclear factor I
NF-κB	nuclear factor kappa-light-chain-enhancer of activated B cells
NF-Y	nuclear factor Y
NNPP	neural network promoter prediction
OD	optical density
ori	origin of replication
PBS	phosphate buffered saline
PCR	polymerase chain reaction
PET	paired-end ditag
PIC	preinitiation complex
Pol II	RNA polymerase II
polyA	polyadenylation
PP$_i$	pyrophosphate
PPP	promoter prediction program
PWM	position weight matrix
RACE	rapid amplification of cDNA ends
RNA	ribonucleic acid
RNAi	RNA interference
Rpl27	ribosomal protein L27 gene
Rpl35	ribosomal protein L35 gene
Rpl6	ribosomal protein L6 gene
Rplp1	ribosomal protein large P1 gene
rpm	revolutions per minute
Rps6	ribosomal protein S6 gene
Rps8	ribosomal protein S8 gene
RSV	Rous sarcoma virus
r-tPA	recombinant tissue plasminogen activator
SAGE	serial analysis of gene expression
SAR	scaffold attachment region
SDS	sodium dodecyl sulfate

SOC medium	Super Optimal Catabolite medium
SP	sense primer
Stard3	StAR-related lipid transfer (START) domain containing 3 gene
SV40	Simian virus 40
TAE	TRIS-acetate-EDTA
TAF	TBP-associated factor
TBA	TATA-binding protein
TF	transcription factor
TFBS	transcription factor binding site
TFIIA	transcription factor for RNA polymerase II A
T_m	melting temperature
Tpt1	tumor protein translationally-controlled 1 gene
TRIS	Tris(hydroxymethyl)-aminomethan
TSS	transcription start site
U	unit
UPR	unfolded protein response
UTR	untranslated region
UV	ultraviolet
V	volt
w/v	weight/volume
XCPE1	X core promoter element 1
XCPE2	X core promoter element 2

8.4 IUPAC nucleic acid codes

A	Adenine
C	Cytosine
G	Guanine
T	Thymine
U	Uracil
R	Purine (A or G)
Y	Pyrimidine (C, T, or U)
M	C or A
K	T, U, or G
W	T, U, or A
S	C or G
B	C, T, U, or G (not A)
D	A, T, U, or G (not C)
H	A, T, U, or C (not G)
V	A, C, or G (not T or U)
N	Any base (A, C, G, T, or U)

8.5 Sequencing data

8.5.1 Putative *Rps6* 5’ flanking region (Rps6 S1AS3)

```
   1 TCACTGGATC AGCACAATCT TACATGCAGA TGAGAATACA GAATGTGGAA
  51 TAAGCATATA GAAAGAAGAA AGGTCTTGTC TGTGTTGGTG GTGCACACCT
 101 TTAATCTCAG AACTTGGGAG GCAGGAGGCA GAGGCAGAGG CAGGCAGATT
 151 TCCAAGTTCT AGGCCAGCCT GGTCTACAAA GTAAGTTCCA GGACAGCCAG
 201 GGCTGAATGT CTTGAAAAAC AAAACAAAAT AAAAATAAAA ATAAAAAAGA
 251 TGAAAAGTCT GTAATAATTT TTAACATAGA GTCATCAATC TTTTTTGTGA
 301 AGTATATTTG GACAAAAAAG TCTCTGTAGT GACACATTAT TCATGATTTA
 351 TTAAAGCTTG CCTGATTCAG AAAGCAGAGT CAGCCACAAG CTATAGAGGC
 401 CAGGCACACA GCTTTAATCC CAGGAGCCAG AAGCTTTAAT CCTAGGACCC
 451 AGGATTAAAA GATAGATCTC TGTGAGCCCA AGGCCACCCA GAAGTACACA
 501 AGAGTGAATC AGTCTAAGAG AGAAGTATAG CTCACACCTT TAATTCCAAC
 551 ACTAGGCAAG TATATAAAAC AGAAGCAAGT TCTCAGAGAA GCATTTGTTC
 601 TCCAGCCACA CTGAGAAGAG GCAGCAGTTT GAGACTTGGT GAAGACCTCG
 651 TTTGGGATCA GCCCTTTTAG TTTGAGCTAG AGGTGAGAGC TAGTGGCTAC
 701 CAGGGTATGT TCGATAAGAA GGCCTGAACA GAGAGAAGCA TGTGTTCTAG
 751 TTTACTGTAG GACAGTCCAA CTGCAGAATA GTGGTTAATA CCCATTATAA
 801 GAACCTAGGT ACACATTCCC ACAGCCAAAC CCTCCCCTCA ACCCTCGTTG
 851 ATCTTGCTGT TCTTCAAGTG TTTTTCCACC TTAGGACCTT TGCACTAGTG
 901 TTTCTCTATA CTCAGATTGC TCTTCCTCAA GGTTTATGGA TGACCTTGTT
 951 CTTCTTCTCA GGTCTCTGGC TGGTTTTCAC TGTGGTAGGA ATGTTCTGCA
1001 TGGACTGGTG TTTCCAGGTT CACATGTTTT ATTCAAATAG TTGAAAAAGT
1051 GCACACAAAT AATGCAAGGG ATTTTAATCA GAAACAAAGT AAAAGTAGAG
1101 TCCATACACT GTAAGGGAGC AATGGGCTCT AAGGAAGCAC ACAGCTCAAG
1151 AATGCTTGGT TTATTCTCTG AGTTTTCCTT TTGTGGGGCT GGGAGAAAGA
1201 AAAGTTCATA ACTAAAGGTA GGGTGAGAGA GATTCTGTTT TTATTGACAG
1251 GCTTGGACTA TGTAAGTCTT TGTTTACTGT GCATGTCATT TCCTATGAAT
1301 AATAATAGGG TGCTTAGTAG GAGTTATCTC ACTCAGGCTC TTTTTCGTGG
1351 CACCTCCTAG GCGGTTGGCT GTGTGAAGAT GAAGCTGAAT TTCTCCTTCC
1401 TGGCCACCAG CTGCCAGAAA CTCATTGAAG TGGACGATGA G
```

