Genome-wide identification and comprehensive analysis of transcriptional desert regions

Ulf Schaefer

Thesis presented in fulfilment of the requirements for the Degree of *Doctor Philosophiae* at the South African National Bioinformatics Institute, University of the Western Cape

April 2009

Advisor: Prof. Vladimir Bajic

Abstract

The initiation of transcription in mammalian genomes predominatly occurs at 5' promoter regions, however increasingly initiation events have been observed within introns, coding exons and 3' UTRs. Nevertheless there are large segments of mammalian genomes that are not prone to transcription initiation. These locations can be understood to be '*transcription initiation deserts*'. It is challenging and useful to demarcate these segments or locations of the genome. The availability of a huge number of transcript data has provided an opportunity to develop a methodology to predict and annotate these genomic segments.

A comprehensive collection of data for *Homo sapiens* and *Mus musculus*, consisting of CAGE tags and other evidence for the existence of transcription was used to develop a methodology that allows the annotation of locations of mammalian genomes as those that are highly likely to initiate transcription and those that are unlikely to harbour transcription start sites (TSSs). The algorithm allows the recognition of TSSs with 100% sensitivity, which makes it the superior choice over other existing algorithms for promoter prediction for the task of annotating TSS deserts.

98,680 and 113,814 transcription start sites were accurately determined for *Mus musculus* and *Homo sapiens* respectively. The properties of the regions immediately surrounding these TSS locations were used to determine features that distinguish genomic transcription initiation segments from those that are not likely to initiate transcription. The algorithm utilises various constraining properties of features identified in the upstream and downstream regions around the TSSs, as well as statistical analyses of these regions. The methodology thus developed was applied in order to analyse the genomes of *Mus musculus* and *Homo sapiens* for areas unlikely to initiate transcription. The analysis suggests that on average more than 40% of the human and mouse genome can be regarded as '*transcription initiation desert*' and thus as highly unlikely to initiate transcription.

The '*transcription initiation desert*' regions that were determined with this methodology were subsequently combined with other available evidence for the existence of transcription to produce '*transcriptional deserts*'. '*Transcriptional deserts*' are set apart from '*transcription initiation deserts*' in so far as the latter comprise regions of mammalian genomes that do not initiate transcription while the former consist of regions that are neither themselves transcribed nor are they likely to initiate transcription. '*Transcriptional deserts*' were examined for their compositional properties, repeat content, occurrences of single nucleotide polymorphisms (SNPs), evolutionary conservation and the presence or absence of binding sites for transcription factors. The results of these analyses suggest that while these regions are not transcriptionally active, they cannot be regarded as devoid of function. The data shows that they have distinct characteristics, harbour a high concentration of remote regulatory elements and are of importance to understanding gene function.

The method introduced in this work represents the first one capable of identifying large parts of mammalian genomes as '*transcription initiation deserts*'. This methodology can significantly localise the search for TSS locations and thus contribute to promoter and gene finding, to more successful experimental designs, as well as to gene annotation. It can also help in the assessment of 5' completeness of expressed sequences. The areas identified in this work as '*transcriptional deserts*' do play an important role when investigating gene function.

Declaration

I declare that "*Genome-wide identification and comprehensive analysis of transcriptional desert regions*" is my own work, that it has not been submitted for degree or examination at any other university, and that all the resources I have used or quoted, and all work which was the result of joint effort, have been indicated and acknowledged by complete references.

Ulf Schaefer August 2009

Signed

Publication arising from this thesis

Hofmann O, Caballero OL, Stevenson BJ, Chen YT, Cohen T, Chua R, Maher CA, Panji S, **Schaefer U**, Kruger A, Lehvaslaiho M, Carninci P, Hayashizaki Y, Jongeneel CV, Simpson AJG, Old LJ, Hide W. Genome-wide analysis of cancer/testis gene expression. ***PNAS***. 2008 Dec 23;105(51):20422-7. Epub 2008 Dec 16 [1]

Sagar S, Kaur M, Dawe A, Seshadri SV, Christoffels A, **Schaefer U**, Radovanovic A, Bajic VB. DDESC: Dragon Database for Exploration of Sodium Channels in Human, ***BMC Genomics*** 2008, 9:622 [2]

Kaur M, Radovanovic A, Essack M, **Schaefer U**, Maqungo M, Kibler T, Schmeier S, Christoffels A, Narasimhan K, Choolani M, Bajic VB. Database for exploration of functional context of genes implicated in ovarian cancer, ***Nucleic Acids Research***, 2009,Vol. 37, Database issue D820-D82 [3]

Essack M, Radovanovic A, **Schaefer U**, Schmeier S, Seshadri SV, Christoffels A, Kaur M, Bajic VB. DDEC: Dragon Database of Genes Implicated in Esophageal Cancer. ***BMC Cancer*** 2009, 9:219 [4]

Schmeier S, MacPherson CR, Essack M, Kaur M, **Schaefer U**, Suzuki S, Hayashizaki Y, Bajic VB. Deciphering the transcriptional circuitry of microRNA genes expressed during human monocytic differentiation. ***BMC Genomics*** [5]

Dawe AS, Radovanovic A, Kaur M, Sagar S, Seshadri SV, **Schaefer U**, Christoffels A and Bajic VB. DESTAF: A Database of Text-Mined Associations for Reproductive Toxins Potentially Affecting Human Fertility, ***Reproductive Toxicology***, **under review** [6]

Schaefer U, Kodzius R, Kai C, Kawai J, Carninci P, Hayashizaki Y, Bajic VB. High sensitivity TSS prediction and its application to TSS desert identification. **In preparation for submission to *Bioinformatics***

Schaefer U, Bajic VB, et al. A comprehensive analysis of transcriptional desert regions in mammalian genomes. **In preparation for submission to *BMC Genomics***

Schaefer U, Bajic VB, et al. PROMEX: A web based tool for evidence based promoter extraction. **Application note in preparation for submission to *BMC Bioinformatics***

Table of Contents

List of tables

List of figures

Abbreviations

CAGE	-	capped analysis of gene expression
cDNA	-	complementary DNA
CV	-	cross validation
DDM	-	Dragon TSS Desert Masker
EST	-	expressed sequence tag
GIS	-	gene identification signature
HS	-	*Homo sapiens*
knt	-	kilonucleotide
LINE	-	long interspersed nuclear element
LTR	-	long terminal repeat
MM	-	*Mus musculus*
nt	-	nucleotide
OSM	-	online supporting material
PPP	-	promoter prediction programme
RACE	-	rapid amplification of cDNA ends
RT	-	reverse transcriptase
SE	-	sensitivity
SINE	-	short interspersed nuclear element
SNP	-	single nucleotide polymorphism
SP	-	specificity
TD	-	transcriptional desert
TF	-	transcription factor
TFBS	-	transcription factor binding site
TIAR	-	transcription initiation active region
TID	-	transcription initiation desert
TSS	-	transcription start site
UTR	-	untranslated region

Preface

The research presented in this dissertation deals with three separable yet connected issues. These three issues are dealt with in the three main chapters of this dissertation and from them the overall structure of this piece of work is derived. Chapter one deals with the initiation of transcription. It establishes a reference dataset of transcription start sites and introduces a methodology that allows the demarcation of locations in mammalian genomes that are unlikely to initiate transcription. The research presented in chapter one is currently being prepared for submission for publication in *Bioinformatics*.

The subject of chapter two is the analysis of transcriptionally inactive regions in the genome of mammals. The methodology that was introduced in chapter one was combined with other existing data to locate regions in the genome that are devoid of any transcripts. These regions are subsequently examined under various aspects like sequence composition and the occurrence of mutations. This analysis gives indications of their genetic function and their role in molecular cell mechanisms. The research presented in chapter two is currently being prepared for submission for publication in *BMC Genomics*.

Chapter three explains the design and implementation of PROMEX, a tool for promoter extraction that was utilised for data procurement in chapter one and has contributed to several other projects that were conducted during my years of doctoral studies. Each chapter is presented as an independent, conclusive and self-sufficient piece of research. The dependencies and interrelations of the chapters are briefly discussed in the final 'Summary' section at the end of this document.

Chapter 1 – The identification of transcription initiation desert regions

INTRODUCTION

Although the full sequence of the human genome as well as other mammalian genomes has now been available for several years, the annotation of these genomes is far from being complete. Especially the variability of the transcriptome and the existence of numerous sometimes hugely different transcripts for a single gene have posed great challenges to the scientific community in deciphering the exact function of all parts of mammalian genomes. A substantial portion of these difficulties arise from the fact that a large number of genes possess many possible transcription start sites. These can be located far upstream or downstream from the 5' end of the gene body.

The sequencing of full-length cDNA libraries, the generation of millions of ESTs, and later tag approaches (CAGE, GIS, etc) [7-11] have provided the scientific community with information on transcripts and the location of their transcription start sites. This data illustrates that transcription in mammalian genomes can initiate at various and unusual positions (e.g. coding exons, 3'UTR) [12,13] and thus contribute to the complexity of mammalian transcriptomes. However, mammalian transcription does not initiate randomly. It is observed that large segments of the genome are not prone to the initiation of transcription. All collections of transcriptional data show that transcription initiation activity in mammalian genomes is concentrated in specific regions. At the same time, the data allows us to conclude that there are vast stretches of DNA where transcription initiation is not observed to occur. A detailed analysis of the TSS neighbourhood [14] shows that there are a lot of regularities in the regions immediately surrounding the TSSs, making these regions more suitable environments for transcription initiation events. The same behaviour can also be observed in the prevalence of genes and their locations on the mammalian genome. It is

known that many regions in these genomes are considered to be gene deserts [15,16]. They constitute regions that contain very few genes. At the same time, many other segments of the mammalian genome are rich with genes, such as human chromosome 22 [17,18]. Traditionally these gene dense and gene desert regions have been interpreted in the convenient terms of GC-richness of isochores on the mammalian genomes [19].

Sometimes transcriptional activity in the form of CAGE tags and/or transfrags [20] is observed within the genomic desert regions. In order to experimentally confirm the existence of such novel and unexpected transcripts, RACE [21] is the method of choice. If it is possible to obtain a full length transcript from the observed tag, the existence of a transcript at this location can be regarded as confirmed. However, the RACE primer design is very difficult if no information on the transcribed regions is available. On the other hand, if a TSS location is known and well-supported by the existence of several CAGE tags or a combination of CAGE tag(s) and other expressed sequences, then by designing primers close to this known TSS, more than half of novel putative ESTs can be experimentally confirmed through RACE. This brings into consideration the preparation of experimental designs for confirming putative transcripts. It would be of great value to computationally filter candidate transcripts before performing RACE, to avoid wasted experiments.

Another issue is that it is difficult to ascertain that ESTs are 5' complete. This problem is one of the burning issues in determining the accurate TSS locations, which impacts follow-up studies on transcriptional regulation. It is thus of practical importance to be able to determine, computationally and in advance, which regions in genomic DNA are likely to be good environments for transcription initiation and which ones are not. The availability of such knowledge would lead to more precise experimental designs and to the elimination of a vast number of false positive transcription candidates. For example, if the 5'end of transcript falls into a

region not likely to initiate transcription, this would signal its potential 5' incompleteness.

The efforts of the scientific community have provided a vast amount of transcript data that has allowed the very precise determination of a large number of TSSs in mouse and human genomes [10,11,22,23]. Based on the analysis of the properties of the upstream and downstream regions immediately surrounding the TSS [14], it was observed that TSS locations in both mouse and human follow certain rules that confine these TSSs to particular genomic regions. This idea was utilised and extended with the aim to develop the Dragon TSS Desert Masker (DDM). This tool can demarcate in a strand specific manner, locations in mammalian (mouse and human) genomes that are highly unlikely to contain sites of transcription initiation. The collection of all these locations is referred to as transcription initiation desert or TID. The Dragon TSS Desert Masker (DDM) is a tool that is capable of annotating a significant portion of the TID. The non-annotated part is likely to contain the vast majority of transcription initiation sites. This non-annnotated part is being referred to as transcription initiation active region or TIAR. DDM is able to perform the distinction between TID and TIAR with high accuracy.

Using DDM, it is possible to mask a part of TID regions in mammalian DNA. The non-masked regions indicate TIAR that is likely to harbour the vast majority of genuine TSSs. The TIAR might possibly support a more precise RACE primer design and can help in estimation of completeness of the 5'-ends of ESTs. Consequently, they can help in annotation of promoter regions in mammalian genomes and moreover, such information can complement promoter and gene finding and help focus on those regions that are of particular interest.

RESULTS

TID is defined as the set of all strand-specific genomic locations that are highly unlikely to initiate transcription. The remaining locations are called TIAR and contain the vast majority of genuine TSSs as well as a remaining portion of locations unable to initiate transcription that were not possible to localise using the current TSS prediction algorithm introduced here. Currently and unfortunately, it is impossible to determine whether the majority of all genuine TSSs for any mammalian genome are known. Therefore, as regrettable as it might be, only estimates of TID can be made. This is because of the aforementioned non-determinability of whether or not all genuine TSS locations are known for any mammalian genome. For the sake of simplicity these estimates of TID are referred to as TID in the further text. A number of key promoter features were determined. These allow the separation of mammalian genomic sequences into active (TIAR) and desert (TID) domains relative to transcription initiation. Due to the fact that TID is only estimated, based on a computational algorithm that is not perfect, TIAR are considered to be a set of genomic locations that contain the vast majority of the known TSS locations, while TID are those genomic locations that contain only a minimal fraction of known TSSs or contain no TSSs at all. Thus the density of known TSS locations in TIAR is expected to be considerably higher than in TID. In order to be able to make the distinction between TID and TIAR regions, one needs to be in possession of a tool that is able to distinguish between genomic positions that are likely to initiate transcription and those that are unlikely, in such a way that no or only very few false-negative statements about TSS locations are made. This is equivalent to a TSS recognition system that operates at or very near 100% sensitivity. Subject to the condition that the TSS set has sufficient coverage, at this level of sensitivity it can be expected that the areas labelled as unlikely to initiate transcription are indeed almost completely devoid of TSSs, because all or nearly all non-TSS statements made by the predictor will be true. In this context, TID are understood to be those locations that

were determined to contain no, or almost no, known TSSs. On the contrary, TIAR is composed of those locations that contain all, or almost all, known TSSs, but it also contains all false positive TSS predictions. It is for this reason that it is not claimed that all locations within TIAR are potential TSSs.

Algorithm

The efforts of the Fantom3 consortium [12] constitute one for the most comprehensive collections of transcription data available. Based on Fantom3 CAGE data and at least one other piece of evidence for the existence of a transcript, two highly accurate sets of TSS locations were compiled. The exact methodology that was employed for the compilation of these data sets is described below in the 'Methods' section of this chapter. For each true TSS location, chromosome, strand and genomic position are recorded. The TSS data set for *Mus musculus* consists of 98,682 accurately determined TSSs, while the TSS data set for *Homo sapiens* contains 113,814 accurately determined transcription start sites. These data sets are used as reference sets of positive samples for genuine true TSSs.

Using these mouse and human TSS data sets, the compositional properties of single-stranded DNA segments covering [-100, +100] nt regions relative to the TSS were analysed. Based on these properties, a system that utilises a variety of different filtering methods in order to filter out those DNA segments that are unlikely to represent genuine TSS positions was designed. When presented with a previously unseen DNA segment, the system is able to accurately determine whether this segment is likely to harbour a genuine true TSS position. Each filtering method employed in the algorithm filters out different fractions of the data by concentrating on different characteristics of the compositional properties of the DNA segments provided. These filtering methods were combined in a multi-staged daisy-

chain algorithm that consists of four different classification phases. The exact method of operation for the complete algorithm is described below in the 'Methods' section of this chapter. On overview of the layout of the algorithm is given in Figure 1.

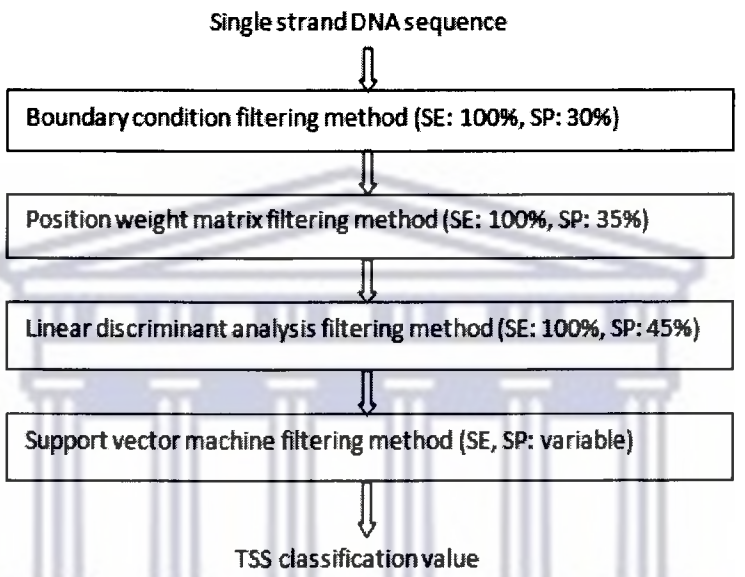

Figure 1: Layout of daisy-chain algorithm, performance estimates after each step in parenthesis

The algorithm analyses nucleotide sequences of length 200 nt. For each segment of DNA that the algorithm examines, it returns a classification value. This output classification value reflects the algorithm's prediction of whether the segment contains a TSS at its centre or not. A threshold is applied to this value in order to determine whether the examined sequence contains a TSS or not. The centre (nucleotide at position +1) of the examined [-100, +100] sequence is masked as a location unlikely to initiate transcription if the classification value is below this threshold. If the classification value is above this threshold, then the centre of the examined sequence is marked as a potential transcription start site.

While this algorithm can be regarded superficially as an ordinary promoter predictor, one has to keep in mind that the accurate prediction of TSS (while desirable) is not the intent or purpose of the presented algorithm. DDM is specifically designed to detect TID, which means that the prediction of TSSs is only implicitly the subject. The DDM algorithm is specifically tuned to operate in such a way that is produces no or very few false-negatives which is a necessary requirement for producing accurate estimates of TID. For this reason DDM is not suitable to be used as a tool for the accurate prediction of TSSs. This point is elucidated further later in the text when DDM is compared to existing promoter predictors in the context of predicting TID.

Performance

The algorithm was applied to data sets from *Mus musculus* and *Homo sapiens*. All genuine TSSs from the reference data sets were used as positive samples. An equal amount of random DNA was extracted from the genome of the respective species. These random DNA sequences served as negative samples. The algorithm was applied to all data and the resulting classification values were collected. After that, a range of threshold values was applied to the classification values and sensitivity and specificity values were determined in order to assess the algorithm's performance. The performance is reported for two separate cases (see Figure 2). Firstly, a 4-fold cross-validation, for which a quarter of the sequences was selected randomly for testing, while three quarters were used for training. The average sensitivity and specificity values for the four test sets are reported in Figure 2 as case 'test' for mouse and human sequences. Secondly, the entire available data sets were used. For this second case no separation between test and training sets is undertaken. Sensitivity and specificity is reported for the entire available data and shown in Figure 2 as case 'training' for mouse and human sequences. The resulting models of the case 'training' are also

used in the algorithm's web implementation which is available at http://apps.sanbi.ac.za/DDM/.

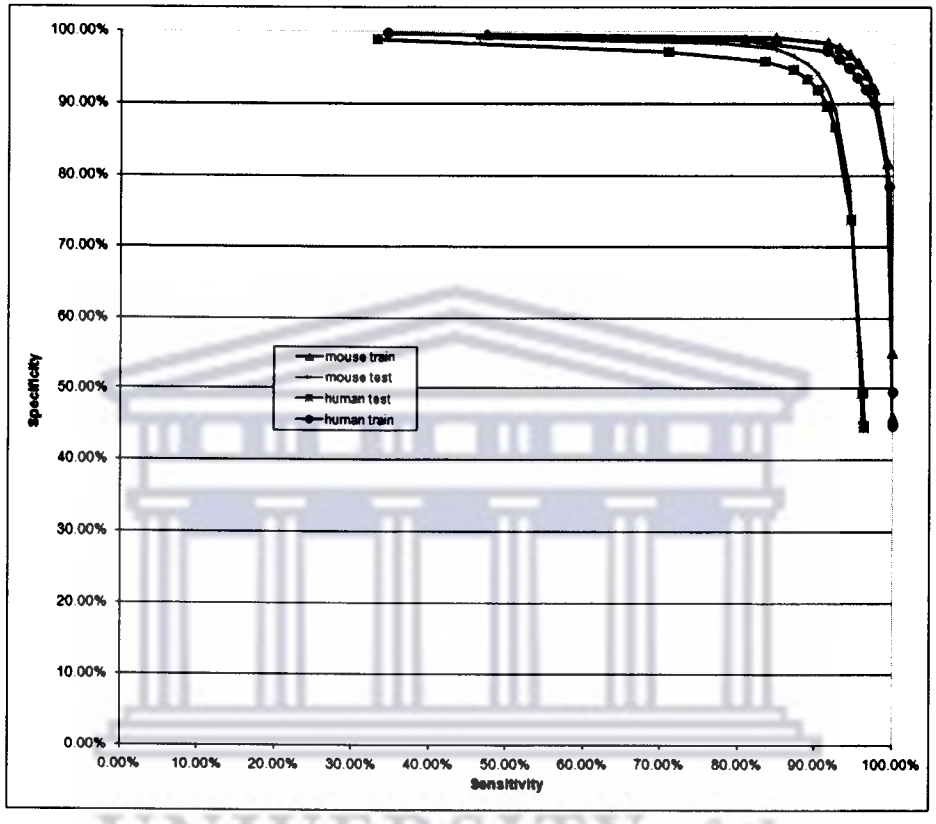

Figure 2: Sensitivity vs. specificity trade-off curve for human and mouse average CV performance, and performance on the whole data sets.

One observes that, with the models that were derived from all available data, which were also used in the implementation of the web server version of DDM, at the level when no known TSSs are lost (which represents a sensitivity of 100% in the system) about 45% of the mouse and human random non-TSS DNA sequences are recognised correctly as those that should not initiate transcription. The fact that DDM delivers this performance at 100% sensitivity is essential to the algorithm's ability to recognise transcription initiation deserts. This is because the entirety of the areas that are annotated by the algorithm as not likely to initiate

transcription can only be regarded as TID if the remaining areas contain all or nearly all genuine transcription start sites. This is only true when the algorithm works at or very close to 100% sensitivity. At lower rates of sensitivity the areas marked as not containing TSS will still contain a portion of the true TSS that is too high to denote these areas as transcription initiation desert regions. As the next section will show, DDM is the only tool that is specifically designed with the intention of detecting TIDs and therefore the only tool that can be used for such purpose.

At a lower sensitivity setting of 99.22% (99.53%), DDM is able to mask a remarkable portion of 81.75% (78.45%) of the mouse (human) random non-TSS DNA sequences. DDM sensitivity of 99.22% means that the system is not able to recognise 0.78% of the real TSSs from the reference data set. With a balanced sensitivity/specificity setting DDM was able to retain 95.44% (95.33%) of true TSSs while at the same time masking 95.63% (93.58%) of the mouse (human) sequences as unlikely to initiate transcription. When 99.16% (98.83%) of the random non-TSS mouse (human) sequences are masked as unable to initiate transcription, a significant portion of 84.76% (80.74%) of the true TSS is recognised as positions likely to harbour a TSS. The performance of DDM at various thresholds for human and mouse sequences is shown in Table 1. The table shows results obtained with models derived from all data and results obtained through a 4-fold CV for both species.

Mouse whole set			Mouse CV			Human whole set			Human CV		
threshold	Sensitivity	Specificity	threshold	Sensitivity	Specificity	threshold	Sensitivity	Specificity	threshold	Sensitivity	Specificity
-2.50	100.00%	45.45%	-2.50	96.10%	45.01%	-2.50	100.00%	44.77%	-2.50	96.36%	44.51%
-2.00	99.99%	45.95%	-2.00	96.07%	45.55%	-2.00	100.00%	44.97%	-2.00	96.36%	44.74%
-1.50	99.91%	55.03%	-1.50	95.72%	54.82%	-1.50	99.98%	49.47%	-1.50	96.20%	49.34%
-1.00	99.22%	81.75%	-1.00	94.30%	78.40%	-1.00	99.53%	78.45%	-1.00	94.62%	73.73%
-0.50	97.47%	92.03%	-0.50	92.49%	89.16%	-0.50	97.59%	89.92%	-0.50	92.46%	86.71%
-0.25	96.50%	94.05%	-0.25	91.41%	92.10%	-0.25	96.42%	91.96%	-0.25	91.47%	89.69%
0.00	95.44%	95.63%	0.00	90.21%	94.11%	0.00	95.33%	93.58%	0.00	90.29%	91.91%
0.25	94.33%	96.86%	0.25	88.80%	95.52%	0.25	94.23%	94.94%	0.25	88.91%	93.41%
0.50	92.98%	97.81%	0.50	86.98%	96.68%	0.50	92.98%	96.17%	0.50	87.12%	94.74%
0.75	91.50%	98.51%	0.75	84.52%	97.66%	0.75	91.46%	97.29%	0.75	83.42%	95.91%
1.00	84.76%	99.16%	1.00	77.36%	98.40%	1.00	80.74%	98.83%	1.00	70.95%	97.22%
1.25	47.32%	99.65%	1.25	46.08%	99.09%	1.25	34.57%	99.76%	1.25	33.19%	98.91%

Table 1: Sensitivity and specificity values for mouse and human test and training cases

For the identification of TID, the performance of the algorithm at very high sensitivity levels (100%) is most relevant. This is because only when the false-negative rate is equal to or very near 0% one can meaningfully speak of the areas that were recognised as unlikely to initiate transcription as being 'transcription initiation deserts'.

Comparison with existing promoter prediction programmes

If the ideal TSS predictor exists (sensitivity = 100%, specificity = 100%), determination of TID will be trivial. However, such TSS predictors do not exist as yet, although many good tools for TSS prediction are available. Albeit, none of these tools has a performance that can solve the problem that

was investigated here (TID estimation). This will be shown later by a comparison analysis.

The novel question that was investigated here, and which has never been attempted before, is the estimation of TID regions. The knowledge of these regions makes as much sense as much as the knowledge of TSSs. If it is agreed that it is useful to have knowledge of regions that can initiate transcription, then it is equally useful to have knowledge about regions that cannot initiate transcription, if for no other reason than for localising the search for promoters/TSSs. One can argue that the task of predicting TIDs cannot be understood to be equivalent to the task of accurately predicting TSSs. To support this notion, consider a TSS predictor that predicts TSSs at 80% sensitivity. This means that 20% of TSSs still remain in TID. This is why it is claimed here that the estimation of TID is not simply the negation of TSS prediction. The best one can currently have is a TSS predictor that is capable of predicting TSS with close to 100% sensitivity, since only then does it makes sense to talk about TID (and estimates of TID) at locations where predictions are not made. In fact it can be argued that TSS predictions and TID predictions are two tasks directly opposite with regard to their aim. While the prediction of TSS concentrates on predicting TSSs with the highest possible accuracy, the aim of predicting transcription initiation deserts endeavours to achieve the opposite, namely to identify regions, as accurately as possible, that do not initiate transcription. While this seems to be the same thing at first glance, the distinct difference in aim has wide-ranging implications on a system's design and its eventual use. Therefore estimating TID is not simply the negation of predicting TSSs. Truly a TID predictor was developed implicitly as a TSS predictor, or as a promoter prediction system. However, in order to have as accurate a prediction as possible of TID locations, the TSS predictor has to operate at a sensitivity level of or very near 100% in order to avoid false-negative TSS predictions. Such false negative prediction would have the effect that the estimated TID still contains a certain amount of genuine TSSs. In such cases one cannot rightly label the locations as belonging to TID.

Many existing promoter prediction programmes (PPPs) have been developed with the general aim to predict TSSs with certain levels of precision and positional accuracy [24]. These tools are designed to forecast the existence of transcription start sites or more generally promoter regions. However, none of these programmes was designed to identify TIDs. The goal that the designers of these programmes had in mind was a rather different one [24,25]. These existing programmes intend to reach the best possible trade-off between sensitivity and false-positive rate. DDM, on the contrary, is designed in such a way that it allows the recognition of all or very nearly all true TSS locations (i.e. ~100% sensitivity). Moreover, in many cases the positional accuracy of the predictions of TSSs by existing PPPs is poor, which in itself makes the identification of TIDs complicated. DDM, on the contrary, annotates genomic locations as likely or unlikely to initiate transcription with the highest possible positional accuracy of no mismatch between the prediction and the real TSS. Consequently, DDM appears to be the only tool available that is specifically designed to demarcate regions unlikely to initiate transcription.

Having asserted the above claims, it can be argued that a comparison of DDM with existing PPPs is dispensable. Such a comparison would compare systems that have distinctly different basic design goals and thus this comparison is questionable. However one could also argue that the same goal that DDM is designed for can, in principle, be achieved with the existing promoter predictors. In principle, any promoter predictor that provides a very high sensitivity level (close to 100%) of predicting TSS locations at a one nucleotide resolution can serve the purpose of estimating TID. To test if this is possible with the currently available promoter predictors, a comparison between the abilities of several such predictors and DDM to accurately predict TID was made. It should be highlighted that the aim of this comparison is not to evaluate how well promoter predictors perform in predicting TSSs (though this aspect is implicitly involved), but rather how capable they are in accurately estimating TID. This comparison

analysis will show that DDM is superior for this task and achieves accuracy that is much better than other systems can achieve.

In order to make a comparison of how well PPPs and DDM perform in identification of TIDs, programmes from [24,25] were evaluated. For this purpose the datasets $HTSS_{compare}$ and $RNDM_{compare}$ were created. For details about the creation and content of these datasets please see the 'Methods' section of this chapter. To make this comparison as fair as possible, a test set was created that contains 1000 randomly selected TSSs ($HTSS_{compare}$) from the original human TSS set, and 1000 randomly selected human DNA sequences, $RNDM_{compare}$. DDM was then retrained with the remaining human TSS sequences and the random DNA sequences that did not contain $RNDM_{compare}$ (see the 'Methods' section of this chapter). Consequently, the test set data is completely independent of the training set for DDM for this comparison.

The datasets $HTSS_{compare}$ and $RNDM_{compare}$ were analysed with Promoter2.0 [26], NNPP2.2 [27], First Exon Finder [28], Eponine [29] and Fprom [30]. N-SCAN [31] and McPromoter [32] do unfortunately only allow very limited online submission and thus were not tested with $HTSS_{compare}$ and $RNDM_{compare}$. Instead only the performance as it is given by the authors of the respective studies could be reported. CpGProD [33], Dragon Promoter Finder [34,35] and Dragon Gene Start Finder [36,37] have specific design constraints that make them unsuitable for this comparison. The constraints of these three PPPs are further elucidated in the 'Discussion' section of this chapter.

The URLs of the PPPs used in this comparison can be found in Table 2.

Promoter2.0	http://www.cbs.dtu.dk/services/Promoter/
NNPP2.2	http://www.fruitfly.org/seq_tools/promoter.html
First Exon Finder	http://rulai.cshl.org/tools/FirstEF/
Eponine	http://servlet.sanger.ac.uk:8080/eponine/
Fprom	http://www.softberry.ru/berry.phtml?topic=fprom&group=programs&subgroup=promoter
N-SCAN	http://mblab.wustl.edu/nscan/submit/
McPromoter	http://tools.genome.duke.edu/generegulation/McPromoter/McPromoter.html

Table 2: URLs of promoter prediction tools used in comparison

Two tests were conducted wherever possible. For the first test (test A), a mismatch of ±100 nucleotides was allowed for a prediction of a TSS to be counted as correct. For the second test (test B), only those predictions were counted as correct that predict the known true TSS with no mismatch. A negative prediction was regarded as correct if there was no prediction within 100 nucleotides of position 801 for each sequence in $RNDM_{compare}$ (for test A) or if there was no prediction at 801 exactly (for test B). The results of these experiments are described below.

The tests were conducted using two data sets, a set of 1000 sequences from human covering [-800, +800] relative to a known true TSS ($HTSS_{compare}$) and a set of randomly chosen human sequences of length 1600 nt ($RNDM_{compare}$). Sensitivity was determined on $HTSS_{compare}$ as the portion of sequences in $HTSS_{compare}$ that were correctly recognised as TSSs by the respective tool, either with (test A) or without positional mismatch (test B). Accordingly, specificity was determined on $RNDM_{compare}$ as the portion of those sequences in $RNDM_{compare}$ that were correctly recognised as not being

a TSS by the respective tool, either with (test A) or without positional mismatch (test B). It is in this way, that DDM achieves a sensitivity of 99.8% and a specificity of 40.1% at threshold -2.0 with no mismatch (test B) and 100.0% and 11.1% for sensitivity and specificity respectively at threshold -1.0 (test A).

Consider for example a PPP that achieves a performance of 85% sensitivity and 80% specificity in identifying TSSs. While this is a respectable performance for TSS prediction, this tool can still not be used for the identification of TID. This is because a sensitivity of 85% means that 15% of true TSSs are not recognised as TSSs by this tool. The consequence of this is that areas labelled as devoid of TSS by this tool would in fact still contain 15% of true TSSs. This would disqualify these areas as 'transcription initiation deserts'.

Promoter2.0

Promoter 2.0 does not allow the setting of any threshold. For test A, a sensitivity of 22.5% and a specificity of 86.6% were achieved on $HTSS_{compare}$ and $RNDM_{compare}$ respectively. Promoter2.0 does not provide predictions with no positional mismatch, so test B was omitted for this tool.

NNPP2.2

For test A the threshold for which NNPP2.2 achieves 100% sensitivity on $HTSS_{compare}$ was determined to be t=0.12. For this value of t, a specificity of 4% on $RNDM_{compare}$ in test A is observed. For test B, a sensitivity of 21% and a specificity of 95% for the same value (t=0.12) is seen.

First Exon Finder

The lowest available threshold t=0.2 for all probabilities was used. This selection of the threshold guarantees the highest possible sensitivity that this tool can achieve. For test A, First Exon Finder achieves a sensitivity of 40.7% and a specificity of 98.6% on $HTSS_{compare}$ and $RNDM_{compare}$

respectively. First Exon Finder does not provide predictions with one nucleotide accuracy, so test B had to be omitted for this tool.

Eponine

The lowest available threshold t=0.9 was used. This threshold selection delivers the highest sensitivity possible for Eponine. For test A, Eponine achieved a sensitivity of 34.3% and a specificity of 91.4% on $HTSS_{compare}$ and $RNDM_{compare}$ respectively. Eponine does not provide predictions with one nucleotide accuracy, so test B had to be skipped for this tool.

Fprom

The thresholds for which the authors report a sensitivity of 100.0% for non-TATA-box promoters and TATA-box promoters respectively (-9.496 and -6.766) were used. For test A, Fprom achieves a sensitivity of 59.3% and a specificity of 99.4% on $HTSS_{compare}$ and $RNDM_{compare}$ respectively. For test B, a sensitivity of 2.4% and a specificity of 100.0% was observed.

N-SCAN

N-SCAN does not allow the submission of multiple sequences simultaneously, so meaningful tests could not be conducted. The authors of this tool do however report the performance of this tool to be 21% vs. 29% for sensitivity and specificity respectively when predicting transcripts and 84% vs. 63% when predicting exons.

McPromoter

McPromoter does not allow the submission of multiple sequences simultaneously, so meaningful tests could not be conducted. The author of this tool does however report the tool to have a sensitivity of 65% at the highest available sensitivity level.

The performance of all PPPs that were examined is summarised in Table 3. The performance of DDM with no mismatch allowed is shown in Table 4 for comparison. For Table 4 threshold values were adjusted to match each

sensitivity and specificity value reported in Table 3 and the corresponding performance of DDM at this threshold is reported.

Tool	SE	SP	threshold	SE	SP
Promoter2.0	22.5%	86.6%	n/a	n/a	n/a
NNPP2.2	100.0%	4.0%	0.12	21.0%	95.0%
First Exon Finder	40.7%	98.6%	0.2	n/a	n/a
Eponine	34.3%	91.40%	0.9	n/a	n/a
Fprom	59.3%	99.4%	0.0	2.4%	100.0%

Table 3: Performances of promoter prediction tools

2.40%	100.00%
4.0%	99.90%
21.00%	99.50%
22.50%	99.20%
34.30%	99.10%
38.40%	98.60%
40.70%	98.50%
53.3%	98.2%
85.50%	95.00%
91.20%	91.40%
92.30%	86.60%
95.20%	58.70%
99.80%	4.00%
100.00%	4.00%

Table 4: Performance of DDM on $HTSS_{compare}$ and $RNDM_{compare}$ with no mismatch (test B)

Based on the results obtained through this comparison, it can be concluded that none of the PPPs described achieves a performance that is good enough to identify, with a high accuracy, locations that are not likely to initiate transcription. Such locations must be guaranteed to be largely devoid of TSS. DDM is the only tool presently available that manages to detect such locations. However, it must however be remarked that the performance of the programmes tested might improve if they were trained with the highly accurate datasets that were used for DDM or if they could be tuned specifically for sensitivity of 100% or close to it. Since in many cases the promoter data sets that have been used in the configuration of the tested PPPs have been limited by today's standards for knowledge about

transcription initiation, the performance of the PPPs examined must be seen in the light of their age and the availability of data at the time of their creation.

Application of DDM to identify TID

In order to make any statement about how much of the mammalian genome is able to support the initiation of transcription, the DDM algorithm needs to be applied to the entirety of all chromosomal sequences. Each position within the chromosomal sequences needs to be examined. To achieve this, a sliding window of length 200 nucleotides is analysed by DDM. The algorithm determines the propensity of the nucleotide at position +1 to be the location of transcription initiation. After that, the window is moved by one nucleotide and the analysis is repeated for the next nucleotide at position +1.

The human chromosomes 21, 22 and 4 were selected as showcases for the analysis of the whole genome. These chromosomes reflect an average, high, and low GC-content with regard to the whole human genome. Since the gene-richness and the amount of transcriptional activity on a certain genomic region is often explained in terms of the GC-content of this region, it is interesting to see how DDM behaves in GC and AT rich environments respectively. As a matter of fact, the amount of known genes on chromosomes 21, 22, and 4 is about average, and relatively high, and low in comparison with the whole human genome as well, when normalised for the size of the chromosomes. The DDM algorithm was applied to the forward and reverse strand of the chromosomes in question. The results are shown in Table 5. They reflect the average between positive and negative strand, keeping in mind that the differences between the two strands are minimal to start with.

Threshold	TID	TIAR	TID	TIAR	TID	TIAR
0.0	91.53%	8.47%	78.18%	21.82%	95.87%	4.13%
-2.5	41.1%	58.9%	27.2%	72.8%	46.84%	53.16%

Table 5: TID and TIAR of three showcase human chromosomes

For the analysis of these chromosomes two different threshold settings (0.0 and -2.5) were used. The latter threshold allows a performance of 100% sensitivity on the complete data sets. Sensitivity in this case refers to the ability of DDM to correctly identify a TSS location. For all chromosomes examined, a certain level of masking was observed, which refers to the proportion of the chromosome that is deemed very unlikely to harbour TSSs. If a sequence is annotated with DDM a threshold of -2.5 should be used to ensure that all or the vast majority of potential TSSs are recognised correctly.

One notices that this level of masking is in correlation with the GC-richness of the chromosomes, as well as with the number of known genes [22] on these chromosomes. This means that the higher the GC-content of (or gene density on) a chromosome, the lower the observed level of masking, and vice versa. In so far the observations made regarding the level of masking through DDM comply with the expectations that arise from information about the GC-content of the examined chromosomes. However, in spite of the observed correlation between level of masking and GC-richness, the TIDs are not confined only to the GC-poor regions and can be found also within the GC-rich areas. Chromosome 21 can be regarded as a showcase example, because it has an approximately average GC-content in comparison with the entire human genome. At a sensitivity of 96.36% on cross-validation, 41.1% of the chromosome is masked as TID. At a sensitivity of 90.29% on CV, 91.53% of human chromosome 21 is masked

as TID. If the locations of the genuine true TSS that are present in the data set for human chromosome 21 are compared with the regions that were marked as likely to initiate transcription, it can be shown that the remaining 8.47% of chromosome 21 contain ~92% of all TSSs on this chromosome. This makes the density of TSS in the TIAR 132 fold higher than that in the TIDs. This justifies the classification of the TID and TIAR domains as active and desert regions relative to transcription initiation that was introduced earlier in this chapter.

For human chromosome 22, which has a relatively high GC-content in comparison with the whole human genome, 21.82% of the chromosomal sequences are marked as TIAR. This portion of the chromosome contains 93.8% of all genuine true TSS for this chromosome. This makes the density of TSS in TIAR 54-fold higher than in TIDs. For human chromosome 4, which has a relatively low GC-content in comparison with the whole human genome, 4.13% of the chromosomal sequences are obtained as TIAR. This portion of the chromosome contains 83.6% of all genuine true TSS for this chromosome. This makes the density of TSS in TIAR 118-fold higher than in TIDs.

Repeats and transcription initiation deserts

Repeat sequences are an abundant genomic element in vertebrates. In mammals they often make up more than 40% of the entire genome [38]. Repeats can be roughly grouped into two categories, tandem repeats and interspersed repeats. Tandem repeats are normally areas of low complexity DNA where a certain motif is repeated a certain number of times. The motifs are usually not longer than 60 nucleotides and often significantly shorter. Interspersed repeats are longer sequences that fall within the larger group of mobile genetic elements. These sequences possess the ability to move from one location in the genome to another by multiplying themselves. There are two basic mechanisms through which this is

achieved: one directly employs an enzyme called 'transposase', which directly transfers the genomic element, while the other one involves the transcription of the interspersed repeat to an RNA intermediate and the subsequent reverse transcription of the intermediate. There is evidence that suggests that repeat elements play a role in evolution by helping to form new genes. They might also play a role in genetic disorders [39-41].

Repeat elements were previously regarded to be of little significance for the characteristics and behaviour of the cell. The opinion that these sequences constituted 'genomic background noise', that had no function, was popular. In fact, repeat sequences are regularly excluded from analyses that deal with gene function, gene regulation and other issues revolving around the functional annotation of eukaryotic genomes (e.g. [42]). With this in mind, it is interesting to examine in how far repeat sequences are able to initiate transcription. The general expectation would be that extremely little to no transcription is initiated from within repeat regions.

In order to facilitate this analysis, RepeatMasker [http://www.repeatmasker.org] [43] was applied with all default settings for *Homo sapiens* to a set of sequences spanning [-100,+100] around all known genuine human TSSs. Of the total length of these sequences, 4.95% are masked as repeats, predominantly as simple repeats and areas of low complexity (Table 6). This means that ~5% of all nucleotides that are less than 100 nt from a known TSS are classified as belonging to a repeat sequence. Of all known genuine TSSs, 18.48% possess a repeat within less than 100 nucleotides upstream or downstream of the position of the TSS. Therefore 81.52% of TSSs do not have a repeat within 100 nucleotides around them. Of the genuine true TSSs themselves, 3.4% were masked by RepeatMasker. It can therefore be concluded that it is incorrect to regard repeats a priori as incapable of initiating transcription, since the analysis shows that 1 in 30 TSS lies within a repeat.

```
==================================================
file name:     human_tss_113814.fa
sequences:     113814
total length: 22762800 bp  (22762800 bp excl N/X-runs)
GC level:      66.37 %
bases masked: 1127476 bp ( 4.95 %)
==================================================
                number of      length   percentage
                elements     occupied  of sequence
--------------------------------------------------
SINEs:               160        12831 bp    0.06 %
      ALUs            89         6265 bp    0.03 %
      MIRs            69         6280 bp    0.03 %

LINEs:               136        16263 bp    0.07 %
      LINE1           49         6179 bp    0.03 %
      LINE2           73         8722 bp    0.04 %
      L3/CR1          13         1298 bp    0.01 %

LTR elements:        198        27489 bp    0.12 %
      MaLRs           46         6234 bp    0.03 %
      ERVL            70         9982 bp    0.04 %
      ERV_classI      69         9565 bp    0.04 %
      ERV_classII      8         1304 bp    0.01 %

DNA elements:         37         3560 bp    0.02 %
      MER1_type       16         1625 bp    0.01 %
      MER2_type       12         1011 bp    0.00 %

Unclassified:          0            0 bp    0.00 %

Total interspersed repeats:     60143 bp    0.26 %


Small RNA:             4          207 bp    0.00 %

Satellites:           20         2220 bp    0.01 %
Simple repeats:     8990       444194 bp    1.95 %
Low complexity:    14315       620766 bp    2.73 %
==================================================
```

Table 6: Repeat analysis for human TSS sequences

The regions masked on human chromosome 21 by RepeatMasker (46.47% masking) and the regions masked by DDM (91.53% masking) were compared. Not all repeats are masked by DDM, which agrees with the fact that a fraction of TSSs was observed to be located within repeat sequences. In fact, only about half of the sequences marked as TID on human chromosome 21 are repeat regions. The other half seems to consist of sequences that are not repeats, but nevertheless not able to initiate

transcription. To elucidate this further DDM (with balanced sensitivity and specificity levels) and RepeatMasker were applied sequentially to human chromosome 21. The area that was left unmasked by either tool corresponds to TIAR that does not contain any repeat sequences. This area covered 7.43% of the chromosome, compared to 8.47% that was left as TIAR after applying DDM alone. It turns out that the area demarcated as TIAR when DDM and RepeatMasker are both applied to human chromosome 21 (i.e. TIAR without repeats) contains 89.1% of true TSS locations. This corresponds to a TSS density that is 102-fold higher in the unmasked area than in the TID region. This compares unfavourably to a density ratio of 132 when only DDM is used.

The observations made from these experiments suggest that the combination of RepeatMasker and DDM is, at balanced sensitivity and specificity levels, not beneficial to the overall performance in masking TIDs. As was shown above, 3.4% of TSSs are masked by RepeatMasker. When using DDM alone, the threshold allows for a more favourable reduction in sensitivity. It appears that the incorporation of repeat information into a system designed to detect genomic regions unlikely to initiate transcription, is not the optimal choice.

Example: DDM masking explains failed amplification by 5'-RACE

As was mentioned in the introduction to this chapter the DDM algorithm can be useful in eliminating false positive evidence for transcription in tag approaches to transcriptional analysis. To elaborate on this point an example was chosen where CAGE tags were further examined for the actual existence of a transcript by 5'-RACE experiments and the areas around those CAGE tags were examined with DDM.

Firstly, the case of two CAGE tags between alternative TSSs in the gene Oprm1 in mouse (opioid receptor, mu 1; coordinates: chr10, negative strand 3,308,332..3,557,942; EntrezGene ID: 18390) was considered. The area around and immediately upstream of the 5' end of this gene is shown in Figure 3. DDM demarcates two major TIAR in this genomic region. The TIAR in this area consist of a large number of consecutive nucleotides that are characterised by DDM as likely to initiate transcription. The larger of the two major TIAR blocks is about 3000 nt in size and contains the 5' end of the gene. This TIAR block can be understood to be the main promoter region of the gene Oprm1. A smaller TIAR block is found about 60,000 nucleotides upstream of the gene and suggests the existence of alternative TSSs. DDM also marked numerous other positions as potential TSSs. Due to the resolution of the figure, these are not shown in Figure 3. A TSS at position 3,557,930 is supported by one CAGE tag (Fantom3 representative tag ID 122BA39P0901, undefined tissue library) and this TSS is not masked as a position unlikely to initiate transcription by DDM. This TSS was confirmed by 5'-RACE experiments in 4 out of 6 tissue samples supporting this prediction. Details about the tissues used can be found in the online supporting materials for the Fantom3 publication [12]. The primer identifier for this TSS is T10F0065AF50. Thus the claim made by DDM is in agreement with the result of the RACE experiment.

Contrary to this, a false positive TSS at position 3,580,940 indicated by one CAGE tag (Fantom3 representative tag ID 119BA53D1906, macrophage tissue library) could not be confirmed by 5'-RACE in any of the 6 tissues used. 5'-RACE experiments with two different primers (T10F006553E1 and T10F006553F9) were conducted, but neither of them succeeded in producing a viable transcript. This false-positive TSS is masked by DDM suggesting it is not likely to promote transcription. As before, the claim made by DDM regarding the location's ability to promote transcription is supported by the 5'-RACE experiment. The masked sequences surrounding these two CAGE tags are shown in Figure 4. Which positions are masked and which positions are likely to initiate transcription in these surrounding areas is also indicated in Figure 4.

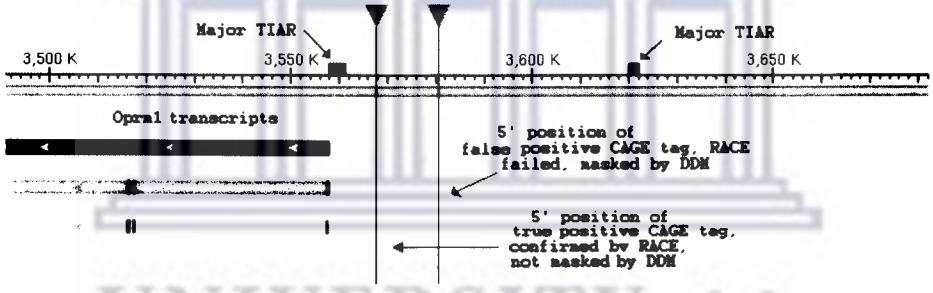

Figure 3: True and false TSSs for mouse gene Oprm1 recognised by DDM

>122BA39P0901 mm5 chr10 – 3559800.. 3559350

aaacaaccctaatgcatggttgttacaaagaaacTTAGagTCTCTCcTCCC
ACCCTCTCTGCAGCATCCCCGCTTCTGCTCCCCCCCCCCCTACCCCAATT
TACACTCCTTTACACGGAAGTGCATAGAGACTGAGGAGGCTGATTCTGAG
TTGCTTTGCGCACAATCCACTCCTTCTCTCTCCTCCCTCCCCTCTAGCCT
CTGGATCCCTCACAGCCCATGCTCCCTCCCTTCCACTCAGAGAGTGGCGC
TTTGGGGATGCTAAGGATGCGCCTCCGTGTACTTCTAAGGTGGGAGGGGG
CTACAAGCAGAGGAGAATATCGGACGCTCAGACGTTCCATTCTGCCTGCC
GCTCTTCTCTGGTTCCACTAGGGCTTGTCCTTGTAAGAAACTGACGGAGC
CTAGGGCAGCTGTGAGAGGAAGAGGCTGGGGCGCCTGGAACCCGAACACT

>119BA53D1906 mm5 chr10 – 3583000.. 3582500

tgaaatgtaagtaaagaaattatctaataaaaatatataaaaataaataaa
taaatggactcatgtgtgtgtgtgtgtgtgagagagagagagagagagag
agagagcgatctgttgatagtttggtgttttgttttctttggggggtatgg
tttgttttgtttcctttggttttttgggggggtatcgctgtattgggttgttt
ggttttttcaggaaaacttaaagttgggtgggtagggaccaggggaaggattt
tgaagaaccatggagaatgggggagaatatggtcaacatatatttaaatttta
aaatagttttaaataataaaaatataatataggtaatataatcattaaaa
atctgccataaataaggagttaaaggaaattacaaagtaagttaaaaaaa
caaaaagcaatgtcatccatatagatacatatatatataactaaatataa
tatatagcataaaagaaagtcaaacttcacaaaagtaaatatgataattaa

CAGE tags are shown in bold and yellow highlight, masked positions are shown in lower case letters and strikethrough, potential TSS are shown in capital letters

Figure 4: DDM masking around CAGE tags T10F0065AF50 and T10F006553E

Because DDM operates with a resolution of a single nucleotide, TID and TIAR locations or regions are sometimes small and clustered. For reasons of image resolution the complete TID and TIAR can therefore not be shown in Figure 3. However the exact masking situation around the CAGE tags from Figure 3 is shown in Figure 4. To illustrate TID and TIAR locations or regions on a wider scale the complete sequence of human chromosome 21 was annotated with DDM. The annotated sequence is shown in OSM to this manuscript.