Figure 8-1: Sequence of the putative 5’ flanking region of the CHO *Rps6* gene
Full-length fragment Rps6 S1AS3; The ATG start codon is marked in bold and underlined.

8.5.2 New 2 kb sequence (Rpl6 A6 2kb)

```
   1 TGCTGGAGAC CAACTGTAAG GGATGGCAAG CTTGTCAGGG GACATGCAGT
  51 ACTGCCAGGG AGCCCTGCCC ATTCTCCCGC ACTGCCCTGG GCTCCTTAGC
 101 TTACCATTCG GTAAAGGTCA AGGGCTTCTT GGAAGACAGC CTGCAGCCTG
 151 TGCGAGACTG ATCTTATACT GCCCAGCTCA AGGAGTGCTG GGGGCAGGAC
 201 CTCAGCTGGG AGCCCAGGCT CCCCACAGTG TCCAGGGCTG CCCGGCATAT
 251 CTGGGGGTAG AGAAGAGTGA ATGACCAAGA TAGTTTGACT TTCTGGATGG
 301 CTTGAGAGAG GCCAGGTGGC TGGGCTAATG ACTGTTGGCC CTGGACCACT
 351 TTCTTCAGTC CCCAAGGCCT CTCTTCCTCA CAGGGTTCTG CTGGATAGAA
 401 ACTTCTCCTT AAATCATGTT GTTGCCATCC TAAAATTGAA GCTACACTGC
 451 TCTCCCCGCC CCTTTCCCCC TGCCTTATTT CTCTCAGTAA CTTGTATCAC
 501 CTTGTGATAC GTCTGACACT TTGGTCTGGG TATCTCCCCT AGCGTAGGAA
 551 TCAACTCTCT CGTTTGTCCT CTTGTTTGTC ACCATGTCTC ATGTGCCCAG
 601 GACCTAGAAC AAAGCCCAGC ATGTGCTCAT TAACTATGTG AATGAATGAA
 651 TGAACGAATG GATGAATGAG TGAATGAATG AATGAGTGAA TGAATAACAC
 701 AGTCCCTCTT GTGGCTAGGG TGAGCGTCCT ACTTGGCAGA GGAGTCGAGC
 751 GGCACAGGGG ACAGGCTAGG GCTCTCTCTA TCCGTGTCCC ATGGTAGTAC
 801 TCACCAGGGG TGGGCTCCAG GAGAGCAGCG GTACAATTAG GCACCCTAGA
 851 CCTGGTCTCT AGAAGTTGGA GGCTGTTGGG ATGGCCAGGA GGGATGAGGG
 901 GCCCAGAGGC TGAGGTCCTT GGGAGAAACA CAGAGCTGTG GAGTCTTGGG
 951 GTGCTAGCAT CTACAGGGCT GGGTGCTGGG AAGCTGGCTA CTGTGGCTGT
1001 CACACCAGGC TGGACATCCA CAGAGTCCTG AGACCACACA CATGTGATGG
1051 GTGGCTCCAG TGGGGGTGAA GACAGGAGCT GGCCACAGAC ACTGGATAAG
1101 GGCAGTTTGA GGCTGGCTCG GGCCTCATGA TTGCCCCAGG AAGGCATGGC
1151 AGGATCCTGG CCTTCGATGC TAGAAGTGAG AGTCACTGTG GTGGGGATGG
1201 CCTGGGGTTC CTGGCCCAGG GAAGCCAGTT CTCCCATGGA TAAGGGTTTG
1251 TGAAGTGACC TGGATGGCTC AGGTGTCACA AGGCCCTCAG TGTCCCCAAG
1301 AGACATGCTA CGTGATATCT TGGCATGTGA GCTAGTCGTC GTCTCCAGGT
1351 AAGAACATGT AGACGGGGCA CAGACACTCT GATCCAGGCC TGTGATAGGT
1401 ACAGATGTGG GTGTTGGAGG CTTCCTGTCT GTGGGTGGCA CCGAGGAGGA
1451 AGAGACACAG AAGGCCAGGC CTGTGAGGGA AGGAGCTGCA AAGGAAGCAG
1501 ATGGCTCACT TGTACACGAG TCTCAGAGCC CTGCCCTAGG AATGCTTCAC
1551 CATCTGAACG GACAATTCCA TGGCTCAGCT GGACACCCAG GGCTGTTCAT
1601 CTCGGATTGG AGACATTTAC TTGGTAGATA CAAGGCCTCT GGAGGCTCCT
1651 CAGCCTTCTG CCCAGAGAGG ACCTCCAAGC CCAGCTCTGT GGTAAAGACA
1701 AGGCCTCCTG GCTTGCAGCC CTGGGTCCTA GGCGTGACCT CAATACTCAC
1751 ATTGGTCCCA TCTGAGGACA TGTAGCCAGT ACCTGCTCTC AGGGACCCTG
1801 CTCTGGACTG GCCGCATCCT TGTCCTGTAG GCTTGACCTC TGGAGACTTC
1851 CTAAGGTGAA GGGGCAGCCG GGGAGGGAAG CTGGGGAAGA AGCCAGGTGG
1901 AAAAGAGAAG TTCACATAGC CTGACCCGAG GGTCACGGAG GGGCGTGGCT
1951 GAGGAAAGGG AAATTGGGGG TCACTGGCTT CTACCTGGAG GCCTTCTGGA
2001 GTCGTGAGAG AAACTGGGTG GAGATGCTCA GCCGGGGATT GAAGAGATGG
2051 TGGTCAGCCT CTG
```