In addition to the above another 5 genomic positions where transcription is indicated by the existence of a single CAGE tag (Fantom3 representative tag IDs 120BA49K1606, 081AA66D1203, 069AE29I1002, 097AA30J2305 and 112BA90K2006) were examined. These CAGE tags are derived from adipose, liver, lung, macrophage, and embryonic tissues, respectively. They

are not all located around the position of the Oprm1 gene, but come from various other locations on the genome of *Mus musculus*. The existence of a transcript could not be confirmed by 5'-RACE for any of these positions. The experiments failed to produce a valid transcript in all of the six tissues used. When checking the areas surrounding these tags with DDM, it is observed that DDM masks all but one of these positions as unlikely to initiate transcription. The unmasked location is the location of CAGE tag 112BA90K2006, which is marked as a potential TSS. This could either indicate a false positive analysis by DDM (which is probably the case) or a problem with the RACE experiment. For the other 4 cases, the claims made by DDM are in accordance with the results of the RACE experiments.

This example illustrates that DDM can help to isolate false-positive candidate-tags for further analysis by determining computationally whether a tag falls into TID or TIAR. Consequently, DDM can contribute greatly to the accuracy of transcriptional studies and, with that, also to the success of follow up studies that look at gene finding or the functional aspects of genomics.

DISCUSSION

This chapter describes the development of the Dragon TSS Desert Masker (DDM) and its application to three showcase human chromosomes. With DDM, a tool was developed that can very accurately identify a portion of DNA sequences that is highly unlikely to promote transcription initiation. This region is set apart from the region of mammalian genomes which is called transcription initiation active region (TIAR) and contains the vast majority of genuine TSSs as well as a remaining portion of sequences unlikely to initiate transcription.

This tool was applied to human chromosomes 4, 21, and 22 in order to produce an initial estimate demarcation of the regions in the human genome

where TSSs are only very sparsely present. The set of these regions was called TID. The results obtained from the application of DDM to these chromosomes suggest that over 40% of mammalian genomes represent TIDs, that is, they are highly unlikely to promote transcription initiation. The algorithm was developed in such a way that it exploits the compositional properties of those short regions of DNA that immediately surround the TSS location. These locations were determined using at least two distinct and independent pieces of experimental evidence that the TSS is in fact positioned at the location in question. Therefore, the reference data sets, containing as many genuine true TSSs for human and mouse as currently possible, are extremely accurate. Moreover, since the TSS sets contain 113,814 human and 98,682 mouse TSS location, these represent to the best of the researcher's knowledge the most comprehensive sets of TSS locations confirmed by at least two independent types of experimental evidence. These two sets contain many alternative TSSs for a large number of genes. In spite of the richness of the TSS data sets, one must be aware that they do not represent the complete TSS complement for human or mouse. Many genuine TSS locations are not included, but there is no way to assess which ones these are and how many there are.

The analysis and results demonstrate that a very large majority of locations capable of initiating transcription in mammals are concentrated within a small fraction of the mammalian genome. This is contrary to the prevalent opinion that the initiation of transcription is a process that can occur at any given place in the genome of mammals and is not restricted to a limited number of dedicated locations. Based on the large collections of transcription data available today, it was shown that transcription in mammals does not initiate randomly over the entire genome. This claim is backed up by the results that were obtained from the application of DDM to the showcase chromosomes mentioned above. Instead of initiating randomly over the entire genome, only a small portion of the genome is likely to initiate transcription for a vast majority of transcripts. For *Homo sapiens* it can be estimated that no more than 10% of the genome is responsible for

more than 90% of transcription initiation. This high concentration of TSS in a relatively small fraction of the genome justifies the separation of DNA sequences into transcription initiation deserts (TID) and transcription initiation active regions (TIAR). As a consequence, the results presented here, and the DDM tool itself, can be used to demarcate, in advance, regions of interest for studies of transcription in mammals. It will serve to eliminate the vast majority of regions where transcription initiation cannot take place and thus significantly enhance the accuracy of those studies and their follow-ups.

One can conclude from the results obtained here that over 40% of mammalian genomes can be estimated to be part of TID, that is, they contain no or almost no genuine TSSs. The remaining portion of the genome should be understood to contain all or the vast majority of genuine TSS locations. It also contains those locations that were incorrectly labelled by DDM as TSSs. This is a consequence of the imprecision of DDM and the incompleteness of the TSS data sets that were used for this study. It is therefore not claimed that every location in the portion of the genome not included in TID as predicted by DDM represents a possible TSS. Instead, it is claimed that the part that constitutes TIAR includes (almost) all genuine TSSs, but it also contains all locations that were falsely recognised as TSS by DDM, according to the specificity level of the algorithm.

In the attempt to combine masking repeats and TIDs, it was found that at the balanced sensitivity and specificity levels, the accuracy of DDM does not benefit from masking repeat regions in TIAR. The performance of DDM was shown in Figure 2 and Table 1 above. The algorithm allows for the performance to be adjusted based on a threshold value. This adjustment makes it possible to achieve a more favourable trade-off between sensitivity and specificity than by combining masking repeat regions with TIDs. This is due to the fact that repeat regions cannot be regarded as incapable of initiating transcription. While it is true that only a minority of genuine TSSs fall within a repetitive DNA sequence, the *a-priori* exclusion of repeats

from studies of transcription initiation leads to inaccuracies. These inaccuracies can be prevented through the usage of DDM.

The DDM programme would be useful for researchers working on several types of problems. These problems include, but are not restricted to promoter identification, gene annotation, data curation from high-throughput experiments and wet-lab experiment designs. All these issues are of broader interest. Promoter identification, although considerably advanced [24,44], still suffers from positionally inaccurate prediction of the actual TSS location. The problem is circular to the accuracy of the data set on which these systems are trained, as well as the coverage of the real TSSs within the data sets. In many cases the systems designed to predict TSS are trained on data sets which determine the TSS position inaccurately, which leads to shortcomings in the positional accuracy of predictions made with those tools. Another problem that exists is the fact that the data sets used for TSS prediction training are incomplete with regard to the reflection of all real existing transcription start sites. This inevitably leads to the second problem of a more accurate annotation of transcripts. The most frequently used methods for full-length cDNAs are Cap-trapper [45,46] and Oligo-capping [47]. Due to the specificity of sequences around mammalian TSSs (generally high GC% and strong secondary structures), under optimal conditions over 90% of full-length cDNA can be generated with the rest of cDNAs being non-full-length [48]. The DDM system could assist in cleaning this data from experimental artefacts and incorrect signals. Bioinformatics approaches, microarray experiments, and other high-throughput data are prone to false-positives. The genuine TSS locations have to be confirmed through wet-lab experiments (Northern hybridization, RACE, RT- or quantitative PCR) and possibly by multiple pieces of evidence. Most low-throughput but high-confidence experimental techniques require advance knowledge of specific genomic regions for probe or oligonucleotide primer design. The design of more accurate probes and oligonucleotide primers can be greatly simplified by the application of DDM before experimental validation. This would benefit the experiments

and the success of follow-up studies undertaken with the data derived from these experiments.

In the attempt to show that the available PPPs are not suitable to predict TIDs with an acceptable level of accuracy, the DDM system was compared to several existing PPPs. These PPPs constitute the standard approaches for promoter and gene start finding that are found in literature. Three of the PPPs examined in [24,25] are by design unsuitable for a comparison with DDM. These tools are CpGProD [33], Dragon Promoter Finder [34,35] and Dragon Gene Start Finder [36,37]. CpgProD is restricted to prediction of CpG-island related promoters, which make up only a subclass of all existing promoters. Therefore a comparison is not feasible, because DDM endeavours to predict all TSS, regardless of their specific structure. At the same time, Dragon Promoter Finder and Dragon Gene Start Finder determine TSSs based on averaging over a number of strong predictions. This way, a TSS prediction is generated that is unlikely to represent the real TSS with one nt accuracy, although it is likely to be very close to the real TSS. Other promoter prediction systems have other types of restrictions, as was discussed in the 'Results' section of this chapter.

The general comparison setup was based on the use of a test set that is completely independent of the training set used to derive the DDM model for the comparison tests. This introduces fairness into the comparison. Furthermore, to be able to estimate TID, the promoter predictors should be able to operate at sensitivities of ~100%. Not all promoter predictors have the possibility to adjust their tuneable parameters to achieve a value close to that sensitivity, and thus it must be concluded that they are not suitable for this task. However, for those promoter predictors that allow the adjustment of their parameters, to let them operate more closely to the high sensitivity levels demanded in this context, such adjustments were made. It is important to note that, after this intervention, such promoter predictors have reached an 'extreme setting' relative to their typical mode of operation.

The observed differences in performance comparison results come from several factors. First, DDM is capable of separating TID and TIAR predictions at the level of a single nucleotide, because the DDM algorithm is trained to pinpoint the actual TSS location. Promoter predictors frequently only indicate a region in which they expect a TSS to be present, thus reducing the resolution of the tools dramatically. Only two of the promoter predictors, NNPP2.2 and Fprom, are capable of pinpointing the TSS location exactly. All other predictors that were used only give an interval in which they claim the TSS location to be. The comparison with DDM when the exact TSS location is to be predicted (test B), shows that both NNPP2.2 and Fprom are not recognising a significant portion of the real TSS locations, making them unsuitable for TID estimation. Another issue is that the design goals of promoter predictors could be different from the design goal of DDM. DDM attempts to achieve 100% sensitivity in recognition of real TSSs and to minimise the predictions of random genomic locations as TSSs. Promoter predictors, on the other hand, generally aim at maximised balanced sensitivity and specificity, usually sacrificing sensitivity in favour of specificity. Systems tuned in such a way are not necessarily suitable for determining TID, as one needs to have guarantees that all (or the vast majority) of TSSs are included in the predicted locations. The current promoter predictors unfortunately do not provide this characteristic. It is for this reason that the settings of some of the promoter predictors had to be changed to make them operate at a very high sensitivity. It should be highlighted that the comparison results have to be interpreted with these issues in mind.

Also, the set of random DNA that was used in the comparison of DDM with other promoter predictors was assumed to contain no TSS locations. For the purpose of comparing DDM with existing promoter predictors, DDM was specifically retrained with a set of all human TSS sequences, excluding those that were used in the comparison experiment ($HTSS_{compare}$). This means that all data used in the comparison was previously unknown to DDM and no advantage for DDM through data selection was obtained. How

the various data sets that were used in this study have been created is explained in detail in the 'Methods' section of this chapter.

METHODS

Data: Transcription Start Sites

The creation of reference data sets is probably the most important step in creating a computational recognition system. Any system can only be as good as the data that was used to train it. It is therefore most essential to obtain data that is as complete and as accurate as possible. Two highly accurate sets of TSS for *Mus musculus* and *Homo sapiens* were compiled. The reference genome builds that were used for these species are mm8 and hg18. The respective surrounding sequences covering [-100,+100] relative to these TSSs were compiled. The sequence was extracted from the same strand that the TSS was reported to be residing on. A TSS was only regarded as genuine and made part of the TSS data set if it was possible to find two pieces of independent supporting evidence for the existence of a TSS in a specific genomic location. The first piece of evidence required was the presence of at least one FANTOM3 CAGE tag. This piece of evidence was considered to be backed up by a second piece of evidence, if the first 5' nucleotide of the CAGE tag coincided exactly with the first 5'nucleotide of either at least one full-length cDNA or at least one mRNA. The cDNA sequences used here are all cDNA sequences that are found either in FANTOM3 or in the UCSC browser [49]. The mRNA sequences used in this process are taken only from the UCSC browser. All TSS locations selected in this way are supported by at least two independent pieces of evidence. A minimum distance between neighbouring TSSs was not enforced as long as two pieces of evidence were present at a location. No mismatch between the two pieces of supporting evidence was allowed. This

means that TSSs which are only one nt apart are considered to be two separate TSSs.

It has been established by [50] and [51] that within promoter regions there exist many alternative TSSs that are often located within a few nucleotides from each other. In the context of this manuscript, these TSSs are regarded as separate, even if they are residing on neighbouring nucleotides. Although TSSs that are located very close to one another are likely to transcribe the same transcriptional unit, even the most minimal difference in the location of the TSS leads to the production of a slightly different transcript. Since the aim of DDM is to pinpoint the exact location of TSS and not only the approximate location of a tag cluster, these transcription events are therefore regarded as separate. Furthermore, even a small positional difference between two TSSs causes the surrounding area of the TSSs to be different, with features of this area residing at different location with regard to the TSS. While in some cases this approach has the effect that TID and TIAR appear in a clustered fashion on the chromosomal sequence (see chromosome 21 in online supporting materials), the belief is held that this could be a more true reflection of the actual biological situation with regard to transcription initiation.

Since the two pieces of evidence that are required for all true TSSs are taken from two completely independent and distinct experiments, the resulting set of genuine true TSSs has an extremely high accuracy. Sequences that contained ambiguous characters ('N') were excluded. In this way, a mouse reference TSS set containing 98,682 sequences and a human reference TSS set containing 113,814 sequences was compiled. These sets were called MTSS and HTSS respectively. From the HTSS set, a subset of 1,000 TSS locations was chosen randomly. For these TSS locations the sequences covering [-800, +800] relative to the TSS location were extracted. These 1,000 sequences were called $HTSS_{compare}$. These are used for the comparison between DDM and existing promoter prediction programmes. The set of all human TSS with all items in $HTTS_{compare}$ removed is called $HTSS_{tc}$.

The best data available at present was utilised and a rigorous methodology in establishing the reference TSS sets was applied, though it must be said that this set is naturally only a subset of the set of all genuine TSSs in mammals. While as many genuine TSSs were included in the reference data set as possible, it cannot be claimed that the reference data set possesses a complete set of all human and mouse TSSs. As a matter of fact being in possession of only a part of the genuine TSSs raises the need for predicting TID in the first place. Since it is demanded that all TSSs in the set have a CAGE tag support, there are high dependencies on the accuracy of that data, and this is one of the reasons why two independent pieces of evidence to support the TSS location were used.

Data: Other sequences

As non-TSS sequences or 'negative' sequences, DNA sequences from human and mouse were selected indiscriminately. These DNA sequences were 200 nt in length and selected randomly from all human and mouse chromosomes. In doing so it was ensured that the number of sequences selected was proportional to the length of the chromosomes. Sequences that contained ambiguous characters ('N') were discarded. If the 5' end of a CAGE tag fell within [-10, +10] relative to the centre of the sequences, the sequence was also discarded. In total 110,000 random human DNA sequences and 100,000 random mouse DNA sequences were selected. In the same manner, an additional 1,000 human DNA sequences 1600 nt in length were extracted to be used as a negative set for the comparison with existing PPPs. This set was called $RNDM_{compare}$. This means that $RNDM_{compare}$ and the negative set (RNDM) used for training of DDM are disjoint sets.

DDM training set for comparison with promoter predictors

To make the comparison of DDM with the other promoter predictors fair, DDM had to be retrained on a training set that was completely independent from the test set used in this comparison. The training set used for this purpose contained $HTSS_{tc}$ as the positive data and RNDM as the negative data. Please note that this training set was completely independent from the test set used in the comparison experiment.

DDM test set for comparison with promoter predictors

The positive and negative data set $HTSS_{compare}$ and $RNDM_{compare}$ respectively formed a test set used to assess the performances of DDM and other promoter predictors. This set is independent from the set used for the training of DDM for the comparison with the promoter predictors.

Algorithm:

To achieve the highest possible accuracy, the presented algorithm utilises a four-stage daisy-chained filtering method. The basic layout of the algorithm was presented in Figure 1. Sequences of length 200 are examined and have to be classified by all four stages as a potential TSS in order to be recognised as part of a TIAR. The algorithm uses a different filtering method at each stage. This way it is possible to exploit different compositional features of the sequence under examination. Thus the overall method achieves the very high discrimination between locations likely and locations not likely to initiate transcription.

In a formal way, the algorithm presented here can be understood to be a multi-classifier system, which is a common approach to classification problems in machine learning. Normally, the same problem is presented to a

number of individual classification modules and the overall statement is derived in some way as a combination of the results of the individual classification modules by some decision logic module. There are numerous ways in which classifiers can be combined and numerous ways to design the decision logic. DDM can be categorised as a parallel multi-classifier system, in which the decision logic outputs a negative result as soon as one of the inputs from the individual modules is negative. This is achieved indirectly by applying the individual modules of the algorithm sequentially. A step is only executed if all previous steps have deemed a sample to represent a possible TSS.

All four steps of the algorithm are performed either on the entire available data sets or on the training part of the data during the 4-fold CV.

Boundaries of k-mer distribution and frequencies of k-mers

A total number of 1,364 k-mers of length 1-5 was considered. The lengths and number of these k-mers are summarised in Table 7. The number of occurrences u of each k-mer K in the upstream segment [-100,-1] was determined, as well the number of occurrences d of K in the downstream segment [+1,+100]. These two numbers were recorded. Both values u and d are from the interval [0, 100+1-k] where k denotes the length of k-mer K. For every sequence in MTSS and HTSS with an upstream occurrence u of k-mer K, the minimum, *min(d)*, and maximum, *max(d)*, occurrence of K downstream of TSS was determined. For every sequence in MTSS and HTSS with an downstream occurrence d of k-mer K, the minimum, *min(u)*, and maximum, *max(u)*, occurrence of K upstream of the TSS was determined. This was done for all possible values u and d from [0, 100+1-k] and for all possible k-mer lengths 1-5.

1	4	4
2	16	20
3	64	84
4	256	340
5	1024	1364

Table 7: Statistics on k-mers used in development of the algorithm

For every k-mer *K*, the collection of all points defined by (*min(d)*, *u*) and (*max(d)*, *u*), as well as (*d*, *min(u)*) and (*d*, *max(u)*), define boundaries of the region that contains all TSS locations. To illustrate this point, please refer to Figure 4 which shows an example for the 1-mer 'C'. The region of all TSS is shown in grey.

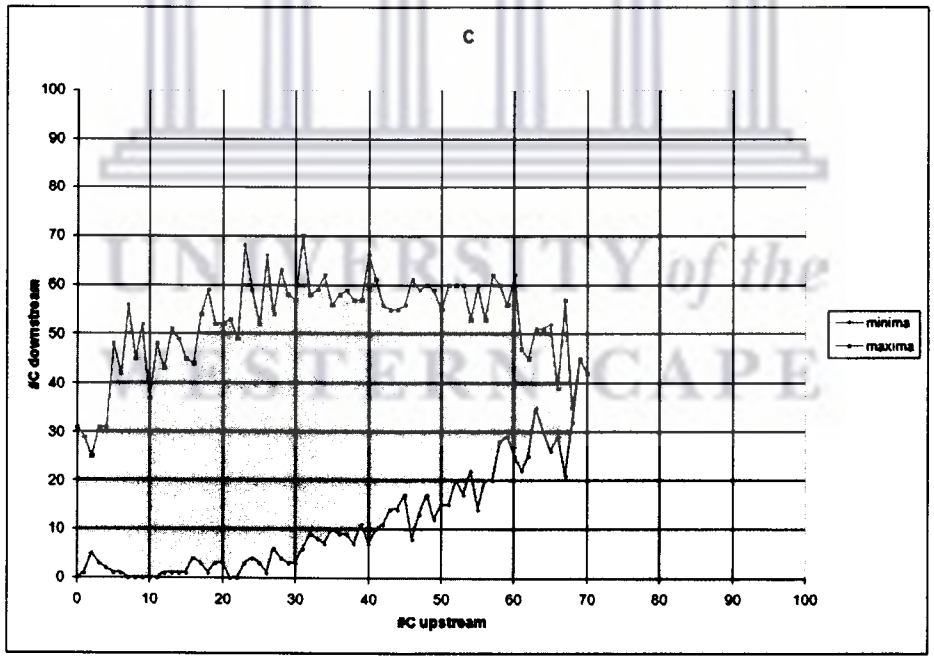

Figure 5: Constraining boundaries for occurrences of 1-mer 'C'

A particular TSS characterized by ($u1$, $d1$) for k-mers K will be recognized, if for $u1\ min(d) \leq d1 \leq max(d)$, and for $d1\ min(u) \leq u1 \leq max(u)$ is obtained. A sequence is considered to contain a TSS on position +1 if for all 1,364 k-mers it satisfies the constraining conditions above.

[-10,+10] PWM thresholding

For this step of the algorithm, the sets MTSS and HTSS are divided into 16 subsets characterised by different dinucleotides at positions [-1,+1]. All sequences with dinucleotide 'AA' at positions [-1,+1] are put together in one subset, etc. For each of these subsets, all sequences of length 20 nt covering the region [-10,+10] were extracted, and for each of the 16 such subsets, a position weight matrix (PWM) [52] was constructed. The PWM of each subset has 20 columns in correspondence with the region it covers. The PWM of a given subset is subsequently used to determine the PWM scores s_s [52] of all [-10,+10] sequences in the subset. Out of these scores the minimum score s_{min} is selected. A sample is considered to contain a TSS on position +1 if its associated PWM score $s_s \geq s_{min}$ in the respective subset.

LDF 40

For this step of the algorithm, the sets MTSS and HTSS are again divided into 16 subsets, as described above. The complete TSS region [-100,+100] for all TSSs in a given subset is divided in 40 consecutive non-overlapping sections of length 5 nt. For each of these 40 sections, a PWM as previously described was determined, using all sequences from a given subset of HTSS and MTSS respectively. For each of the sequences from all 16 subsets of HTSS and MTSS, a feature vector comprising of 40 PWM scores was determined. Each score was determined using all 40 sections of the sequence and the corresponding PWM. In this way 16 sets of 'positive' data

with one 40-element feature vector for each sample is produced. The 'negative' data was processed with the same PWMs derived from the MTSS and HTSS subsets to create 16 sets of 'negative' data with one 40-element feature vector for each sample.

Linear discriminant analysis [53] is used on these sets of 'positive' and 'negative' data to determine 16 linear discriminant functions (LDFs), one for each of the 16 subsets. A linear discriminate function for a 40-element feature vector possesses 40 coefficients, to be multiplied with the individual features, and one constant. These 41 elements of the function are meant to be summed up. The 16 LDFs that were determined as described above are then used to calculate LDF values. An LDF value is calculated using the 40 coefficients c_i, $i=1,2,...,40$, plus one constant c_0.

All sequences in HTSS and MTSS are subjected to the LDF that corresponds to the dinucleotide at positions [-1;+1] and the score s_{LDF} is calculated for each sequence ($s_{LDF} = c_1x_1 + \ldots + c_{40}x_{40} + c_{const}$, where x_i are the corresponding scores of the respective PWMs). A threshold value is determined for each of the 16 subsets in MTSS and HTSS by selecting LDF_{min} so as to preserve 100% sensitivity in the recognition of real TSSs.

A sample is considered to contain a TSS on position +1 if $LDF_{sample} \geq LDF_{min}$ for the respective subset. Otherwise, the sample is classified as not containing a TSS on +1.

SVM

For this step of the algorithm, the sets MTSS and HTSS and the 'negative sets' are processed as described above to produce positive and negative data containing 40 values for each sample. A support vector machine (SVM light: http://svmlight.joachims.org/), with a radial basis kernel function, is trained as a classifier. The radial basis gamma value 1.28 delivered the highest accuracy for this data. The class for sequences containing a genuine

TSS is labelled 1, the class for random non-TSS DNA sequences is labelled -1. The two resulting models M_{HS} and M_{MM} are derived.

A threshold value t_{SVM} is then applied. A sample is considered to contain a TSS at position +1 if the SVM score $s_{SVM} > t_{SVM}$.

The threshold t_{SVM} is the only adjustable input parameter to the tool implemented on the web server (http://apps.sanbi.ac.za/DDM/). It can be used to manipulate the sensitivity / specificity behaviour of the algorithm. For details on the sensitivity / specificity behaviour of this algorithm please refer back to the 'Results' and 'Discussion' sections of this chapter and in particular to Table 1. All other parameters of the algorithm are fixed at a level that experimentally provided maximum sensitivity. In particular, the threshold values for [+10,-10] PWM and for LDF40 functions were fixed at levels that allow the retention of 100% of the genuine true TSSs. Although it is possible to use these thresholds to manipulate the sensitivity / specificity behaviour of the algorithm, it was experimentally determined that the SVM is the step that allows the most beneficial trade-offs, and moreover, contribute to the simplicity of the parameter adjustment process. Because the SVM classification of sequences is computationally the most time consuming of all steps, it is also beneficial for the overall speed of the algorithm to place it at the end of the daisy chain.

CONCLUSIONS

In this chapter a new algorithm for masking transcription initiation deserts in mammalian genomes was presented. The algorithm has the ability to mask a significant portion of the genome as containing a minimal fraction of genuine TSS locations while retaining a vast majority of the genuine TSSs in the non-masked regions. It was shown that it can be estimated that for *Homo sapiens*, less than 10% of the genome are responsible for over 90% of all transcription initiation. This enables the focusing of research attention to

narrow segments of the genome. These segments could otherwise be difficult to identify. The great advantage of the algorithm is that it can identify transcription initiation deserts at the resolution of a single nucleotide. The server with this algorithm is freely available at: http://apps.sanbi.ac.za/DDM/. It is believed that this resource could be of wide use to researchers in different fields of life sciences. The work presented in this chapter is currently being prepared for submission for publication in *Bioinformatics*.

Chapter 2 – The analysis of transcriptional deserts

BACKGROUND

The transcription of DNA sequences into messenger RNA is the first important step on the way from DNA to the production of proteins, which determine the biological behaviour of the majority of cells. Vast collections of transcription data that are now widely available have enabled researchers to examine closely the mechanisms involved in transcription [7-12]. These studies have shown that for a large number of genes there exist numerous alternative transcription start sites which contribute greatly to the fact that a single gene can produce multiple sometimes vastly different transcripts and in turn different gene products [54]. It was also shown that a large number of genes have the ability to produce various gene products by means of post transcriptional modifications, such as alternative splicing [55,56]. Here, the location of splice sites is crucial. It has been proposed that errors in the selection of the correct transcription start site under specific circumstances, disturbances in the process of transcription regulation, or erroneous splicing activities, are involved in the development of diseases and genetic disorders [57-59]. Much research attention has been given to regions that promote the initiation of transcription and to regions that are themselves transcribed. However, the exact control mechanisms that lead to the usage of one or another alternative TSS remain elusive. It is suggested that a variable number of different control elements work together to produce a specific transcript under specific circumstances. It appears that in this process the distances between the various control elements and the gene being regulated can be large and that regulatory elements do not necessarily regulate genes in their immediate neighbourhood [60]. The malfunctioning of one of those control elements might have consequences for the entire transcription procedure.

While research attention is concentrated on transcriptionally active regions of the genome, not much attention at all is given to regions of the genome that are not transcriptionally active. In the last few years, it has been accepted that 'most' of the mammalian genome is in fact transcribed [61-63] and that only a minority of those transcripts are translated. In this context, research attention is also concentrated on 'non-coding regions'. There is, however, a portion of the genome that is not transcribed and which has hitherto been largely ignored. Researchers regularly do not consider those regions to be important and ignore them in their studies. While this is a valid and correct assumption in many cases, it does not contribute to the elucidation of the specific functions of these regions. The genomes of all organisms living today have evolved over several billions of years [64], so the existence of regions in the genome that do not play an active role in transcription, implies that they might have a function after all that has so far escaped the grasp of the research community. This is especially interesting considering that in many cases distal regulatory DNA elements have a direct or indirect influence on gene transcription and expression. It seems appropriate to assume there is indeed a reason for the existence of those DNA stretches. It is also worth considering that the complexity of transcription and the amount of protein interaction is thought to be mainly responsible for the complexity of higher organisms such as mammals, and not the size of their genomes or the number of genes contained within these genomes [65,66]. Therefore, the role of regions that are not directly transcriptionally active is worth investigating. The suggestion that genomic regions, which are not immediately involved in transcription, have only structural roles or are evolutionary leftovers might be an underestimation and too simple a view-point considering the vastly complex process of transcription regulation and gene expression.

Fully understanding the role that transcriptionally passive DNA has in the genome, and trying to illuminate the connection between this DNA and protein-coding genes, might help to understand the ways in which certain types of genetic disorders originate and proliferate. Once the regular

function of transcriptionally passive DNA is determined, it can be explored what effects a malfunctioning of the processes involved has on the normal operation of the cell.

This study contributes an initial estimate towards an investigation of potential roles that transcriptionally inactive DNA might play in mammalian genomes. A comprehensive methodology for accurately determining part of those regions of the genome that are not directly involved in transcriptional activity was developed. This methodology allows the relatively accurate distinction of regions that are either evidently transcribed, or that might promote the initiation of transcription from those regions that possess neither property. The methodology was applied to the genomes of *Mus musculus* and *Homo sapiens* and these regions were extracted. These regions are termed 'transcriptional deserts' (TDs) and they are studied in this chapter.

In previous studies that have examined 'gene deserts' the areas under investigation were define simply as intergenic regions. Results of these studies have been contradictory with some suggesting that gene deserts have no particular function and can be deleted without consequence for the viability of the organism [67] while others come to the conclusion that these regions contain remote control elements for gene expression [68-71]. In contrast to the above approach, the present study is much stricter in its definition of 'deserts' and analyses only those regions which are not directly involved in transcriptional activity. These regions are termed 'transcriptional deserts' (TDs), and contain sequences of genomic DNA that are neither themselves transcribed nor represent the locations of TSSs.

Subsequently these transcriptional deserts are subjected to various kinds of analysis in order to investigate their specific characteristics. Special attention is given to the way in which transcriptional deserts display distinct properties that differentiate them from transcriptionally active genomic regions. TDs are examined with respect to their compositional properties and their GC content. GC-richness or AT-depletion is a property of DNA

that is frequently employed by researchers to explain gene-richness and gene-depletion of genomic regions. Here it is shown that, while it is known that the GC-richness is correlated with the higher gene density in a DNA region, GC-richness itself is not sufficient to fully explain the presence or absence of transcriptional activity.

Furthermore, TDs are examined for the occurrence of single nucleotide polymorphisms (SNPs) and their rate of evolutionary conservation. It is generally accepted that the rate of evolutionary conservation gives an indication of the level of function between different areas within the genome. It is normally thought that regions that are evolutionarily conserved are more functional than those that are not. While it does not necessarily follow that regions that are not evolutionarily conserved are only of minor significance, knowledge about the rate of evolutionary change in TDs, in comparison with non-TDs, shows to what extent TDs are involved in processes that are conserved over time in mammals.

Transcriptional deserts are also examined for the existence of transcription factor binding sites (TFBSs) [72], which could produce important insights into the regulatory activity of TDs. Consequently, combining the knowledge about the existence of TFBSs and SNPs, several candidate regions are presented where a collection of mutations might cause differences in the way in which the transcription factors that bind to the region influence the transcription of genes. Such disturbances might be involved in irregular transcriptional activity and might, among other things, contribute to explaining the origin of neoplasia.

METHODS

To identify transcriptional deserts in the genomes of *Homo sapiens* and *Mus musculus*, the genomic sequences of these species was subjected to a multi-staged analysis. The aim of this analysis was to extract genomic regions that

have two distinct properties. Firstly, these regions must, with a very high probability, be incapable of initiating transcription. This assures that TDs do not contain any transcription start sites, which are supposed to be excluded from the analysis. To do this, the Dragon TSS Desert Masker (DDM) was applied to the forward and reverse strand of all chromosomes of human and mouse.

The Dragon TSS Desert Masker (DDM) was introduced in Chapter 1 and has the ability to determine very accurately those regions of mammalian genomes that are highly unlikely to initiate transcription. A sensitivity level of 99% was chosen and all TSS deserts in the forward and reverse strand of all chromosomes of *Homo sapiens* and *Mus musculus* were masked. At this level, DDM has a specificity of 87% for mouse and 86% for human. This means that regions recognised as 'transcription initiation deserts' (TID) at these performance levels will contain only 1% of genuine TSSs and will cover 87% or 86% of all genomic locations in human and mouse genomes respectively. At the same time regions recognised as 'transcription initiation active regions' (TIAR) at the above mentioned performance levels will contain 99% of all genuine TSS and will be localised within 13% and 14% of all genomic locations in human and mouse respectively. Subsequently all those regions recognised as TID were studied. Regions that were part of the TIAR were eliminated from the chromosomal sequences. This was done for the forward and reverse strand of each chromosome separately.

Secondly, it must be ensured that the transcriptional deserts are not themselves transcribed. For this purpose a comprehensive collection of transcription data that is known to exist in humans or mice today (April 2008) was compiled. These compilations of transcript data are summarised in Table 8.

ESTs	6,564,306	4,072,781	UCSC
l-SAGE	275,021	527,129	UCSC
CAGE	2,808,513	1,776,667	FANTOM3
flcDNA	1,118,025	416,489	DBTSS
mRNA	171,372	195,600	UCSC

Table 8: Summary of all transcripts used in TD production

This compilation contains all known ESTs for human and mouse on the one hand, as well as two different types of full-length transcript data (l-SAGE and flcDNA) on the other. The data in these collections is taken from three different sources. For these two reasons, it can be assumed that these compilations contain as complete a collection of human and mouse transcripts as can be obtained at the time this analysis was performed (April 2008). Data that is published in the time since this analysis has been performed (e.g. [61-63]) and data that will be published in the future are of course more up-to-date than the data used.

The genomic coordinates, with start and end position, as well as chromosome and strand, are extracted for every transcript shown in Table 8. These positions are also eliminated from the chromosomal sequences, in order to mask out all positions in the genome for which the existence of a transcript can be shown. A position is eliminated as soon as there is one transcript that occupies the position in question. This was done for the forward and reverse strand of each chromosome separately. From the remainder of the chromosomal sequences, all those regions are extracted where more than 518 consecutive nucleotides are neither characterised as likely to initiate transcription nor part of any known transcript. This effectively makes 518 a required minimal length for transcriptional deserts. Since a strong clustering behaviour is observed in the occurrence of potential TSS, a minimum lengths for TDs needs to be enforced. The threshold of 518 nucleotides was chosen, because 95% of all human full-

length cDNA sequences are longer than 518 nucleotides. Therefore, the shortest transcriptional desert can only accommodate the shortest 5% of all human flcDNA sequences.

The resulting regions exist for the forward and the reverse strand of each chromosome separately. They constitute all those regions that are a) highly unlikely to initiate transcription; b) for which no know transcript exist; and c) which have a minimal length of 518 nucleotides.

Here, TDs are not considered to be strand-specific. Instead, in the context of this study, a TD is defined as a region of the chromosome where a transcriptional desert exists on both strands. This means that corresponding nucleotides on both strands are neither transcribed nor likely to initiate transcription. If this condition is satisfied, the chromosomal position in question is considered part of a TD. Again, only those TDs that consist of at least 518 consecutive nucleotides are considered. These non-strand-specific TDs are extracted for all chromosomes in the genomes of mouse and human and subsequently subjected to various types of analysis.

In order to investigate in how far the application of a specific minimal length affects the occurrence of TDs, the complete process of creating TDs was repeated. This time a minimal length of only 259 nucleotides was enforced. Statistics regarding the number of TDs identified, as well as the chromosomal coverage of these shorter TDs, have been composed in order to compare them with the original TDs of minimal length 518. It has to be admitted at this point that the selection of a specific minimal length for TDs is, to some degree, arbitrary. Unless stated otherwise, all subsequent analysis was conducted with the TDs of minimal length 518.

The occurrence of all possible k-mers of length 1 to 8 (monomers to octamers) was determined in TDs. This yielded the GC-content as a by-product by producing the proportion of 1-mers 'G' and 'C'. In order to investigate if there exists a specific k-mer composition in TD regions, the k-mer composition of TD regions was compared with the k-mer composition

of randomly extracted DNA. For *Homo sapiens*, a number of DNA sequences were randomly extracted from the genome, with the number of sequences extracted from each chromosome proportional to the size of the chromosome. Only random DNA sequences that had a similar (±1%) GC-content to the TD regions in *Homo sapiens* were considered. The total number of sequences extracted corresponds to the total number of TD regions identified in *Homo sapiens*. The length of each extracted sequence corresponds to the average TD length in *Homo sapiens*. The same was done for *Mus musculus*. Using the chi-square method, the p-values and corresponding chi-square values were calculated for the distribution of all k-mers between TDs and random DNA, assuming 4 categories of k-mers, k-mer x in TD, non-x in TD, x in random DNA and non-x in random DNA. P-values were calculated based on the null-assumption that no difference can be observed between the k-mer distributions in TDs and random DNA. The selection of random DNA of the same amount and with the same (average) length and GC-content as the TDs in the corresponding species makes it unlikely that possible differences observed in k-mer composition caused by sequence properties other than lack of transcriptional activity in TDs.

RepeatMasker [http://www.repeatmasker.org] was used to analyse the repeat content in human and mouse TDs, in comparison with the repeat content in their entire genomes.

All known human and mouse single nucleotide polymorphisms were extracted from dbSNP (NCBI built 129) and separated into SNPs lying within TDs and those lying outside of TD regions. Rate matrices for SNPs lying inside and outside of TDs were determined. The proportion of SNPs that fall within TDs was compared to the proportion that fall outside transcriptional deserts. The proportion of SNPs that fall within TDs was then set in relation to the proportion of the human and mouse genomes that, according to this analysis, are covered with TDs.

The alignment between the human genome (version hg18) and the mouse genome (version mm8) was downloaded from the University of California

in Santa Cruz (UCSC Genome Browser, http://genome.ucsc.edu/ [49]). It is estimated that a common ancestor for mice and humans lived during the Cretaceous period [73]. This means that the mouse and human genomes have been subjected to more than 70 million years of independent evolution. For the context of this study, sequences that are conserved in the genomes of both species are for this reason considered as evolutionarily conserved. This alignment between the human and mouse genomes was used in this study and was compiled with the tool BLASTZ [74]. The alignment data comprises all sequences in the human genome that can be matched to corresponding sequences in the mouse genome. The genomic locations, as well as the sequences themselves, are contained in the downloaded data. For each matched sequence, a similarity score is given. This similarity score is specific to BLASTZ and explained in the corresponding publication [74]. In order to investigate the extent to which TD regions are evolutionary conserved, it was determined what portion of human TDs fall into regions for which a match can be found in the mouse genome. This was done for three different minimum BLASTZ similarity scores (0, 5000 and 10000), which refer to a weak, a medium, and a strong conservation.

All TDs were analysed with 'MATCH' [75] (TRANSFAC 11.4 [76]), with standard settings for vertebrate and optimised for the minimisation of false-positive matches. This delivered the number and density of possible binding sites for matrices derived from the binding sequences of a group of known transcription factors (TFs). Only high quality matrices were used in this analysis. For comparison, the same was done for the entire sequence of human chromosome 21, the collection of human and mouse cDNAs as shown in Table 8 above, as well as a sequence of randomly generated DNA of length 1,000,000. The latter was created by randomly selecting and concatenating one million, randomly selected, single letters out of A, C, G and T.

In order to evaluate the extent to which TFBSs and SNPs coincide within TD regions, the results of the analyses regarding SNP and TFBS were

subsequently combined, thus determining how many binding sites harbour a SNP in desert regions. As part of this analysis, the positions of SNPs and TFBS in TD regions were matched and then discriminated between SNPs located in the peripheral area of the TFBS or in the core area of this binding motif. It was similarly examined how strongly clustered TFBSs appear in TD regions. For this purpose the occurrences of all possible TFBS in a sliding window of size 200 nt were counted. The sliding window was moved by one nucleotide along all TD regions of the respective organism. The mean value for TFBS occurrences in all sliding windows and the standard deviation from this mean value was calculated. It was also calculated to what extent there are significant outliers present from this statistic. The results from these investigations were compared to the respective characteristics in promoter regions.

RESULTS

DDM was applied to the human and mouse genome at a sensitivity level of 99%. This means that only 1% of true TSS from the test set is not correctly recognised. At this level, 78.6% of human chromosome 21, 57.6% of human chromosome 22 and 86.4% of human chromosome 4 are marked as positions where transcription is very unlikely to initiate. A complete overview of what portions of the individual chromosomes are recognised as unlikely to initiate transcription at this level and what portions are likely to be TSSs is presented in Table A1 in the appendix. For the determination of all of values in Table A1, only those positions have been taken into consideration which do not possess a character other than 'A', 'C', 'G' or 'T' (e.g. 'N') within 100 nucleotides upstream and downstream have been taken into consideration. Any sequences of length 200 containing characters other than 'A','C','G' or 'T' (e.g. 'N') have been left out of the calculation of the portions likely and unlikely to initiate transcription.

Desert size and coverage

Based on the results regarding regions unlikely to initiate transcription, obtained through the application of DDM, all those regions that are neither likely to initiate transcription nor are part of a known transcript were determined. All DNA sequences from the genomes of human and mouse that meet these criteria and were of a minimal length of 518 nucleotides were extracted. These regions where termed 'transcriptional deserts' (TDs). Details on the procedure that was followed to achieve this can be found in the 'Methods' section of this chapter. Statistics on the TDs for all chromosomes of the mouse and human genome can be seen below in Tables 9a. The actual TD regions for all chromosomes for mouse and human are presented in the online supporting materials (OSM) to this study.

species and chr	nof deserts:	tot. length:	longest:	shortest:	avg. length:	median length:	% of chr covered:	size:	GC-content:
HS chr1	6147	4776186	6435	518	776	682	1.93	247249719	29.21%
HS chr2	8815	6852882	4520	518	777	680	2.82	242951149	29.40%
HS chr3	6660	5177077	2947	518	777	685	2.60	199501827	28.92%
HS chr4	12195	9538252	4657	518	782	685	4.99	191273063	28.68%
HS chr5	8153	6270066	3179	518	769	673	3.47	180857866	29.04%
HS chr6	7109	5483736	4084	518	771	679	3.21	170899992	29.25%
HS chr7	4921	3775130	3370	518	767	679	2.38	158821424	29.35%
HS chr8	4794	3703951	5061	518	772	677	2.53	146274826	29.82%
HS chr9	3578	2703174	2834	518	755	667	1.93	140273252	29.63%
HS chr10	3473	2699576	4292	518	777	676	1.99	135374737	30.38%
HS chr11	4392	3367264	6891	518	766	674	2.50	134452384	29.61%
HS chr12	3323	2592040	4529	518	780	677	1.96	132349534	29.84%
HS chr13	7017	5536876	21313	518	789	696	4.85	114142980	28.72%
HS chr14	2905	2244172	2934	518	772	679	2.11	106368585	29.29%
HS chr15	1130	858415	3283	518	759	665	0.86	100338915	30.02%
HS chr16	1160	889421	5178	518	766	666	1.00	88827254	31.65%
HS chr17	804	603956	3335	518	751	656	0.77	78774742	30.16%
HS chr18	3515	2751064	3311	518	782	686	3.61	76117153	29.41%
HS chr19	267	216084	3814	518	809	670	0.34	63811651	35.37%
HS chr20	824	618171	3717	518	750	669	0.99	62435964	31.10%
HS chr21	2005	1538810	4396	518	767	679	3.28	46944323	29.37%
HS chr22	328	243684	2577	518	742	656	0.49	49691432	36.61%
HS chrX	10265	7955221	9966	518	774	678	5.14	154913754	30.15%
HS chrY	2774	2291630	25805	518	826	688	3.97	57772954	31.53%
HS whole genome	106554	82686838	25805	518	776	680	2.68	3080419480	30.27%
MM chr1	7651	5388830	4600	518	704	648	2.73	197069962	32.07%
MM chr2	5282	3663748	6780	518	693	639	2.01	181976762	32.72%
MM chr3	6939	4841853	3900	518	697	644	3.03	159872112	32.06%
MM chr4	5342	3773523	12330	518	706	647	2.43	155029701	32.67%
MM chr5	4218	2941789	3771	518	697	641	1.94	152003063	33.00%
MM chr6	4814	3346059	4008	518	695	640	2.24	149525685	32.74%
MM chr7	2566	1739880	3521	518	678	633	1.20	145134094	33.93%
MM chr8	4145	2849009	3355	518	687	639	2.16	132085098	33.14%
MM chr9	2989	2118934	21148	518	708	634	1.71	124000669	34.16%
MM chr10	4389	3065051	3401	518	698	645	2.36	129959148	32.50%
MM chr11	2257	1548624	2171	518	686	634	1.27	121798632	33.48%
MM chr12	3938	2751254	4136	518	698	645	2.28	120463159	32.16%
MM chr13	2908	2036070	2561	518	700	643	1.69	120614378	33.33%
MM chr14	5481	3942343	11230	518	719	654	3.18	123978870	30.87%
MM chr15	3935	2787665	11194	518	708	645	2.69	103492577	31.78%
MM chr16	3574	2522850	2692	518	705	646	2.57	98252459	31.34%
MM chr17	2196	1545791	5144	518	703	645	1.62	95177420	33.01%
MM chr18	2831	1965569	7148	518	694	637	2.17	90736837	33.04%
MM chr19	1003	700953	6885	518	698	635	1.14	61321190	34.32%
MM chrX	8385	6008969	17549	518	716	652	3.63	165556469	31.37%
MM chrY	93	74811	7375	519	804	629	0.47	16029404	31.52%
MM whole genome	84936	59613575	21148	518	701	644	2.25	2644077689	32.63%

Table 9a: Statistics on TDs for *Homo sapiens* and *Mus musculus* for minimal TD length 518

Table 9a shows, for each human and mouse chromosome, the number of TD regions located on that chromosome, the total length of the chromosomal sequence covered by TDs, as well as the percentage of the chromosomal sequence that is covered by TDs. It also shows the longest and shortest TD for each chromosome as well as the mean and median TD lengths. It is observed that the minimum required length for TDs (518 nucleotides) was

applied everywhere and makes up the shortest TDs in all cases. It is further observed that for *Homo sapiens* a total of 106,554 distinct TD regions were located, while for *Mus musculus* fewer regions (84,936 distinct TD regions) were found. This is partly due to the smaller genome size of *Mus musculus* in comparison with *Homo sapiens*. Looking at the proportions of the genome that are occupied by TDs, one notices that the proportion of the human genome that is covered by deserts (2.68%) is larger that of the mouse chromosome (2.25%). This concordance is in fact artificial as will be shown when the minimal TD length is reduced.