Figure 8-2: Identified new 2kb CHO genomic sequence
Full-length sequence of the fragment Rpl6 A6 2kb

8.5.3 Putative *Rps8* 5' flanking region (Rps8 A6)

```
   1 ACTTGACAAT TACAGGAACT CATGTATGTA GAGTTCTTAG CACCAAGCCT
  51 GGCCTTTAGT AAGAAATCCA TTGGTGGTAC TATTGATATT CTGATAATGA
 101 TGATCATGAA GAAAACAAAA TTCTCTGAGC AAAAGATATT TGTGAGCCTA
 151 ATAGTATTAA TGCAGTGTTA ATACAAACAG TAGATGAATC ATGCAGCCCA
 201 CATGTTGAGA CAATAGAGGA GCATGAAATG ATTAAATGAG GAAATTCTGC
 251 CCTCTAGTGG CAGAGATTGG ATGCTTATTA AACACCCATA TGCTCATCAA
 301 TTTCCCAGGC AGACTTGAAG GGCGGGGCAT GTGAACAGTT CTGACTCATG
 351 ACATGCTTGC AGAAGGACCA ACCAGTCCTC TAAATTTCCA TAAATCTTCC
 401 CTATGTTAAG CTACTGGAGT TTCATGTTCT GCCTGTTGTG GTACTTGACA
 451 TGAATCATTT TAATGAACAC ACTTGGGTCA CCTCCCTTAC AATCCTTCTT
 501 GCCTCATCCC ACTGGAAAAT TTTGGGGAAA GAGAAGAAAA GATGTCTCTC
 551 ATTGAATTAT ACTGGAAGCA TTATTCAGAT CATAATTTGG CCCAAATGGA
 601 ACCTGAATCT TTCCAAAATG CTCAAATAGA GGAGAGATAG TGAGAGCTGC
 651 AGCCAAAGGA GGCTTCCTGA GTTTCCTGAT GATCAGTCAA CATCCTCAGA
 701 GGAATTTCAC CATCAATGCA CCCTAGAAAA GATGAATGTT TTCAACTATC
 751 CCTTTCTCTG TCCTCAAAAA TCCTGCCCAA GAAAGGCCTT TGCTGTAATG
 801 ATATTGGGGA GATACTTTCA GTGGATTCAA TGAAGTAATG ACAAACAAAA
 851 CTCTATCTTT ATAGTATGAC CAAGAAATCA ACAACAACAA AAAAAAAAAA
 901 AAAAACAGGA AATTGTCAAC AACAGTGTTT TTTGAAAATT CAGCCTAACT
 951 AGGAATTGAA GAAATTCCAA TTTTCTTGAG TCTATTTCAT GGCGGGCATT
1001 GTAATGGGTG ATGACACCGG GACTAGCAAG AATCTGAGAG GCGACACTCT
1051 ACTCACCTTT ACAGCAGGAA ACTAAAGAAG GGAACCTTTG GCAGCCAGCA
1101 CCGAGCAATG AGCATCTCTC GGGACAACTG GCAC
```

Figure 8-3: Sequence of the putative 5' flanking region of the CHO *Rps8* gene
Full-length fragment Rps8 A6; The ATG start codon is marked in bold and underlined.