For comparison, statistics were also created for TDs of a minimal length of 259 nt. This constitutes half the minimal size that was originally applied (518 nt). These statistics can be seen below in Table 9b.

species and chr	nof deserts:	tot. length:	longest:	shortest:	avg. length:	median length:	% of chr covered:	size:	GC-content:
HS chr1	21848	10349685	6435	259	473	391	4.19	247249719	34.96%
HS chr2	31923	15053872	4520	259	471	390	6.20	242951149	34.64%
HS chr3	23793	11272909	2947	259	473	392	5.65	199501827	34.32%
HS chr4	40462	19670073	4657	259	486	402	10.28	191273063	33.66%
HS chr5	29234	13797019	3179	259	471	394	7.63	180857866	34.38%
HS chr6	24940	11849582	4084	259	475	395	6.93	170899992	34.41%
HS chr7	17813	8366937	3370	259	469	390	5.27	158821424	34.89%
HS chr8	18054	8390458	5061	259	464	384	5.74	146274826	35.00%
HS chr9	13820	6340563	2834	259	458	385	4.52	140273252	35.02%
HS chr10	12754	5998437	4292	259	470	390	4.43	135374737	35.29%
HS chr11	16600	7712609	6891	259	464	388	5.74	134452384	34.92%
HS chr12	12726	5918566	4529	259	465	384	4.47	132349534	35.22%
HS chr13	22419	11055662	21313	259	493	406	9.69	114142980	33.54%
HS chr14	10364	4889981	2934	259	471	390	4.60	106368585	34.50%
HS chr15	4713	2117113	3283	259	449	372	2.11	100338915	35.80%
HS chr16	4978	2224895	5178	259	446	370	2.50	88827254	36.66%
HS chr17	3420	1517511	3335	259	443	369	1.93	78774742	36.07%
HS chr18	13162	6181336	3311	259	469	388	8.12	76117153	34.53%
HS chr19	1121	513396	3814	259	457	368	0.80	63811651	37.86%
HS chr20	3590	1586352	3717	259	441	369	2.54	62435964	36.28%
HS chr21	6810	3254438	4396	259	477	400	6.93	46944323	34.01%
HS chr22	1273	576353	2577	259	452	380	1.16	49691432	38.39%
HS chrX	35031	16811220	9966	259	479	397	10.85	154913754	34.84%
HS chrY	9000	4519493	25805	259	502	405	7.82	57772954	35.73%
HS whole genome	379848	179968460	25805	259	473	392	5.84	3080419480	35.20%
MM chr1	37504	15761002	4600	259	420	360	8.00	197069962	36.41%
MM chr2	27167	11278344	6780	259	415	358	6.20	181976762	36.47%
MM chr3	33531	14083351	3900	259	420	360	8.81	159872112	36.07%
MM chr4	26286	11059780	12330	259	420	360	7.13	155029701	36.53%
MM chr5	20396	8566437	3771	259	420	362	5.64	152003063	36.88%
MM chr6	24380	10148861	4008	259	416	360	6.79	149525685	36.46%
MM chr7	14899	5976614	3521	259	401	349	4.12	145134094	37.45%
MM chr8	21424	8840201	3355	259	412	357	6.69	132085098	36.73%
MM chr9	16044	6611094	21148	259	412	352	5.33	124000669	37.35%
MM chr10	22056	9203924	3401	259	417	360	7.08	129959148	36.36%
MM chr11	13129	5310276	2171	259	404	353	4.36	121798632	37.22%
MM chr12	19718	8239654	4136	259	417	360	6.84	120463159	36.27%
MM chr13	16091	6585115	2561	259	409	353	5.46	120614378	36.95%
MM chr14	23381	10225045	11230	259	437	372	8.25	123978870	35.78%
MM chr15	18823	7986669	11194	259	424	364	7.72	103492577	36.33%
MM chr16	16522	7036822	2692	259	425	364	7.16	98252459	35.86%
MM chr17	10743	4508497	5144	259	419	359	4.74	95177420	36.72%
MM chr18	15533	6366836	7148	259	409	354	7.02	90736837	36.50%
MM chr19	5820	2353703	6885	259	404	348	3.84	61321190	37.17%
MM chrX	38303	16408161	17549	259	428	363	9.91	165556469	35.78%
MM chrY	468	204213	7375	259	436	357	1.27	16029404	36.56%
MM whole genome	422218	176754599	21148	259	418	359	6.68	2644077689	36.56%

Table 9b: Statistics on TDs for *Homo sapiens* and *Mus musculus* for minimal TD length 259

Using the minimal TD length of 259, it is observed that the number of TDs on the genomes of *Homo sapiens* and *Mus musculus* sharply increases to 379,848 distinct TD regions for human and 422,218 distinct TD regions for mouse, while the average and median TD length fall below the previously applied minimal length of 518. The portion of the genomes that are covered by TDs increases accordingly to 5.82% for *Homo sapiens* and 6.68% for *Mus musculus*. Evidently the portion of the genome that is covered by TDs

is now larger in mouse than in human, while, when looking at larger TD regions, the human genome is to a larger extent covered by TDs. It can therefore be concluded that the minimal length that is applied in the creation of the TD regions has a strong influence on the occurrence of TD regions on the genomes of *Homo sapiens* and *Mus musculus*. It can also be concluded that there are differences in TD occurrence between the two species under investigation and that these differences are related to the minimal TD length that is applied during the creation of the TD regions.

It is remarkable that for both mammalian species under examination the chromosome that is most rich in TDs is chromosome X. While this can in part be explained by the relative AT-richness and the relative gene-depletion of the X chromosome in the two mammalian species under examination, there are – for both species – chromosomes in their genome which are even more AT-rich and have even fewer genes relative to their size in comparison to the species' X chromosome. It can therefore be speculated that the relative richness in TDs in the X chromosome is connected to the distinct characteristics of the sex chromosomes.

GC-content

Another value that is shown in Table 9a is the GC-content of the transcriptional deserts on the individual chromosomes. Genome-wide, the GC-content of the TDs is ~30% for human and about 32% for mouse. The GC-content increases to on average 35.2% in human and 36.6% in mouse when smaller TDs are considered (Table 9b). This means that larger TD regions are more likely to be AT-rich than smaller ones. Overall the GC-content in TDs is significantly lower compared with the overall GC-content of the human and mouse genomes. The overall GC-content of the human genome is 41.5% while that of the mouse genome is 41.7%. While these values only differ by 0.2%, the values for GC-content of the TDs on the

respective genomes differ by 10 times as much. This means that while the overall genomes are very similar with regard to their GC-content, there are differences in the GC-content of the TDs on those genomes. The fact that the transcriptional deserts in the mouse genome are richer in GC-content than those of the human genome is therefore not a consequence of a high GC-content in the whole mouse genome. The TDs on the mouse genome are not richer in GC, because the mouse genome in general is richer in GC. It seems more likely that there is a connection between the lower TD content of the mouse genome and the GC-content of the TDs that can be identified on the mouse genome. It can be speculated that the lower TD content of the mouse genome signifies that transcriptional activity is denser in the mouse genome compared to the human one. Similarly, as the GC-content of the TDs in the mouse genome is higher than the GC-content in the TDs in human genome, it can be speculated that GC-content plays a larger role for transcriptional activity in the human genome than it does in the mouse genome. This is due to the observed positive correlation between GC-content and transcriptional activity.

GC-content is normally used as a convenient way of explaining gene-richness or the absence of genes in DNA of vertebrates. AT-rich regions are understood as not likely to transcribe, while GC-rich regions are considered to be regions of interest for studies of transcription. Our observation that the GC-content of transcriptional deserts is, with around 31% when applying 518 nt minimal TD length, significantly lower than that of the overall mammalian chromosomes (human genome: 41.5%, mouse genome: 41.7%.) is consistent with this well-established fact. However, if one looks closely at the individual TD region, it becomes obvious that GC-richness is neither a necessary nor a sufficient condition for a region of DNA to be transcriptionally active. While GC-richness is in many cases a good indication of transcriptional activity, it cannot serve as anything more than that. Many TDs can be identified that are very rich in GC nucleotides and nevertheless not transcribed. On the other hand some AT-rich regions are also part of transcriptionally active regions. Figure 6 plots the number of

TDs in the human and mouse genome against their GC-content. The TDs with a minimal length of 518 nt are used here.

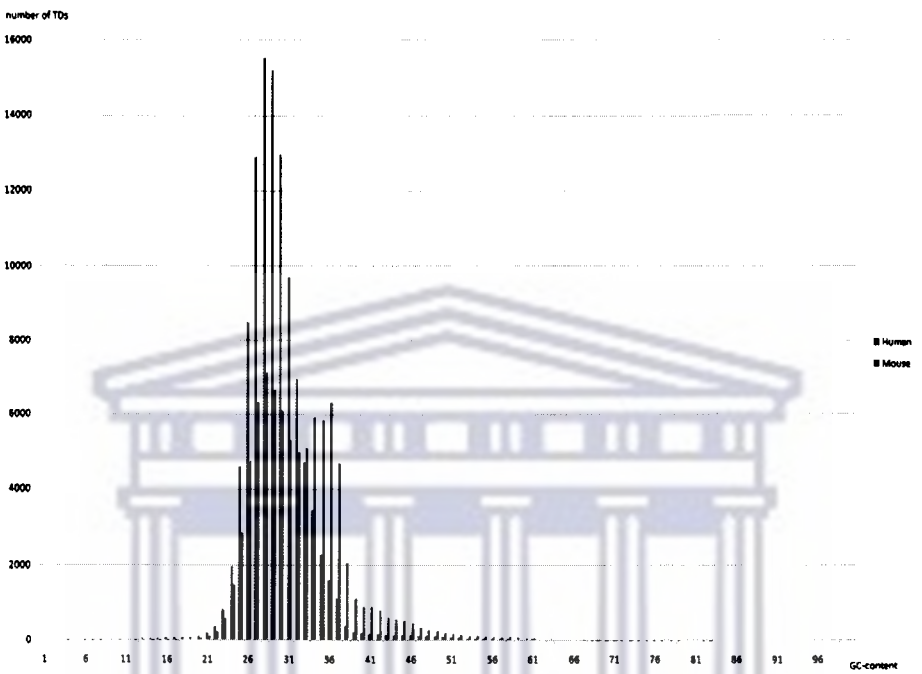

Figure 6: GC-content in human and mouse TDs

It can be seen that most TDs have a GC-content lower than the genome-wide average of about 41%. However, there exist numerous TDs whose GC-content is higher than the genome-wide average. Again there are differences between mouse and human, with the significance of high GC-content TDs in human being lower than in mouse, which is in agreement with the observation that the GC-content of mouse TDs is in general higher than in human. It is also consistent with the claim that GC-content is more relevant for transcriptional activity in humans than in mice. Figure 6 also shows that the distribution of GC-content is unimodal for human and bimodal for mouse. However, it must be said that this might be an artefact of applying a minimal desert length of 518 nt.

An example for a transcriptional desert with a high GC-content is on human chromosome 21 between nucleotides 14,280,730 and 14,281,941. The GC-content in this region is 67%.

K-mer composition

A comparison of k-mer composition in TD regions and randomly extracted DNA was analysed. The random DNA that was used in this comparison was chosen to have a similar GC-content to the TD regions of the corresponding organism. For this analysis, the TD regions, for which a minimal length of 518 nt was enforced, were used. The GC-content of these regions is ~30% for human and 32% for mouse, so the GC-content of the randomly extracted DNA was chosen to be between 29% and 31%, and between 31% and 33% respectively. As a consequence, the composition of 1-mers ('A', 'C', 'G' and 'T') is very similar between TDs and the random DNA. Tables with the complete analysis of k-mer composition for *Homo sapiens* and *Mus musculus*, including their comparison with random DNA of the respective organism, can be found in the online supporting materials to this dissertation (OSM Tables 1 and 2). The enrichment of all k-mers between TDs and random DNA sequences is also shown in these tables, along with p-values that reflect the likelihood that these distributions could be obtained by chance given the assumption that the k-mers distribution in TDs is equal to the k-mer distribution in random DNA. If this hypothesis is to be rejected for $p<0.05$, it can be determined that out of 87,380 evaluated k-mers the distribution of 59,302 and 73,225 k-mers is statistically significantly different between human and mouse TDs and random DNA respectively.

It was found that there are 89 k-mers that do not appear in human TDs, and 22 that appear neither in TDs nor in random DNA, which leaves 67 k-mers that do not appear in TDs but do appear in random DNA. In no case, however, is the non-appearance of these 67 k-mers deemed to be statistically significant. 342 k-mers do not appear in mouse TDs and 56 appear neither in TDs nor in random DNA, which leaves 286 k-mers that do

not appear in mouse TDs but do appear in random DNA. Out of these 286 k-mers only 38 are deemed to be statistically significant at $p < 0.05$.

For *Homo sapiens*, there are 108 k-mers that appear in TDs more than 10 times as often than in random DNA and 2026 k-mers that appears more than twice as often. Similarly, there are 1049 k-mers that appear more than twice as often in random DNA than in TD regions. For *Mus musculus*, there are 220 k-mers that appear in TDs more than 10 times as often than in random DNA and 3035 k-mers that appear more than twice as often. Also, there are 7457 k-mers that appear more than twice as often in random DNA than in TD regions. This means that the majority of k-mers (84,305 for human and 76,888 for mouse) are relatively evenly distributed between TD regions and random DNA. Their relative enrichment in TDs compared to random DNA is between 2.0 and 0.5. Bearing in mind that the k-mer composition of TDs was compared with the k-mer composition of sequences with a similar GC-content, it can be concluded that the k-mer composition is not extravagantly special in TD regions beyond the differences that can simply be explained with the TD regions' low GC-content.

Repeats

RepeatMasker was used to analyse the content of repetitive DNA in transcriptional deserts. Repeats are generally not thought to be of high significance to gene expression and cell activity. Their role was already discussed in the 'Repeats and transcription initiation deserts' section in Chapter 1. It is observed that for the human genome ~56% of TDs consist of repeats. For mice the repeat content of TDs is ~50%. This means that roughly half of the transcriptional deserts consist of sequences that cannot be classified as repetitive DNA. Table 10 shows a comparison between the repeat analysis of the TDs on human chromosome 21 and the entire chromosomal sequence of this chromosome. This can serve as a showcase for the repeat situation in the entire human genome and the related TDs.

file name: human chromosome 21
sequences: 1
total length: 46944323 bp (34170146 bp excl N/X-runs)
GC level: 40.88 %
bases masked: 15877801 bp (33.82 %)

	number of elements*	length occupied	percentage of sequence
SINEs:	16181	3883491 bp	8.27 %
ALUs	11935	3253853 bp	6.93 %
MIRs	4211	624582 bp	1.33 %
LINEs:	9771	6258173 bp	13.33 %
LINE1	6371	5272852 bp	11.23 %
LINE2	2848	854718 bp	1.82 %
L3/CR1	396	84766 bp	0.18 %
LTR elements:	7665	3798212 bp	8.09 %
MaLRs	4234	1853897 bp	3.95 %
ERVL	1592	767625 bp	1.64 %
ERV_classI	1615	1074433 bp	2.29 %
ERV_classII	79	70039 bp	0.15 %
DNA elements:	3716	1050750 bp	2.24 %
MER1_type	2020	452969 bp	0.96 %
MER2_type	629	348296 bp	0.74 %
Unclassified:	74	40323 bp	0.09 %
Total interspersed repeats:		15030949 bp	32.02 %
Small RNA:	102	9528 bp	0.02 %
Satellites:	66	208305 bp	0.44 %
Simple repeats:	5276	414117 bp	0.88 %
Low complexity:	5072	225798 bp	0.48 %

* most repeats fragmented by insertions or deletions have been counted as one element

file name: human chromosome 21 TDs
sequences: 2005
total length: 1538810 bp (1538810 bp excl N/X-runs)
GC level: 29.37 %
bases masked: 718706 bp (46.71 %)

	number of elements*	length occupied	percentage of sequence
SINEs:	241	24976 bp	1.62 %
ALUs	172	16087 bp	1.05 %
MIRs	69	8789 bp	0.57 %
LINEs:	1117	491307 bp	31.93 %
LINE1	1083	486242 bp	31.60 %
LINE2	32	4772 bp	0.31 %
L3/CR1	0	0 bp	0.00 %
LTR elements:	177	25369 bp	1.65 %
MaLRs	88	8247 bp	0.54 %
ERVL	41	4424 bp	0.29 %
ERV_classI	45	12490 bp	0.81 %
ERV_classII	2	141 bp	0.01 %
DNA elements:	122	30091 bp	1.96 %
MER1_type	37	7071 bp	0.46 %
MER2_type	49	12694 bp	0.82 %
Unclassified:	1	757 bp	0.05 %
Total interspersed repeats:		572400 bp	37.20 %
Small RNA:	4	220 bp	0.01 %
Satellites:	71	57703 bp	3.75 %
Simple repeats:	552	58077 bp	3.77 %
Low complexity:	717	30834 bp	2.00 %

* most repeats fragmented by insertions or deletions have been counted as one element

Table 10: Repeat analysis human chromosome 21 TDs and whole sequence

The repeat content of the TDs in human chromosome 21 is found to be 46.71%. The overall repeat content of the whole chromosome was determined at 33.82%. However if one excludes un-sequenced nucleotides (N/X runs), which do not appear in TD regions, from consideration, the repeat content of the entire sequence of human chromosome 21 can be interpreted to be as high has 46.47%. In the same way the repeat content of human chromosomes 4 and 22 was determined to be 49.64% and 48.5% respectively. The repeat content of the transcriptional deserts on these chromosomes was determined to be 45.5% and 69.02% respectively. Since human chromosomes 4, 21, and 22 can be regarded as showcases for chromosomes of low, average, and high gene density and GC-content, this can be interpreted as meaning that the overall repeat content is very similar between the TDs on a chromosome and the entire chromosomal sequence.

Only for chromosome 22, which is of very high density in terms of transcriptional activity, is the repeat content of TDs significantly higher.

While it is observed that the overall content of repeats is not considerably different when comparing the TDs and the entire chromosomal sequence, the composition of repeat sequences displays some characteristic differences. It can be seen that the type of repetitive DNA that make up repeats in TD regions is distinctly unlike the overall repeat composition. This can be seen above in Table 10. The role of Long Interspersed Nuclear Elements (LINE) has gained particular importance in transcriptional deserts in comparison to all other types of repetitive DNA. In fact, LINEs do make up almost 80% of all repetitive DNA in TDs, while overall their portion does not exceed 40%. Almost all interspersed repeats in transcriptional deserts are LINEs. Short Interspersed Nuclear Elements (SINEs) and Long Terminal Repeats (LTRs) have disappeared almost entirely. DNA elements have also reduced their proportion, but not as considerably as SINEs and LTRs. This applies equally to human and mouse. Satellite sequences, simple repeats and regions of low complexity have gained importance in TDs in relation to their occurrence in the entire chromosome, but still play only a minor role in the composition of repeats in transcriptional deserts. They do, however, occur more often in mouse TD sequences than in human TD sequences.

Single nucleotide polymorphisms

A single nucleotide polymorphism (SNP) is a type of DNA sequence variation. It consists of a discrepancy in the DNA sequence between the individual members of one species, in which only a single nucleotide is different from one individual to the other. SNPs have previously been implicated in disrupting the process of gene regulation and the development of genetic disorders. All known SNPs in the human and mouse genome were collected and examined to see how many of them fall within the

transcriptional desert regions. Tables 11 and 12 below show an analysis of single genomic SNPs in *Homo sapiens* and *Mus musculus* with regard to their location within or outside of TD regions.

chrom	total SNPs	SNPs in deserts	% of SNPs in deserts	SNPs out of deserts	% of SNPs out of deserts	total length of deserts	tot len of non deserts	% of chr covered by desert	chr size	SNPs/Knuc (desert)	SNPs/Knuc (non-desert)	SNPs/Knuc (all)
chr1	1009418	19855	1.9670	989563	98.0330	4776186	242473533	1.9317	247249719	4.1571	4.0811	4.0826
chr2	921727	25573	2.7745	896154	97.2255	6852882	236098267	2.8207	242951149	3.7317	3.7957	3.7939
chr3	729150	18598	2.5506	710552	97.4494	5177077	194324750	2.5950	199501827	3.5924	3.6565	3.6549
chr4	768943	38268	4.9767	730675	95.0233	9538252	181734811	4.9867	191273063	4.0121	4.0206	4.0201
chr5	685532	23341	3.4048	662191	96.5952	6270066	174587800	3.4668	180857866	3.7226	3.7929	3.7904
chr6	730709	23304	3.1892	707405	96.8108	5483736	165416256	3.2087	170899992	4.2497	4.2765	4.2757
chr7	656470	15329	2.3351	641141	97.6649	3775130	155046294	2.3770	158821424	4.0605	4.1352	4.1334
chr8	585312	14075	2.4047	571237	97.5953	3703951	142570875	2.5322	146274826	3.8000	4.0067	4.0015
chr9	707810	15352	2.1689	692458	97.8311	2703174	137570078	1.9271	140273252	5.6792	5.0335	5.0459
chr10	600193	14460	2.4092	585733	97.5908	2699576	132675161	1.9942	135374737	5.3564	4.4148	4.4336
chr11	572341	15785	2.7580	556556	97.2420	3367264	131085120	2.5044	134452384	4.6878	4.2458	4.2568
chr12	529824	10779	2.0344	519045	97.9656	2592040	129757494	1.9585	132349534	4.1585	4.0001	4.0032
chr13	402675	23025	5.7180	379650	94.2820	5536875	108606104	4.8508	114142980	4.1585	3.4957	3.5278
chr14	355283	8785	2.4727	346498	97.5273	2244172	104126413	2.1098	106368585	3.9146	3.3277	3.3401
chr15	357766	3603	1.0071	354163	98.9929	858415	99480500	0.8555	100338915	4.1973	3.5601	3.5656
chr16	411450	4159	1.0108	407291	98.9892	889421	87937833	1.0013	88827254	4.6761	4.6316	4.6320
chr17	320919	2338	0.7285	318581	99.2715	603956	78170786	0.7667	78774742	3.8711	4.0754	4.0739
chr18	312005	12098	3.8775	299907	96.1225	2751064	73366089	3.6142	76117153	4.3976	4.0878	4.0990
chr19	256293	1202	0.4690	255091	99.5310	216084	63595567	0.3386	63811651	5.5627	4.0111	4.0164
chr20	306801	3712	1.2099	303089	98.7901	618171	61817793	0.9901	62435964	6.0048	4.9029	4.9139
chr21	172588	9230	5.3480	163358	94.6520	1538810	45405513	3.2779	46944323	5.9981	3.5978	3.6764
chr22	218189	2185	1.0014	216004	98.9986	243684	49447748	0.4904	49691432	8.9665	4.3683	4.3909
chrX	421470	25996	6.1679	395474	93.8321	7955221	146958533	5.1353	154913754	3.2678	2.6911	2.7207
chrY	92955	10306	11.0871	82649	88.9129	2291630	55481324	3.9666	57772954	4.4972	1.4897	1.6090
whole	12125823	341358	2.8151	11784465	97.1849	82686838	2997732642	2.6843	3080419480	4.1283	3.9311	3.9364

Table 11: Human SNPs and TD analysis

chrom	total SNPs	SNPs in deserts	% of SNPs in deserts	SNPs out of deserts	% of SNPs out of deserts	total length of deserts	tot len of non deserts	% of chr covered by	chr size	SNPs/Knuc (desert)	SNPs/Knuc (non-desert)	SNPs/Knuc (all)
chr1	746295	15274	2.0466	731021	97.9534	5388830	191681132	2.7345	197069962	2.8344	3.8137	3.7870
chr2	687895	7729	1.1236	680166	98.8764	3663748	178313014	2.0133	181976762	2.1096	3.8144	3.7801
chr3	575751	13576	2.3580	562175	97.6420	4841853	155030259	3.0286	159872112	2.8039	3.6262	3.6013
chr4	625325	9564	1.5294	615761	98.4706	3773523	151256178	2.4341	155029701	2.5345	4.0710	4.0336
chr5	495364	7403	1.4945	487961	98.5055	2941789	149061274	1.9353	152003063	2.5165	3.2736	3.2589
chr6	552038	8458	1.5321	543580	98.4679	3346059	146179626	2.2378	149525685	2.5277	3.7186	3.6919
chr7	507856	3779	0.7441	504077	99.2559	1739880	143394214	1.1988	145134094	2.1720	3.5153	3.4992
chr8	415589	7377	1.7751	408212	98.2249	2849009	129236089	2.1569	132085098	2.5893	3.1587	3.1464
chr9	344962	3981	1.1540	340981	98.8460	2118934	121881735	1.7088	124000669	1.8788	2.7976	2.7819
chr10	254930	3940	1.5455	250990	98.4545	3065051	126894097	2.3585	129959148	1.2855	1.9779	1.9616
chr11	380675	2987	0.7847	377688	99.2153	1548624	120250008	1.2715	121798632	1.9288	3.1409	3.1254
chr12	268072	4346	1.6212	263726	98.3788	2751254	117711905	2.2839	120463159	1.5796	2.2404	2.2253
chr13	264026	2646	1.0022	261380	98.9978	2036070	118578308	1.6881	120614378	1.2996	2.2043	2.1890
chr14	306878	9342	3.0442	297536	96.9558	3942343	120036527	3.1799	123978870	2.3697	2.4787	2.4752
chr15	309001	6683	2.1628	302318	97.8372	2787665	100704912	2.6936	103492577	2.3973	3.0020	2.9857
chr16	181623	3549	1.9540	178074	98.0460	2522850	95729609	2.5677	98252459	1.4067	1.8602	1.8485
chr17	158826	1811	1.1402	157015	98.8598	1545791	93631629	1.6241	95177420	1.1716	1.6769	1.6687
chr18	289405	4159	1.4371	285246	98.5629	1965569	88771268	2.1662	90736837	2.1159	3.2133	3.1895
chr19	224635	1560	0.6945	223075	99.3055	700953	60620237	1.1431	61321190	2.2255	3.6799	3.6633
chrX	308519	10727	3.4769	297792	96.5231	6008969	159547500	3.6296	165556469	1.7852	1.8665	1.8635
chrY	322	14	4.3478	308	95.6522	74811	15954593	0.4667	16029404	0.1871	0.0193	0.0201
whole	7897987	128905	1.6321	7769082	98.3679	59613575	2584464114	2.2546	2644077689	2.1623	3.0061	2.9870

Table 12: Mouse SNPs and TD analysis

As can be seen in Tables 11 and 12, there are only minor variations in the number of SNPs per kilonucleotide within and outside of TD regions on all chromosomes of mouse and human. It is also true that the percentage of TD coverage for each chromosome is only slightly different from the percentage of SNPs that are located in the TD regions. It can therefore be concluded that a statistically significant difference between SNPs occurring within and outside deserts cannot be established. The data presented in Tables 11 and 12 show that the hypothesis that there is no correlation between occurrence of a SNP and presence of a transcriptional desert at the same location cannot be rejected. It has to be concluded that SNPs are equally distributed over TD and non-TD regions of mammalian DNA. The above analysis is restricted to single genomic SNPs. The same analysis was repeated for insertion and deletion events with very similar results. The conclusion is that neither single genomic SNPs nor insertion-deletion events occur in correlation with transcriptional desert regions.

SNPs can be grouped into two classes, depending on the kind of substitution that is observed between the bases A, C, G and T. Given their chemical structure, A and G are characterised as purines, while C and T are pyrimidine molecules. Substitutions that do not change the chemical structure of a nucleotide, that is, purine to purine or pyrimidine to pyrimidine, are called transitions. Substitutions that change a purine into a pyrimidine or vice versa are called transversions. Overall, roughly two thirds of all SNPs are transitions. Since transversions change the chemical structure of the molecules involved their effects are normally more severe than those of transitions.

The rates in which nucleotides change in the human and mouse genome were determined and these changes were displayed as substitution rate matrices. Table 13 shows these rate matrices.

H.Sapiens				
SNPs rate per kNuc whole genome				
	A	C	G	T
A	0	0.15	**0.47**	0.15
C	0.15	0	0.17	**0.6**
G	**0.6**	0.17	0	0.15
T	0.15	**0.46**	0.15	0

M.Musculus				
SNPs rate per kNuc whole genome				
	A	C	G	T
A	0	0.12	**0.47**	0.12
C	0.13	0	0.1	**0.52**
G	**0.53**	0.1	0	0.13
T	0.12	**0.47**	0.12	0

SNPs rate per kNuc deserts				
	A	C	G	T
A	0	0.2	**0.56**	0.23
C	0.16	0	0.14	**0.53**
G	**0.54**	0.14	0	0.16
T	0.23	**0.57**	0.2	0

SNPs rate per kNuc deserts				
	A	C	G	T
A	0	0.11	**0.33**	0.14
C	0.1	0	0.07	**0.31**
G	**0.31**	0.06	0	0.1
T	0.14	**0.34**	0.11	0

bold = transition
regular = transversion
lines -> columns

Table 13: SNP rate matrices in human and mouse TDs and whole genome

Transitions are shown in bold and transversions in regular font. The matrices are to be interpreted in such a way that the nucleotides in lines turn into the nucleotides in the respective column, with the rate given in substitutions per kilonucleotide. As with the examination of all SNPs presented above, the results are inconclusive. It cannot clearly be said that transition or transversions appear predominantly in TD regions. Nor can any single nucleotide be singled out that is predominantly substituted by another nucleotide within or outside of transcriptional deserts. Overall, the SNP analysis in transcriptional deserts implies that no strong relationship between any kind of SNP and the existence of a TD region can be established.

Transcription factor binding sites

TRANSFAC 11.4 [76] is a knowledge base and software system for the purpose of analyses related to transcription factor (TF) binding sites in DNA

sequences. The modules of TRANSFAC 11.4 that were used in the analysis of transcriptional deserts are a collection of matrices that describe possible binding sites for transcription factors (TFBS). Each matrix was constructed from a number of experimentally proven binding sites for transcription factors. Each binding site consists of a core section and sections flanking it, possibly from both sides. Similar binding sites have been grouped together to form one matrix. There are in total 834 distinct matrices. The conformance between the individual binding sites that were used to construct a matrix is very high in the core section of the matrix and less high in the marginal section.

MATCH [75] is a program that scans DNA sequences and identifies possible binding sites for transcription factors. For each possible TFBS, a core and a marginal match score is reported. MATCH was used to analyse transcriptional desert regions for possible TFBSs. It was decided to restrict the search to those matrices that were constructed from binding sites that have been proven to exist in vertebrates. Furthermore, a predefined configuration for the exclusion of binding sites with low match scores was chosen, so as to minimise the occurrence of false-positive binding site predictions. This included restriction to the use of only high quality matrices. There are 196 high quality matrices which represent binding sites for 1251 different TFs. In order to compare the incidence of TFBS in TD regions, the same analysis was conducted on the complete sequence of human chromosome 21, the selection of human cDNA as shown in Table 8, and a sequence of randomly generated DNA junk.

Table A2 in the appendix shows the occurrence of binding sites for individual matrices in these four types of data in matches per kilonucleotide. It also shows the n-fold enrichment in the occurrences of each matrix between TD regions and one of the data sets used for comparison. In total, there are on average 27.39 potential binding sites for any of the total of 196 binding site matrices present per kilonucleotide in transcriptional desert regions. For cDNA sequences, there are 21.44 such binding sites and for the

entire sequence of human chromosome 21 there are 24.87. Interestingly, there are also 25.47 binding sites per kilonucleotide in randomly generated junk DNA. There are, however, 13 matrices for which no binding site can be found in randomly generated junk DNA, while there are only two that cannot bind in cDNA sequences. TD regions and human chromosome 21 provide binding sites for all matrices with only one exception. There are five matrices (for FOXP1, CART1, POU1F1, HNF6 and POU6F1) which occur more than ten times more often in TD regions than in cDNA. Four out of these five also occur more than twice as often in TDs than in human chromosome 21 and more than six times more often in TDs than in randomly created junk DNA. FOXP1 does not occur in randomly generated junk DNA and occurs 1.6 times more often in TD regions than in human chromosome 21. Figure 7 show the occurrence of matrix binding sites in the various data sets used. The order of TF matrices has been sorted according to occurrence in TD regions (blue).

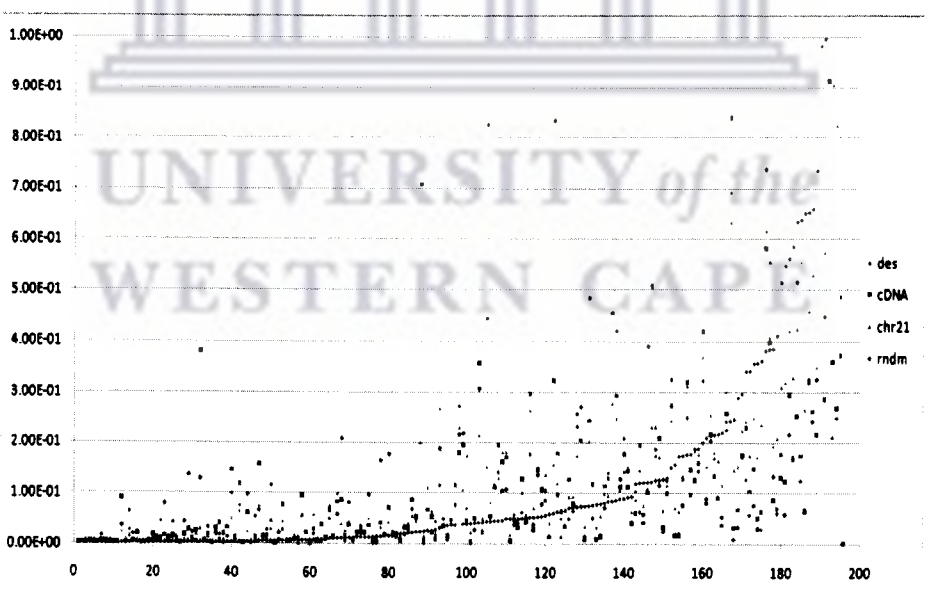

Figure 7: Occurrence of binding sites for TF matrices in 4 types of DNA data

The frequency with which individual transcription factors can bind to the four types of DNA sequences used in this analysis was also examined. This is distinctly different from investigating the binding frequencies of TF matrices. One TF can have several binding domains, that is, several parts of the TF can bind to different DNA sequences. The individual binding domains of a single TF can be very different from each other, which means that the same TF can bind to largely different DNA sequences, using one or the other of its binding domains. Different binding domains were used in the creation of different binding matrices. Therefore investigating which TFs bind to a DNA sequence and which TF matrices bind to a DNA sequence is not the same and might yield different insights from the previous analysis.

Table A3 in the appendix shows the frequency with which individual transcription factors can bind in the four types of DNA sequences used in this analysis. The 1251 TFs under examination can together bind in TDs, cDNA, human chromosome 21 and randomly generated junk DNA 182.57, 166.97, 198.08 and 166.27 times per kilonucleotide respectively. The individual transcription factors bind on average 0.15, 0.13, 0.16 and 0.13 times respectively per kilonucleotide. It is noticeable that these values are very similar for cDNA and randomly generated junk DNA. On the other hand binding frequencies for TFs are also similar between TD regions and human chromosome 21. There are 48 TFs for which no binding site can be located in randomly generated junk DNA, while there are only four that cannot bind to the cDNA sequences and two that cannot bind to human chromosome 21. In transcriptional desert regions TFBSs for all transcription factors can be identified.

Transcriptional deserts are rich in the occurrence of TFBS when compared with cDNA and randomly generated junk DNA. On average, binding sites for an individual transcription factor are found 1.48 times more often in TDs than in cDNA and 1.91 times more often in TDs than in randomly generated junk DNA. TFBS appear to be denser on human chromosome 21 than in TDs. A binding site for an individual TF factor is found on average 0.75

times less often in TDs when compared to human chromosome 21. However, this number only takes into consideration the portion of human chromosome 21 that is not covered by unsequenced nucleotides ('Ns'). 'Ns' make up 27% of the chromosomal sequence of human chromosome 21.

There are 18 TFBSs which occur in TDs more than 10 times as often than in cDNA, while at the same time there are 94 TFBSs for which less than one tenth of possible binding sites in TDs than in cDNA are found. For human chromosome 21, there are 28 TFBSs which occur more than twice as much in TD regions than in this chromosome. There are also 60 TFBSs which bind in TDs with less than one tenth of the frequency they bind within human chromosome21.

This serves to show that there are distinct differences between the landscape of TFBSs in TDs and other types of DNA sequences. This implies that it can be speculated that, although TDs are transcriptionally inactive, some TD regions might play a distinct role in the regulatory process of genes.

Clusters of TFBSs in transcriptional desert regions

In order to examine further the possibility that transcriptional desert regions play a role in remote regulatory processes, it was subsequently examined to what extent TFBSs appear in TDs in a clustered fashion. For this purpose a sliding window of length 200 nt was assumed. This window was moved along the TDs in the mouse and human genome and it was determined how many TFBS are located in each window. The window was slid along by one nucleotide at a time. The mean (m) and the standard deviation (σ) of TFBS occurrences across all windows were determined for each chromosome separately in *Homo sapiens* and *Mus musculus*. The results of this analysis are shown in Tables 14 and 15.

#chr	mean (m)	stddiv (σ)	> m + 2*σ	> m + 3*σ	> m + 4*σ	> m + 6*σ	> m + 8*σ	> m + 10*σ
chr1	5.73	2.94	136543	27586	7681	1094	114	0
chr2	5.65	3.00	179337	40035	14469	3675	1283	558
chr3	5.71	2.95	136991	29835	8697	2624	1394	703
chr4	5.76	2.92	263270	51239	12240	1717	409	67
chr5	5.72	2.94	169229	35537	10711	2326	368	154
chr6	5.67	2.93	143834	30764	9190	1762	359	25
chr7	5.67	2.96	103457	21249	5534	770	167	0
chr8	5.67	3.04	104045	26270	10393	2634	270	165
chr9	5.67	2.95	71657	16469	5226	1036	28	0
chr10	5.51	3.13	74818	19030	5496	1847	668	134
chr11	5.64	2.98	90707	18837	5985	1907	709	105
chr12	5.69	3.09	77467	20316	5881	2298	717	8
chr13	5.84	2.93	155907	30891	8608	2106	276	0
chr14	5.69	2.93	59044	11570	3408	340	19	0
chr15	5.46	2.87	18576	3778	1336	158	0	0
chr16	5.75	3.57	23772	10710	4533	988	0	0
chr17	5.76	3.33	13106	4967	2517	733	143	0
chr18	5.73	3.05	88260	22572	5550	785	55	0
chr19	5.53	3.50	6739	1584	674	80	0	0
chr20	5.46	3.06	17720	3491	872	179	0	0
chr21	5.51	3.05	41408	9218	2531	267	48	0
chr22	5.04	3.90	8452	1761	909	287	118	0
chrX	5.53	2.99	209223	54066	17579	3745	1170	364
chrY	5.30	3.35	49932	16223	5882	1882	771	333
HS whole genome	5.61	3.10	2243494	507998	155902	35240	9086	2616
outliers/knt			27.132	6.144	1.885	0.426	0.110	0.032

Table 14: TFBS cluster in human TDs

#chr	mean (m)	stddiv (σ)	> m + 2*σ	> m + 3*σ	> m + 4*σ	> m + 6*σ	> m + 8*σ	> m + 10*σ
chr1	5.34	3.15	148877	42600	14090	2131	334	82
chr2	5.22	3.19	96558	26906	9958	1459	465	176
chr3	5.29	3.10	123175	32197	9520	1492	373	157
chr4	5.29	3.20	108841	33319	7811	1021	126	0
chr5	5.30	3.42	64041	26036	12516	2395	289	0
chr6	5.17	3.11	81217	22015	7779	1557	217	42
chr7	5.03	3.24	42580	14825	6584	1553	421	75
chr8	5.17	3.22	77126	24178	6379	897	87	0
chr9	5.18	3.17	53323	16151	5691	178	0	0
chr10	5.24	3.20	86728	23763	6478	1114	144	0
chr11	5.18	3.22	41868	11402	3470	937	160	0
chr12	5.37	3.21	82324	19891	6728	1030	0	0
chr13	5.16	3.41	54502	13229	5788	2908	1379	403
chr14	5.56	3.19	128006	22249	6485	1201	272	5
chr15	5.40	3.19	80755	23198	4641	483	89	0
chr16	5.40	3.14	71181	19199	5845	396	42	0
chr17	5.21	3.18	41332	12864	4547	487	142	0
chr18	5.24	3.22	56302	16908	3296	605	187	9
chr19	5.03	3.36	19025	4630	2316	827	72	0
chrX	5.40	2.95	146608	35882	9620	1186	242	76
chrY	5.43	3.68	2204	1190	463	87	0	0
MM whole genome	5.27	3.23	1606573	442632	140005	23944	5041	1025
outliers/knt			26.950	7.425	2.349	0.402	0.085	0.017

Table 15: TFBS cluster in mouse TDs

The results show that on average there are between five and six TFBS per 200 nt predicted by MATCH in the TD regions of human and mouse. The standard deviation from this mean value is on average three TFBS per window. These values are consistent throughout the genomes of human and mouse. Also shown in Tables 14 and 15 are values quantifying the number of significant outliers from these mean values. The numbers in columns 4 to 9 of Tables 14 and 15 give the number of 200 nt windows in TD regions that harbour more than the mean plus a multiple of the standard deviation. For the genome of *Homo sapiens* there are in total 2616 windows of size 200 nt that are located in TD regions and that harbour more than 35 TFBSs ($m + 10 * \sigma = \sim 35$). For the whole genome of *Mus musculus* there are 1025 such enriched windows. The bottom row of Tables 14 and 15 shows how many significant outliers there are per kilonucleotide in TD regions.

For comparison 10,000 promoter regions have been analysed with MATCH. These promoter regions were obtained by randomly selecting 10,000 TSSs from the set of 113,814 TSSs that were described in the 'Methods' section of Chapter 1. Sequences covering the interval [-3000, 200] around these TSSs were extracted and used in this comparison analysis. These promoter regions can be regarded as the main control regions for transcriptional activity. Investigating the clustering of TFBSs in them and then comparing the results of this investigation to the results obtained regarding the clustering of TFBSs in TD regions, gives in indication as to how TD regions might contribute to gene regulation as silencers or enhancers. The results of this analysis are shown in Table 16.

	mean (m)	stddiv (σ)	> m + 2*σ	> m + 3*σ	> m + 4*σ	> m + 6*σ	> m + 8*σ	> m + 10*σ
10000 promoters	5.27	3.04	1136852	312114	85256	7273	1707	364
outliers/knt			37.895	10.404	2.842	0.242	0.057	0.012

Table 16: TFBS cluster in 10,000 randomly selected promoter regions

As seen in Table 16, in 10,000 randomly selected promoters the mean (m) frequency of TFBSs in a 200 nt sliding window is 5.27 and the standard deviation (σ) from this mean is 3.04. These values are very similar to the values obtained in the respective analysis in TD regions. It is also observed that the number of outliers per kilonucleotide is lower in TD regions for weak outliers (> m + 2 * σ, > m + 3* σ and > m + 4* σ) and higher in TD regions for strong outliers (> m + 6* σ, > m + 8* σ and > m + 10* σ). This allows the conclusion that a number of transcriptional desert regions might in fact be active as transcriptional remote control elements because some TD regions appear to be harbouring more TFBSs than evidently transcriptionally active promoter regions.

A p-value was calculated for the occurrence of more than m + 10* σ TFBSs in TDs in relation to the occurrence in promoters. This p-value is very small (~5.5 * 10^{-130}) which means that the enrichment in TFBS in TDs can be regarded as stistically significant.

Mutations and TFBS occurrence

The results regarding single nucleotide polymorphisms (SNPs) and transcription factor binding sites (TFBS) have been combined for this part of the study to determine transcriptional desert regions (TDs) in which SNPs and TFBSs are strongly clustered. All SNPs that reside within a TD region and fall within a TFBS as well have been reported. A distinction was made between SNPs that fall within the marginal or peripheral region of the binding motif and those SNPs that fall within the core region of the motif. The preservation of the motif is many-fold stronger in the core region of the binding site than in the marginal areas. Therefore, a SNP occurring in the core region has a much stronger impact on the ability of the site to bind a certain transcription factor. In fact it can be assumed that a mutation in the core region of a TFBS is likely to hinder the further binding of a transcription factor. At the same time, a mutation in the peripheral areas of

the TFBS will have a much weaker influence on the ability of the motif in question to bind its transcription factor.

The genome of *Homo sapiens* contains 106,554 TD regions with a total length of 82,686,838 nucleotides (see Table 9a). 90,944 of those regions harbour at least one SNP. In 37,467 TD regions, the location of a SNP coincides with the location of a TFBS. 28,473 TDs harbour at least one SNP that falls in the peripheral region of a TFBS and 17,270 SNPs in TDs fall within the core binding motif of a TFBS. In total, there are 66,465 TFBSs in TD regions that co-occur with a SNP (22,764 core and 43,701 peripheral). This makes a density of TFBS-SNP co-occurrence in TDs of 0.804 per kilonucleotide (0.275 core and 0.539 peripheral).

The genome of *Mus musculus* contains 84,936 TD regions with a total length of 59,613,575 nucleotides (see Table 9a). 41,828 of those regions harbour at least one SNP. In 15,694 of these the location of a SNP coincides with the location of a TFBS. 11,779 TDs harbour at least one SNP that falls in the peripheral region of a TFBS and 7,033 SNPs in TDs fall within the core binding motif of a TFBS. In total, there are 25,293 TFBSs in TD regions that co-occur with a SNP (8731 core and 16,561 peripheral). This makes a density of TFBS-SNP co-occurrence in TDs of 0.424 per kilonucleotide (0.146 core and 0.278 peripheral).

These results have to be seen in the context of the number of SNPs that are available for *Mus musculus* in comparison with the number of SNPs available for *Homo sapiens* (7,897,987 vs. 12,125,823) as well as the percentage of these SNPs that fall within transcriptional desert regions (1.6% vs. 2.8% see Tables 11 and 12). Nevertheless clusters of SNPs and TFBS are more frequent in the human genome compared to the mouse genome. The complete set of results showing in which TD region which TF binding matrix is subjected to a mutation is shown in Tables 3 and 4 in the online supporting materials to this manuscript.

For comparison, all 1129 promoter regions on human chromosome 21 have been analysed. For this purpose, all TSSs that were described in the 'Methods' section of Chapter 1 and reside on human chromosome 21 where identified and the sequences covering [-3000, 200] relative to these 1129 TSSs where extracted. Subsequently it was investigated to what extent SNPs and TFBSs coincide within these promoter regions. Since promoter regions are the main control regions for gene expression, this analysis can show if the co-occurrence of TFBSs and SNPs in TD regions is to be regarded as notable.

1129 promoter regions on human chromosome 21 with a total length of 3,612,800 nt were investigated. In 883 promoter regions, a SNP co-occurring with a TFBS could be identified. In 728 promoter regions, a SNP fell within the peripheral region of the binding motif and in 553 in the core region. Furthermore, there are 2874 TFBSs within promoter regions on human chromosome 21 that co-occur with a SNP (882 core and 1992 peripheral). This makes a density of TFBS-SNP co-occurrences in TDs of 0.796 per kilonucleotide (0.244 core and 0.551 peripheral).

The results of the analysis of co-occurrences of TFBSs and SNPs in transcriptional desert regions and promoter regions is summarised in Table 17.

	nOf regions	tot. length	nOf core co-occurence	nOf peripheral co-occurence	nOf co-occurance	core co-occurence / knt	peripheral co-occurence / knt	total co-occurence / knt
HS TDs	106554	82686838	22764	43701	66465	0.275	0.539	0.804
MM TDs	84936	59613575	8731	16561	25293	0.146	0.278	0.424
HS chr21 promoters	1129	3612800	882	1992	2874	0.244	0.551	0.796

Table 17: TFBS and SNP co-occurrence

It can be seen that the co-occurrence per kilonucleotide is very similar between human TD regions and human promoters. It can be surmised that the reasons for this are twofold. On the one hand, it was established in a previous subsection that there is no significant difference between the frequency of SNPs within and outside of TD regions. On the other hand, it

was also previously established that the overall frequency of TFBSs is similar between TDs and promoter regions. The similar frequency of co-occurrences is therefore within the bounds of reasonable expectations.

Evolutionary conservation

The evolutionary conservation has been examined using the alignment of hg18 with mm8. It was determined what portion of human TD regions fall into areas that are conserved between the genomes of humans and mouse. Only matched sequences that have a minimum BLASTZ similarity score were taken into consideration. Three minimum scores were considered. The results are shown per chromosome in Table 18. For comparison the overall portions of the respective chromosomes that are conserved between mouse and human are shown as well.

	similarity threshold					
	TDs	hg18	TDs	hg18	TDs	hg18
#chr	0	0	5000	5000	10000	10000
chr1	20.21	37.02	18.85	33.64	15.84	28.39
chr2	18.97	39.68	17.60	36.05	14.81	30.29
chr3	20.89	40.31	19.40	36.83	16.60	31.04
chr4	21.70	36.21	20.10	32.79	17.11	27.13
chr5	21.71	39.86	20.13	36.43	17.13	30.83
chr6	21.47	38.27	19.91	34.63	16.84	28.64
chr7	21.19	37.51	19.54	34.08	16.29	28.57
chr8	19.21	37.17	18.02	33.61	14.91	27.72
chr9	19.33	34.07	17.60	30.93	14.90	26.01
chr10	17.25	39.71	15.68	35.71	12.77	29.44
chr11	15.50	39.90	13.82	36.58	11.33	31.28
chr12	20.56	36.93	18.86	33.14	15.85	27.32
chr13	22.60	30.32	20.93	27.19	17.41	22.30
chr14	20.49	33.87	19.13	30.79	16.25	25.98
chr15	18.29	34.61	17.11	31.52	15.33	26.55
chr16	21.99	35.08	20.39	31.04	17.09	25.39
chr17	20.49	41.50	18.94	37.05	15.44	31.00
chr18	19.51	38.18	17.96	34.46	14.93	28.41
chr19	14.07	21.44	11.43	18.54	9.48	15.19
chr20	19.88	39.18	18.76	35.50	15.60	29.63
chr21	18.25	24.23	16.96	21.31	13.69	17.00
chr22	12.06	24.53	11.28	21.43	8.13	17.25
chrX	8.96	29.67	8.43	27.40	7.30	23.50
chrY	5.86	4.73	5.17	4.05	3.67	2.97
avg	18.35	33.92	16.92	30.61	14.11	25.49

Table 18: Portion of evolutionary conserved human TDs

It can be seen that just fewer than 20% of all TDs are evolutionary conserved between humans and mice when only a weak similarity score is required. When stricter requirements for sequence similarity are demanded, this portion is reduced to about 14% when averaged over 24 human chromosomes. Based on these numbers it can be estimated that between one in five and one in seven TDs regions have been conserved in the genomes of humans and mice. Compared to the overall conservation between human and mouse, TDs are roughly half as often conserved.

DISCUSSION

In this chapter a methodology for the identification of transcriptional desert regions was presented. This methodology unifies the identification of transcription initiation deserts through DDM, which was presented in Chapter 1, and a comprehensive collection of known transcripts. The regions thus identified are devoid of transcription start sites and are not part of any known transcript. In total, 82.6 million nucleotides for *Homo sapiens* fall into these transcriptional deserts. This corresponds to 2.7% of the human genome. For *Mus musculus*, 59.6 million nucleotides or 2.3% of the genome could be identified. These values increase when the minimum length requirement for TDs is reduced. The longest consecutive stretch of DNA that is not capable of initiating transcription and for which a transcript is not known to exist is 25 kilonucleotides long.

It is generally accepted knowledge that almost half of human and other mammalian genomes consist of repetitive DNA which is not assigned major functionality. It is also generally accepted knowledge that only a minor fraction of the human and other mammalian genomes consist of protein-coding exons. Together, this creates the idea that large areas of mammalian genomes do not have any function. Such thinking is only slowly overcome in today's research approaches.

The existence of a transcript and with it the initiation of transcription implies that the DNA from which the transcript was produced is functional. If a DNA sequence is turned into an RNA sequence that is present in the cell at some point in time some function or other can be assumed to exist. Even DNA that is not transcribed cannot *a priori* be regarded as non-functional.

It is nevertheless surprising to establish that only between 2 and 3% of the genomes of mouse and human constitute transcriptional desert regions. If the notion were correct that only a fraction of mammalian DNA were functional, the proportion that makes up transcriptional deserts would be considerably higher. The analysis and results presented here do in this

context suggest that the view that the human and other mammalian genomes possess large stretches of DNA that are devoid of any function needs to be re-evaluated. There are previous studies that have already suggested that almost all of the human genome is transcribed [77,78]. The results and analysis presented here confirm this point of view and quantify the propositions made.

The GC-content is usually used as a convenient way to interpret and explain the gene-richness or depletion of genes of a DNA sequence [79]. A low GC-content or AT-richness of a region is normally associated with a low rate of expressional activity, while a high GC-content or AT-depletion is associated with a high rate of activity. The observation made regarding GC-content in TD regions does therefore match the expectation that one would derive from the common knowledge between GC-richness and transcription. It was observed that the GC-content in transcriptional deserts (on average 31%) is roughly 10% lower than the overall GC-content of mammalian genomes. However, simply employing the GC-content to detect or explain the presence or absence of transcriptional activity is insufficient. The analysis and results presented here indicate that while there is certainly a strong correlation between the GC-content and the presence of transcription, a high GC-content is neither a necessary nor a sufficient condition for a DNA sequence to be transcribed in a mammalian genome. It is not necessary because there are numerous AT-rich DNA sequences for which the existence of a transcript can be shown and it is not sufficient because a number of GC-rich sequences were identified as TDs. This means that other factors must exist that work in conjunction with GC-content that determine whether a sequence of DNA is transcribed of not. While the GC-content gives a clue as to the possible existence of a transcript it is not the sole determining factor.

It was seen that the average GC-content of TD regions in the mouse genome is 2% higher than in the human genome. It was also seen that TD regions are slightly rarer in mouse than in human, and that TD regions with a high

GC-content play a slightly larger role in mice than in humans. All this suggests that there are subtle differences in the interrelationships between GC-content and transcriptional activity between mice and humans. The data presented here can be interpreted in such a way that the GC-content has, to some extent, a larger impact on a sequence ability to transcribe in human than it would have in mice.

The composition of k-mers observed in transcriptional deserts in comparison with randomly extracted DNA of similar GC-content reveals that the k-mer composition of TD regions can to a large extent be explained by the low GC-content of these region. When the k-mer composition of TDs is compared to random DNA of similar GC-content, the differences observed are relatively modest. It is not too far-fetched to speculate that a comparison with random DNA of a GC-content that is 'average' for the respective genome would yield the observation that TDs are significantly enriched in many AT-rich k-mers. This observation, however, could be explained solely with the different GC-content of the compared sequences and would not be caused by other compositional characteristics that are typical for TD regions. This can be concluded from the analysis performed here, in which TDs were compared with random DNA of very similar GC-content. The fact that this analysis does not yield any prominently enriched or depleted k-mers in TD regions shows that relative AT-richness is sufficient to explain the k-mer composition of transcriptional desert regions.

When studies of transcription and gene expression are conducted repetitive DNA is often *a priori* excluded. It is assumed that repeats do not play a significant role in these processes. Several studies (for example [42]) have restricted their analysis to genomic sequences in which repeats were masked. Such an approach might be correct in many cases. Genes containing protein-coding exons are certainly very unlikely to be located within a repeated DNA sequence. However, suggesting that repetitive DNA is deprived of function might be inaccurate, as the results and analysis presented here imply. The repeat content of transcriptional deserts was

determined to be of a similar magnitude than the repeat content of the overall genome for human and mouse. In both cases the repeat content can be determined to be roughly around 50%. On the one hand, this entails that half of the regions that are neither capable of initiating transcription nor part of a known transcript are composed of non-repetitive DNA. Repeats can therefore not serve as an explanation for the transcriptional passivity of the examined TDs. On the other hand, this implies that the relationship between repetitive DNA and transcriptional DNA is weak at best. If it were true that repetitive DNA is transcriptionally less active than the genome on average, the repeat content of a TD region would be much higher than that of the overall chromosome. As a matter of fact, most repeat regions are transcribed in some way and, as was shown in Chapter 1, might even be the location of transcription initiation. The results and analysis shown in this suggest that the role of repeats in the process of gene expression might be underestimated. While a number of studies propose an evolutionary role for repetitive DNA and transposable elements especially [80-83], an influence on regulatory processes cannot be *a priory* excluded.

While the overall repeat content of the TD regions is very similar to the overall repeat content of the respective genomes, it was seen that the composition of repetitive DNA is distinctly different in transcription deserts and in the genome in general. It was shown that long interspersed nuclear elements (LINEs) play a much more important role in TDs in comparison to their overall role in the genome. There are two families of LINE elements in the human and mouse genome: LINE2 which is an older family and has been lying inactive since before the evolutionary split of mammals; and LINE1 which is still active at the present time. LINEs are a subgroup of retrotransposons, which again are a subgroup of transposons or all transposable DNA elements. The distinct characteristic of retrotransposons is the mechanism that is employed by them in order to gain mobility within the genome. Retrotransposons use an RNA intermediate and a reverse transcriptase (RT) to do so. In many cases the retrotransposons carry the DNA sequence that encodes for the RT itself. In contrast to that, other types

of transposons do not utilise an RNA intermediate and instead copy themselves directly from one genomic position to another. The role of retrotransposons in mammalian transcription was investigated in one recent study [84] which concluded that "retrotransposon transcription has a key influence upon the transcriptional output of the mammalian genome".

Since LINEs replicate themselves through a RNA intermediate and are thus transcribed, it is at first glance not clear why they should appear in such proportion in transcriptional desert regions. It is true, however, that most retrotransposons have become inactive through accumulated mutations and do not transpose anymore. The fact that LINE elements, and especially LINE1 elements, make up the majority of repetitive DNA in transcriptional deserts does nevertheless constitute an interesting observation that deserves further inquiry.

It was also observed that satellites, simple repeats and areas of low complexity have gained importance in TD regions in comparison to the overall genome. Unlike the observation made for LINE elements, an explanation for this can be found in a straight forward way. Mini- and micro-satellites have an extremely low propensity to be either transcribed or to initiate transcription by themselves, whereas they might well be part of larger primary transcripts (in introns, UTRs, etc.). They therefore appear more frequently in TD regions, although not all of them appear there.

Transcriptional desert regions have also been analysed for the presence of absence of Single Nucleotide Polymorphisms (SNPs). SNPs are sequence variations in corresponding locations between two DNA sequences. They consist of single nucleotide differences and the vast majority of SNPs do only produce two alleles [85]. In principle they can occur at any genomic location. Due to redundancies in the genetic code, a SNP that falls within a coding exon does not necessarily lead to the production of a different amino acid, but instead constitutes a 'silent' mutation. However, it is possible that SNPs in coding regions introduce a premature stop codon into the reading frame and with that, constitute a mutation that renders a gene or part of a

gene non-functional. While SNPs that fall within coding exons are surely the most interesting to look at, because they might have a direct influence on the type of amino acid and protein being produced by a gene, it is also interesting to look at SNPs that fall elsewhere on the genomic sequence. These SNPs might have different effects. SNPs that lie within the promoter region of a gene especially might have influence on the transcriptional activity of a gene [86,87]. In the most extreme case a sequence variation in the promoter region of a gene can lead to the silencing of the complete gene. But SNPs that are located distally from any gene location might also have influence in activities of enhancers or silencers including processes such as transcription factor binding and control of remote genes.

It was observed that there exists no significant correlation between the occurrence of SNPs and the existence of a transcriptional desert in the same region. Currently, the occurance of SNPs are understood to be random events. They are mutations that are introduced during the replication of the DNA before mitosis. It is sufficiently proven that environmental factors such as radiation or the presence of toxins have an influence on the development of mutations [88,89]. However, this makes them appear irrespective of the function of the DNA sequence they appear in. Given the fact that all DNA sequences are subjected to selective pressures in the context of evolution, it can be assumed that fewer mutations are observed in functional regions. Since we observe here that SNPs appears with equal frequency in TDs and transcriptionally active regions, it can be concluded that at least the occurance of SNPs suggests a functionality of TDs.

Recent studies have connected SNPs with susceptibility to disease and to responses to medication and vaccines [90-92]. However, most of the studies conducted in this field concentrate on SNPs that are located in coding regions and that directly modify gene products. The effects of these SNPs are most easily detected. It is equally likely that SNPs that fall within intergenic regions have an influence on gene expression by disrupting gene regulation. This influence can be more subtle than a change in the produced

amino acid, but nevertheless important for the overall cell function. The effect of the disruption of remote activators or repressors could be important for the regulation of a number of genes. Which of the SNPs that were identified in transcriptional deserts are candidates for this, depends on their co-occurrence with transcription factor binding sites (TFBSs). In fact, a recent study [93] showed SNPs occurring distally from any gene location that disrupt the expression regulation of the gene PTGER4 (Entrez Gene ID 5734) and thus contributes to the development of Crohn's disease.

The results and analysis presented here show that the concentration of TFBSs in transcriptions desert regions is slightly elevated when compared to other types of DNA sequences, which allows the conclusion that TDs might be active in terms of remote gene regulation by harbouring remote activator or repressor elements. More importantly, there exists a group of TFs that can be found in TD regions in a much higher concentration than in other DNA sequences. To be more precise, five TF binding matrices and 18 TFBSs were identified that occur in TDs 10 times more frequently than in other types of DNA sequences. This suggests that TD regions, more than other DNA sequences, constitute areas that provide the ground for a certain types of remote regulatory processes. The fact that it was shown - that some TD regions are significantly elevated with regard to the density with which TFBSs occur - can also be seen as an indication that some TD regions could play a role in remote regulatory processes. A comparison with the occurrence of TFBSs in promoter regions supports this hypothesis.

Desert377 on human chromosome 3 can be taken as an example, to illustrate the occurrence of TFBS clusters in TD regions. This TD region is 1222 nt long and located between nucleotide positions 20,872,124 and 20,873,345. The nearest known gene (ZNF385D, Entrez Gene ID 79750) to this location is located on the reverse strand more than 400 knt upstream. Nevertheless, this TD region harbours 58 windows of size 200 nt that each contain more than 35 possible binding sites for transcription factors. In total, there are 19 different TFBS matrices that can be identified to be located in this TD

region. Based on the relatively dense clustering of different TFBS in this TD region, it can be hypothesised that this region might have a regulatory function, despite its relative remoteness from any known genes.

To determine possible locations where the interaction of SNPs and TFBSs leads to changes in the behaviour of regulatory processes, the location of TFBSs and SNPs in transcriptional desert regions were matched. Table 3 and 4, of the online supporting materials show that SNPs overlap frequently with TFBS throughout the TD regions. In fact, the majority of TDs possess both SNPs and TFBSs, which is to be explained by the abundance with which both of these appear. It is open to speculation whether the occurrence of a single SNP in a potential binding site of a transcription factor will have a noticeable influence on cell behaviour, irrespective of whether such a SNP appears in the peripheral or the core area of the binding site. It seems clear, however, that the likelihood of the influence of such SNPs on cell behaviour drastically increases when the number of SNPs that appear in TFBS increases. Therefore TDs that harbour multiple TFBSs that are also the location of multiple SNPs, constitute areas of interest when it comes to irregularities in remote gene regulation processes. A comparison with promoter regions shows that the co-occurrence of TFBSs and SNPs is of a similar magnitude in promoter regions and TDs. Many studies have examined the effects of SNPs on transcription factor binding in promoter regions [94-96]. It can be assumed that for those transcription events whose control is contributed to by remote elements, the co-occurrence of a SNP and a TFBS in those remote elements could potentially have a similar effect to those described in these studies.

For example, a number of human TD regions that harbour a noteworthy number of TFBSs, with overlap between SNPs and TFBSs were selected.

TD region no. 3456 on human chromosome 10 is 4,292 nt long and is located between position 134,798,853 and 134,803,144. This area is located 29,817 nt upstream of gene KNDC1 (Gene ID: 85442) and 3684 nt downstream of gene GPR123 (Gene ID: 84435). Within this area there are

151 TFBSs, mainly for matrices CACD1, HAND1E47 and SPZ1. Within this area, there are also 38 SNPs that coincide with predicted TFBSs HAND1E47 and SPZ1 (14 core and 24 peripheral). TFs binding to those sites are suspected to have an important role in spermatogenesis and embryonic development [97,98].

Another example is TD region no. 820 on human chromosome 20. This area is 2899 nt long and located between position 62,226,987 and 62,229,885. It is located 18359 nt upstream of gene NPBWR2 (Gene ID: 2832) and 36386 nt upstream of gene MYT1 (Gene ID: 4661). Both neighbouring genes NPBWR2 and MYT1 play a role in the central nervous system development [99]. Within this area there are 181 TFBSs predicted, mainly for matrices TBP, OCT1 and PEBP. There are also 20 SNPs that coincide with TFBS matrix PEBP (11 core and 9 peripheral) and two SNPs that coincided with binding sites of OCT1 (1 core and 1 peripheral). Proteins potentially binding to PEBP binding site also belong to the AML family and are involved in the course of acute myeloid leukaemia [100,101].

There is also TD region no. 1142 on human chromosome 16. This area is 1489 nt in size, located between positions 87,015,823 and 87,017,311, and 30,204 nt upstream of gene ZFPM1 (Gene ID: 161882). In this region, there were 107 TFBSs detected, mainly by matrices PBX, PAX8, PAX6 and SPZ1. In this area, there are 17 SNPs that interfere with binding sites for TFBS matrix PBX (5 core and 12 peripheral). There are also five SNPs (1 core and 4 peripheral) that interfere with TFBS matrix SPZ1 (see above).

Also of interest is TD region no. 1713 on human chromosome 18, which is 1694 nt in size and resides between positions 37,182,312 and 37,184,005. Within this region there are 107 potential TFBSs, mainly for CART1 and POU3F2 binding matrices. Within this region, there are 12 SNPs that overlap with the location of TFBS matrix CART1 (7 core and 5 peripheral) and seven SNPs that interfere with the binding sites for POU3F2 (6 core). POU family TFs are brought into connection with mammalian brain development [102]. The region is located distally from any known-protein

coding gene. However the nearest gene is PIK3C3 (Gene ID: 5289). Irregularities in the transcription regulation of this gene are implicated in mental disorders [103,104].

These examples serve to illustrate the type of data that could be extracted from the analysis of TFBSs in conjunction with SNPs in TD regions of mammalian genomes. While it is by no means guaranteed that these mutations have any influence on gene expression or on the development of disturbances, they form candidate locations where a closer investigation might yield discoveries regarding transcriptional regulation or the deviation from normal gene expression regulatory patterns. The fact that these areas constitute regions of the genome for which no existing transcript can be found, and that transcription is also not likely to initiate in these locations, does not make them less likely to exhibit characteristics that have a potential influence on the regulation of gene expression. Two things should be kept in mind when investigating these regions for transcriptional remote control elements. Firstly, remote regulation seems to be an abundant feature in controlling gene expression [105] and secondly, disruption of TF activity is known to be linked to the genetic component of carcinogenesis [106].

Another important aspect is the evolutionary status of transcriptional desert regions. It was determined to what extent these regions have been conserved between the mouse and the human genome, which describes more than 70,000,000 years of independent development of those genomes. The fact that sequences are conserved during evolution has been linked to functions, such as long-range enhancing of flanking genes, regulating splicing, and transcriptional co-activation [107-109].

This study shows that the majority of TD regions that were determined are not conserved between the mouse and human genome. It also shows that conservation of TDs is about half as strong as the overall conservation between mouse and human. However, between one in five and one on seven TDs is in fact conserved, depending on the requirements for sequence similarity. Studies such as [42] have suggested that non-coding regions can,

in principle, be divided into those regions that display an inherent function and those that do not. If evolutionary conservation is taken as an indication of the existence of a function of a region with regard to gene expression, the fact that an estimated one in seven TD regions are strongly conserved between *Homo sapiens* and *Mus musculus* can be interpreted in such way that the existence of a function can be assumed for these non-transcribing elements.

All main findings for the various types of analyses performed on TDs are summarised in Table 19.

Desert size and coverage	Roughly between 2% and 7% TD coverage in human and mouse genomes, depending on TD creation parameters
GC-content	On average, significantly lower than whole genome, but high GC TDs as well as high AT transcriptionally active regions are frequent
K-mer composition	Specific k-mer composition of TDs can be sufficiently explained with low GC-content
Repeats	Similar repeat content to whole genome, but distinctly different repeat composition with special emphasis on LINEs
Single nucleotide polymorphisms	No correlation between SNPs and TD occurance found (suggesting function)
Transcription factor binding sites	Differences in TFBS composition between TDs and other types of DNA sequences can be identified. Some TFBSs appear predominantly in TDs.
Mutations and transcription factor binding sites	Co-occurrence of SNPs and TFBSs is frequently observed.
Transcription factor binding site clusters	Some TDs display a dense clustering of TFBSs that is significantly above average.
Evolutionary conservation	Between 1 in 5 and 1 in 7 TDs are evolutionary conserved between human and mouse.

Table 19: Summary of main TD findings

CONCLUSIONS

In this chapter a method was introduced that, based on the results of DDM and a comprehensive set of transcription data, identifies areas in mammalian genomes that are neither likely to initiate transcription nor are they part of any known transcript. These areas were named transcriptional deserts (TDs). It was shown that, using a minimal TD length as an artificial parameter, the area of mammalian genomes covered by TDs is relatively small, with more than 93% of genomes transcriptionally active in some way. The transcriptionally active area decreases if the minimal length requirement for transcriptional deserts is reduced. It was further shown that TD regions display a pattern of composition that is aligned with expectations derived from knowledge about GC and repeat content. It was nevertheless also shown that existing knowledge about repeat and GC-content is not sufficient to explain transcriptional activity in mammals. Analyses regarding SNPs and TFBSs demonstrated that TDs are heavily involved in remote regulation of gene activity and that some are candidates for examination regarding disease development, because they represent potential locations in which DNA sequence variations disturb remote gene regulation activity. The areas that were identified as transcriptional desert for *Homo sapiens* and *Mus musculus* are available from the online supporting materials to this dissertation (http://apps.sanbi.ac.za/dissertations/ulf/). The research presented in this chapter is currently being prepared for submission for publication in *BMC Genomics*.

Chapter 3 – Promoter extraction

INTRODUCTION

Every study performed in any field of science can only be as informative and insightful as the data it was performed on. Analyses performed on erroneous or incomplete data will always be deficient, no matter how robust the methodology. For all studies of transcription initiation, it is for this important reason to be able to extract promoters with high accuracy and precision, and to achieve the greatest possible coverage of promoters relevant for the study. For the context of this manuscript, a promoter is defined as a strand-specific DNA sequence that is located around an existing transcription start site (TSS). Therefore, a promoter has a certain length while a TSS always refers to a single nucleotide position on either the forward or the reverse strand of any chromosome.

The extraction of promoters is paramount to various kinds of studies. Firstly, there are studies of transcription in general, which rely on the determination of all existing promoters and/or TSSs with high accuracy and precision. For these types of studies, all existing TSSs need to be identified and added to the pool of data that is analysed. Missing genuine promoters and/or TSSs from the data or inaccurately determining them would impact negatively on the results of those kinds of analyses. The analysis that was described in Chapter one and the development of the DDM tool is an example for this kind of study. None of the claims made in Chapter one would have any validity if there was any doubt about the soundness with which the reference TSS set was established.

Closely related to this type of study are two types of analyses of transcription initiation that deal with specific subsets of promoters. One is the case in which only promoters for a certain group of genes are meant to be examined. In such cases, it is demanded that promoters for all genes in

this group are identified and allocated to their respective genes. Examples of such gene groups can be genes that share a common property, that are over- or under-expressed in a certain kind of stress situation or tissue type, or that are found to be associated with a certain type of disease. Once such a group of genes has been identified, it is the role of promoter extraction to identify TSSs for all genes in the group and to retrieve promoters for analysis. For example we studied features of ovarian cancer promoters [3]. Again, incomplete or erroneous determination of these promoters would impact very negatively on the results of the study.

Another type of analysis is the examination of promoters in a gene-independent yet tissue specific manner. For this purpose, the process of promoter extraction must be able to identify all TSSs that are reported within a certain tissue type. For this purpose it is necessary to assign tissue specific information to TSSs.

This chapter introduces PROMEX, a promoter extraction tool that achieves of the above mentioned requirements. PROMEX is a web-based promoter extraction tool that allows for the extraction of general, gene-specific or tissue-specific promoters. This tool was used to identify promoters for analysis in Chapter one as well as in [3-5]. A screenshot of the PROMEX tool can be seen in Figure 8.

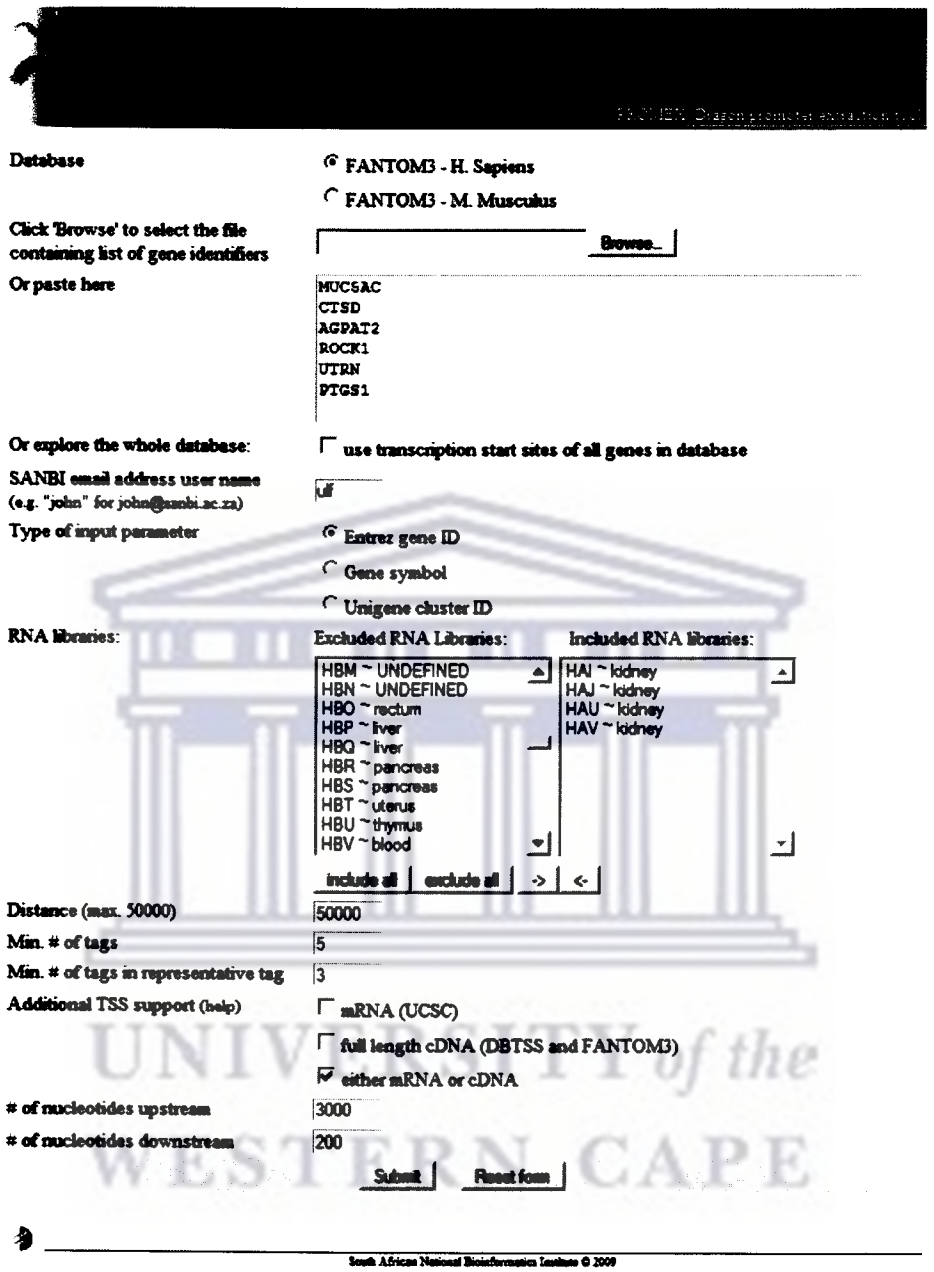

Figure 8: Screenshot of PROMEX tool

METHOD

PROMEX provides promoter extraction capabilities for the species *Homo sapiens* and *Mus musculus*. Following the above definition, the

identification of a promoter is primarily dependent on the identification of a TSS. Once a TSS has been identified, the extraction of the corresponding promoter is trivial. Starting from a given TSS location, a promoter is merely a sequence of DNA covering a certain number of nucleotides upstream and downstream of the TSS in question. PROMEX provides the means by which the user can specify this upstream and downstream length. If nothing is specified, it uses 3000 nt upstream and 200 nt downstream of each TSS.

CAGE

The identification of TSSs is, as mentioned above, an essential step when extracting promoters. PROMEX primarily uses cap analysis of gene expression (CAGE) data that was published in [12]. This data consists of a collection of several million DNA sequence tags that are between 18 and 25 nucleotides in length. These tags constitute the basic data that was obtained experimentally and they each represent the 5' end of a primary transcript. By mapping the tags back to the genome, the genomic origin of these transcripts could be identified. In this sense, each CAGE tags represents one piece of experimental evidence for the existence of a TSS at its corresponding chromosomal location.

However, the analysis performed as part of the CAGE effort goes further. In order to establish a measure of how strongly represented each single TSS is in the data, CAGE tags have been clustered together (for details on this clustering please refer to [12]). Eventually each TSS in the data is represented by two important characteristics. These are the total number of tags that were grouped together in this cluster, on the one hand, and, on the other hand, the number of tags that support the 'representative tag'. The 'representative tag' can be understood to be the strongest CAGE tag, in the cluster and the number of tags in the representative tag tells us how many tags are found at the location of this representative tag. The exact location of the 5' end of the representative tag of each tag cluster is interpreted to be the

location of the TSS. There are a lot of tag clusters that consist only of one tag. In this case, the only tag present is naturally at the same time the representative tag for this cluster.

PROMEX uses these two measures - the overall number of tags in the tag cluster and the number of tags in the representative tags - to let the user specify how strongly supported the TSS is, and with that how strongly supported the corresponding promoters are supposed to be. If the user does not change the default setting of the system, PROMEX applies a threshold of at least five tags in the overall tag cluster and at least three tags in the representative tag.

In addition to the exact location in the genome, each CAGE tags is also associated with the tissue library from which it was experimentally extracted. This allows for the extraction of promoters in a tissue specific manner. When determining how strong the support for a given TSS is, it is possible, with the help of this tissue information, to disregard all tags that did not originate from a certain type of tissue. This way, only promoters that are found in a certain environment can be extracted.

Other evidence for transcription

Since it is not optimal to depend on only one type of data, which was extracted in a specific experimental way, PROMEX introduces other types of evidence for the existence of a transcript in a specific location. These types of evidence are completely independent from the data that was obtained as part of the CAGE experiments. Any type of information that is based on two independent pieces of evidence can be regarded as much more accurate and much less error prone than information that is backed up only by only one piece of evidence. The data would only be incorrect in the extremely unlikely event that exactly the same error was repeated at two different points in time by two different types of data retrieval.

For the purpose of supporting TSSs, PROMEX utilises a collection of mRNA sequences extracted from the University of California in Santa Cruz (UCSC Genome Browser, http://genome.ucsc.edu/ [49]) and a collection of full-length cDNA sequences obtained from DBTSS [23] (both downloaded in April 2008). The exact location of all 5' ends of these sequences has been recorded, using the mappings provided by UCSC for genome build hg18. Each 5' end is interpreted as a TSS location. In addition to support by CAGE tags, the user of PROMEX has the option of selecting additional support by mRNA, or cDNA, or support by both. Currently PROMEX retains the locations of 1,615,187 cDNA sequences and 199,681 mRNA sequences for *Homo sapiens* and 458,321 cDNA sequences and 225,807 mRNA sequences for *Mus musculus*.

A TSS is considered to be supported by an additional piece of evidence if the 5' end of one of those sequences coincides exactly with the 5'end of the representative tag of the corresponding tag cluster. A mismatch is not allowed.

Assignment of gene identifiers

In order to be able to extract promoters for given user-specified genes, the system must retain gene information and allocate TSSs to as many genes as possible. PROMEX understands a TSS to be associated with a specific gene if the TSS is located on the gene body of this gene or within 50 knt upstream of the 5' end of the gene body. The distance of 50 knt is shortened accordingly if there is another gene located within this distance. As a consequence one gene can have no TSSs, one or more than one TSS, while a TSS always belongs to no gene at all or exactly one gene. There are a minor number of exceptions in cases of overlapping gene bodies, but they do not play a significant role in the system. The distance between the 5'end of the gene body and the TSS is reported as part of the promoter delivery. A TSS located downstream of the 5' end of the gene body (that is, on the gene

body) will be reported to have a negative distance to the 5' end of the gene body. The user can select to have only those promoters delivered that have a maximum distance from the gene. However, no distances greater than 50 knt will be reported.

PROMEX was equipped with the gene locations for three types of gene identifiers. A user can select any of these three to specify which gene promoters are desired. These types of identifiers are Entrez Gene ID, gene symbol, and Unigene cluster ID. PROMEX holds a list of these identifiers together with the chromosomal locations of the gene body for each gene in each of the three lists. Currently PROMEX retains the locations (including chromosome, strand, start, and end position) of 28,876 Entrez Gene IDs, 18,445 gene symbols and 22,873 Unigene cluster IDs for *Homo sapiens*, and 28,583 Entrez Gene Ids and 23,764 gene symbols for *Mus musculus*.

Depending on the number of gene identifiers submitted, the results are either returned immediately ("promoters while-u-wait") or are extracted offline and delivered to the user by email after extraction is finished. The latter is also employed if promoters are extracted in a gene independent manner.

DISCUSSION

The system described here provides several options for the extraction of promoters. The location of these promoters is based on the location of their corresponding TSS, which is determined on the basis of CAGE data. In addition to the CAGE data, other independent evidence for the existence of a transcript has also been added to the system. That makes the TSS locations determined by PROMEX highly reliable.

PROMEX offers a number of options for the user, to make the selection of promoters as flexible as possible and to cater for a number of types of

analysis for which promoters are required. Firstly, the user can choose how strongly the TSSs that correspond to the extracted promoters are represented by the data. Secondly, the user can specify how large the sections should be that are extracted as promoters around each TSS. Thirdly, the user can restrict the CAGE data that is supporting the desired TSS to a number of tissue libraries, thus enabling the user to extract promoters in a tissue-specific manner. Fourthly, the user can choose to extract promoters that have been allocated to one or more than one gene. These genes can be specified by one of three different gene identifiers. Alternatively the user can choose to extract the promoters in a gene independent manner. The following examples will illustrate the use of PROMEX.

Examples

A) Extract all promoters from *Homo sapiens* that have at least 1 CAGE tag in the overall tag cluster (and thus 1 CAGE tag in the representative cluster) and have at least one additional piece of independent evidence, either mRNA or cDNA. Extract 100 nt upstream and downstream of all TSSs and include all possible tissue libraries. This requests to PROMEX returns 113,814 promoters, exactly those that were used as the reference TSS set for humans in Chapter 1.

B) Extract all promoters from *Mus musculus* that are evident in liver tissue and have at least 2 tags in the overall tag cluster as well as 2 tags in the representative tag. Select only those promoters that also have at least one other piece of evidence of transcription. Extract promoters of length 3200 nt, 3000 nt upstream and 200 nt downstream of each reported TSS. This request to PROMEX returns 4825 promoters.

C) Extract all promoters for human gene FOXP2 (chromosome 7; location 7q31; 113,842,228 – 114,118,328) that have at least 5 tags in the overall tag cluster and at least 3 tags in the representative tags as well as either mRNA or cDNA support. Include all tissue libraries and extract 3000 nt upstream and 200 nt downstream of the TSS. This request to PROMEX returns 1 promoter. The corresponding TSSS for this promoter is located at 113,842,354. It has 14 CAGE tags in the overall tag cluster and 4 tags in the representative tag. This promoter is also supported by a full-length cDNA whose 5' end is located at 113,842,354. This kind of promoter extraction was done for [3], where a set of 379 identifiers for genes associated with ovarian cancer was submitted to PROMEX and the resulting promoters have been included in the analysis published as part of this study. The same applies for [4], where a group of 529 genes associated with oesophageal cancer was examined. For [5], a specially customised version of PROMEX was used to extract regulatory regions not for genes but for the transcription of miRNAs.

CONCLUSION

This chapter introduces PROMEX, a web-based promoter extraction tool. Since obtaining accurate data is paramount to the success of each study, PROMEX applies a rigorous methodology to determine the location of TSSs and extract the corresponding promoters. The location of TSS is mainly based on the location of CAGE tags, but is also optionally supported by other pieces of independent evidence. This technique makes the location of promoters highly reliable.

PROMEX offers high flexibility through a number of user-adjustable options. These include species selection between *Homo sapiens* and *Mus musculus*, promoter size upstream and downstream of TSS, strength of TSS support, tissue library information, and gene-specific promoter extraction.

Promoters extracted with PROMEX are highly flexible and reliable and have benefitted a number of studies including the study described in Chapter one, as well as several others [3-5].

Overall summary

The main subject of this dissertation is transcription initiation deserts and transcriptional deserts. The first chapter dealt with the detection of transcription initiation deserts (TID) and introduced a method for doing that. This method constitutes the only method available that is able to detect TID. It was shown how this method performs on a number of showcase chromosomes and it can be concluded that only a small fraction of the mammalian genomes is capable of initiating transcription. It was highlighted that this method is useful for researchers in a wide spectrum of life science research. The study presented in this chapter is currently in preparation for submission for publication in *Bioinformatics*.

The second chapter revolves around areas that include TID, but they are extended to include all known transcripts to make transcriptional deserts (TDs). The TD regions were analysed for a variety of aspects and their properties and potential functions have been highlighted. It was shown that these regions only cover a small fraction of mammalian genomes and that the GC-content of these regions is not sufficient to explain why they are transcriptionally silent. It was also shown that they possess a number of interesting characteristics that make them candidates for studies in remote gene regulation. The content presented in Chapter two can be seen as the direct consequence of the system proposed in Chapter one. In fact, Chapter two is the continuation and application of the design ideas from Chapter one on a whole-genome scale. The research presented in this chapter is currently in preparation for submission for publication in *BMC Genomics*.

The third and final chapter describes the process of promoter extraction and the tool that was developed for this purpose. It explains how the extraction of promoters is performed. The numerous options that are associated with the selection of promoters and the reasoning behind the methodologies applied were elucidated. It was also described how the data that was

obtained with the promoter extraction tool contributed to a number of analyses that were conducted during the time of my doctoral studies [1-6].

ONLINE SUPPORTING MATERIALS

- **FASTA files:** TID on human chromosome 21 (forward strand) at 3 different levels of sensitivity (OSMCD:/TID on human chr21/)
- **FASTA files**: TD regions for *Mus musculus* with 518 and 259 minimal length (OSMCD:/MM TD regions/)
- **FASTA files**: TD regions for *Homo sapiens* with 518 and 259 minimal length (OSMCD:/HS TD regions/)
- **Table 1**: k-mer analysis *Homo sapiens* (OSMCD:/Tables/OSM_table1_kmerPVal_hs.xlsx)
- **Table 2**: k-mer analysis *Mus musculus* (OSMCD:/Tables/OSM_table2_kmerPVal_mm.xlsx)
- **Table 3**: SNP TFBS cluster *Homo sapiens* (OSMCD:/Tables/OSM_table3_snptfbsClusters_hs.xlsx)
- **Table 4**: SNP TFBS cluster *Mus musculus* (OSMCD:/Tables/OSM_table4_snptfbsClusters_mm.xlsx)
- **Table 5:** Binding frequency for TFs in TD, cDNA, chr21 and random DNA (OSMCD:/Tables/OSM_table5_tf_bind_freq.xlsx)

APPENDIX

A) **Table A1:** DDM Masking at 99% SE for all human and mouse chromosomes

species	chromosome	% of valid positions likely to be TSS	% of valid positions unlikely to be TSS
Homo sapiens	chr1	23.02	76.98
Homo sapiens	chr2	18.98	81.02
Homo sapiens	chr3	17.21	82.79
Homo sapiens	chr4	13.54	86.46
Homo sapiens	chr5	16.87	83.13
Homo sapiens	chr6	17.03	82.97
Homo sapiens	chr7	20.07	79.93
Homo sapiens	chr8	18.35	81.65
Homo sapiens	chr9	21.69	78.31
Homo sapiens	chr10	22.59	77.41
Homo sapiens	chr11	22.80	77.20
Homo sapiens	chr12	19.96	80.04
Homo sapiens	chr13	14.77	85.23
Homo sapiens	chr14	20.52	79.48
Homo sapiens	chr15	24.34	75.66
Homo sapiens	chr16	31.01	68.99
Homo sapiens	chr17	34.11	65.89
Homo sapiens	chr18	17.68	82.32
Homo sapiens	chr19	40.73	59.27
Homo sapiens	chr20	30.09	69.91
Homo sapiens	chr21	21.40	78.60
Homo sapiens	chr22	42.40	57.60
Homo sapiens	chrX	15.86	84.14
Homo sapiens	chrY	16.11	83.89
Mus musculus	chr1	17.56	82.44
Mus musculus	chr2	21.09	78.91
Mus musculus	chr3	15.62	84.38
Mus musculus	chr4	22.07	77.93
Mus musculus	chr5	22.68	77.32
Mus musculus	chr6	18.46	81.54
Mus musculus	chr7	23.34	76.66
Mus musculus	chr8	21.80	78.20
Mus musculus	chr9	22.28	77.72
Mus musculus	chr10	18.64	81.36
Mus musculus	chr11	26.69	73.31
Mus musculus	chr12	19.83	80.17
Mus musculus	chr13	18.45	81.55
Mus musculus	chr14	17.63	82.37
Mus musculus	chr15	21.04	78.96
Mus musculus	chr16	17.54	82.46
Mus musculus	chr17	23.26	76.74
Mus musculus	chr18	18.11	81.89
Mus musculus	chr19	22.57	77.43
Mus musculus	chrX	12.40	87.60
Mus musculus	chrY	12.16	87.84

B) Table A2: Occurrences of TF binding matrices in TD, cDNA, chr21 and random DNA

AHR_Q5	0.001	0.008	0.184	0.005	0.305	0.022	0.065
AHRARNT_01	0.001	0.008	0.114	0.006	0.142	0.043	0.020
AHRHIF_Q6	0.001	0.013	0.070	0.005	0.196	0.041	0.023
AIRE_02	0.341	0.103	3.324	0.174	1.963	0.056	6.087
AP1_Q2_01	0.043	0.053	0.806	0.081	0.530	0.053	0.810
AP1_Q4_01	0.033	0.035	0.961	0.039	0.863	0.048	0.696
AP2_Q6	0.002	0.144	0.017	0.049	0.049	0.098	0.025
AP2_Q6_01	0.004	0.155	0.026	0.074	0.054	0.066	0.061
AP2ALPHA_01	0.000	0.089	0.003	0.021	0.013	0.034	0.008
AP4_01	0.005	0.076	0.067	0.044	0.116	0.027	0.187
AR_Q2	0.049	0.046	1.061	0.058	0.844	0.037	1.330
AREB6_02	0.021	0.028	0.734	0.039	0.532	0.028	0.742
ARNT_01	0.013	0.035	0.354	0.050	0.249	0.096	0.130
ATF6_01	0.015	0.071	0.213	0.042	0.357	0.177	0.085
BACH2_01	0.001	0.002	0.360	0.005	0.179	0.012	0.071
BRACH_01	0.000	0.000	NaN	0.000	1.222	0.000	NaN
BRCA_01	0.652	0.320	2.040	0.457	1.427	0.324	2.012
CACD_01	0.031	0.188	0.164	0.267	0.115	0.114	0.270
CART1_01	0.651	0.062	10.478	0.264	2.464	0.068	9.568
CBF_02	0.003	0.011	0.317	0.008	0.440	0.053	0.065
CDP_02	0.734	0.215	3.414	0.348	2.110	0.324	2.266
CDPCR1_01	0.121	0.102	1.189	0.107	1.125	0.389	0.311
CDPCR3_01	0.633	0.252	2.508	0.422	1.500	0.513	1.234
CDXA_02	1.990	0.267	7.457	0.822	2.421	0.248	8.025
CEBP_Q3	0.561	0.294	1.906	0.419	1.340	0.215	2.609
CEBPA_01	0.296	0.174	1.700	0.223	1.326	0.171	1.728
CEBPDELTA_Q6	0.066	0.033	2.003	0.065	1.013	0.034	1.935
CEBPGAMMA_Q6	0.547	0.123	4.444	0.317	1.722	0.057	9.588
CETS1P54_01	0.038	0.194	0.195	0.106	0.357	0.218	0.174
CETS1P54_02	0.083	0.129	0.649	0.095	0.879	0.139	0.600
CETS1P54_03	0.043	0.442	0.097	0.151	0.284	0.824	0.052
CHX10_01	0.006	0.001	6.550	0.004	1.424	0.010	0.579
CIZ_01	0.035	0.013	2.627	0.018	1.938	0.009	3.900
CMAF_01	0.063	0.178	0.352	0.118	0.529	0.097	0.646
COUP_DR1_Q6	0.023	0.052	0.445	0.044	0.527	0.048	0.479
CP2_02	0.045	0.195	0.229	0.129	0.346	0.141	0.317
CREB_02	0.011	0.085	0.125	0.039	0.276	0.208	0.051
CREB_Q4_01	0.014	0.041	0.338	0.032	0.435	0.163	0.085
CRX_Q4	0.052	0.026	1.969	0.070	0.735	0.037	1.397
DBP_Q6	0.187	0.124	1.503	0.183	1.023	0.126	1.484
DEAF1_02	0.001	0.016	0.083	0.006	0.233	0.042	0.031
DEC_Q1	0.006	0.021	0.291	0.023	0.271	0.022	0.279
DR1_Q3	0.015	0.020	0.746	0.023	0.656	0.017	0.899
DR3_Q4	0.072	0.256	0.282	0.174	0.416	0.231	0.313
DR4_Q2	0.006	0.028	0.232	0.057	0.113	0.019	0.338
E2_01	0.000	0.004	0.028	0.001	0.091	0.015	0.006
E2_Q6_01	0.000	0.004	0.112	0.002	0.262	0.017	0.026
E2A_Q2	0.000	0.000	0.397	0.000	0.081	0.001	0.012
E2F_03	0.001	0.015	0.058	0.003	0.252	0.078	0.011
E2F_Q6_01	0.001	0.026	0.053	0.006	0.229	0.136	0.010
E4BP4_01	0.003	0.001	4.461	0.002	1.091	0.002	1.361
EBF_Q6	0.001	0.004	0.179	0.002	0.283	0.001	0.689
EBOX_Q6_01	0.010	0.082	0.127	0.098	0.106	0.057	0.182
EGR1_01	0.003	0.059	0.055	0.024	0.137	0.097	0.034
ER_Q6	0.044	0.112	0.391	0.118	0.373	0.113	0.388
ETF_Q6	0.002	0.380	0.004	0.029	0.056	0.128	0.012
ETS_Q6	0.050	0.128	0.389	0.100	0.496	0.061	0.814
EVI1_03	0.006	0.001	5.499	0.003	1.952	0.000	NaN
FAC1_01	0.659	0.261	2.528	0.527	1.250	0.241	2.735
FOX_Q2	0.245	0.030	8.053	0.145	1.696	0.009	27.235
FOXP1_01	0.004	0.000	23.925	0.003	1.621	0.000	NaN
FXR_Q3	0.129	0.100	1.298	0.115	1.129	0.125	1.035
GATA_C	0.215	0.133	1.620	0.180	1.197	0.168	1.281
GATA4_Q3	0.355	0.147	2.421	0.248	1.435	0.075	4.738
GATA6_01	0.006	0.002	3.270	0.004	1.582	0.007	0.812

GCNF_01	0.002	0.019	0.121	0.006	0.360	0.001	2.286
GFI1_Q6	0.049	0.031	1.588	0.049	1.011	0.039	1.267
GLI_Q2	0.001	0.006	0.151	0.007	0.140	0.011	0.082
GRE_C	0.213	0.216	0.986	0.253	0.842	0.203	1.051
GZF1_01	0.000	0.000	NaN	0.000	NaN	0.000	NaN
HAND1E47_01	0.200	0.419	0.477	0.368	0.543	0.322	0.620
HEN1_01	0.000	0.000	0.000	0.000	0.000	0.000	NaN
HIC1_02	0.000	0.061	0.005	0.014	0.022	0.025	0.013
HIF1_Q3	0.001	0.018	0.030	0.006	0.084	0.022	0.025
HLF_01	0.026	0.009	3.077	0.018	1.452	0.043	0.615
HMGIY_Q6	0.340	0.229	1.483	0.242	1.407	0.126	2.701
HNF1_Q6	0.127	0.031	4.099	0.065	1.958	0.024	5.297
HNF3ALPHA_Q6	0.361	0.063	5.724	0.235	1.538	0.028	12.881
HNF3B_01	0.356	0.049	7.249	0.170	2.094	0.029	12.281
HNF4_Q6_01	0.015	0.017	0.856	0.017	0.898	0.018	0.831
HNF4ALPHA_Q6	0.080	0.113	0.708	0.098	0.815	0.071	1.129
HNF6_Q6	0.173	0.017	10.252	0.063	2.727	0.018	9.588
HOXA7_01	0.089	0.111	0.796	0.106	0.836	0.128	0.692
HSF1_01	0.017	0.020	0.832	0.018	0.929	0.013	1.292
HSF1_Q6	0.001	0.001	1.225	0.002	0.823	0.000	NaN
IPF1_Q4	0.174	0.075	2.304	0.127	1.364	0.130	1.337
IRF_Q6	0.024	0.007	3.551	0.013	1.836	0.000	NaN
IRF2_01	0.014	0.008	1.683	0.013	1.032	0.001	13.811
ISRE_01	0.014	0.005	3.068	0.010	1.380	0.001	13.847
KAISO_01	0.010	0.013	0.721	0.014	0.664	0.011	0.875
KROX_Q6	0.001	0.016	0.052	0.011	0.072	0.005	0.162
LEF1_Q2_01	0.036	0.046	0.776	0.038	0.947	0.038	0.944
LEF1TCF1_Q4	0.177	0.146	1.210	0.149	1.186	0.097	1.823
LHX3_01	0.039	0.004	9.470	0.022	1.799	0.001	38.990
LMO2COM_01	0.002	0.039	0.054	0.018	0.115	0.007	0.301
LRF_Q2	0.003	0.117	0.023	0.063	0.043	0.074	0.037
LUN1_01	0.000	0.001	0.216	0.007	0.044	0.000	NaN
LXR_Q3	0.000	0.001	0.294	0.002	0.107	0.000	NaN
LYF1_01	0.011	0.013	0.807	0.025	0.433	0.006	1.814
MAF_Q6_01	0.091	0.112	0.816	0.151	0.606	0.143	0.639
MAZ_Q6	0.024	0.067	0.366	0.110	0.222	0.025	0.978
MEF2_03	0.058	0.012	4.961	0.040	1.459	0.015	3.873
MEIS1_01	0.002	0.004	0.379	0.007	0.245	0.007	0.230
MEIS1BHOXA9_02	0.216	0.090	2.395	0.164	1.319	0.092	2.346
MINI19_B	0.006	0.094	0.063	0.057	0.105	0.054	0.110
MRF2_01	0.513	0.131	3.925	0.309	1.661	0.176	2.916
MTF1_Q4	0.000	0.001	0.035	0.002	0.022	0.003	0.012
MYB_Q3	0.149	0.273	0.547	0.242	0.617	0.325	0.460
MYCMAX_03	0.000	0.001	0.244	0.001	0.172	0.004	0.048
MYOD_Q6_01	0.002	0.029	0.067	0.026	0.073	0.019	0.102
MYOGNF1_01	0.242	0.839	0.288	0.632	0.383	0.691	0.350
NF1_Q6_01	0.021	0.086	0.246	0.054	0.392	0.067	0.317
NFAT_Q4_01	0.065	0.062	1.058	0.050	1.301	0.012	5.424
NFKB_Q6_01	0.004	0.014	0.294	0.014	0.286	0.010	0.404
NFY_01	0.006	0.009	0.606	0.016	0.366	0.009	0.634
NFY_Q6_01	0.008	0.012	0.697	0.013	0.637	0.009	0.895
NKX22_01	0.019	0.006	3.022	0.014	1.358	0.004	4.747
NKX25_Q5	0.071	0.078	0.916	0.092	0.770	0.064	1.111
NKX3A_01	0.049	0.005	9.950	0.018	2.793	0.007	7.023
NRSF_Q4	0.002	0.030	0.079	0.014	0.173	0.012	0.197
OCT_Q6	0.410	0.069	5.931	0.207	1.976	0.060	6.828
OCT1_02	0.584	0.167	3.494	0.328	1.784	0.172	3.398
OCT1_03	0.996	0.286	3.488	0.573	1.739	0.446	2.233
OCT1_07	0.219	0.035	6.205	0.115	1.914	0.053	4.138
OCT1_Q5_01	0.156	0.017	9.367	0.071	2.204	0.017	9.159
OCT4_01	0.288	0.066	4.338	0.151	1.906	0.031	9.282
P53_02	0.012	0.031	0.401	0.029	0.437	0.022	0.567
PAX_Q6	0.176	0.320	0.552	0.312	0.566	0.249	0.708
PAX2_01	0.383	0.396	0.966	0.405	0.946	0.554	0.691
PAX3_B	0.060	0.323	0.185	0.156	0.384	0.833	0.072
PAX4_04	0.637	0.173	3.676	0.553	1.153	0.124	5.137
PAX5_01	0.040	0.357	0.113	0.213	0.190	0.307	0.132
PAX5_02	0.084	0.454	0.185	0.277	0.302	1.159	0.072
PAX6_01	1.548	0.914	1.693	1.238	1.250	1.534	1.009
PAX6_Q2	0.005	0.020	0.225	0.025	0.184	0.007	0.648
PAX8_01	0.380	0.581	0.654	0.615	0.618	0.736	0.516
PBX_Q3	0.121	0.041	2.963	0.075	1.619	0.058	2.080
PBX1_03	0.093	0.040	2.336	0.062	1.493	0.043	2.165
PEBP_Q6	0.050	0.043	1.172	0.056	0.893	0.044	1.139
PLZF_02	1.806	0.359	5.031	0.904	1.997	0.209	8.641
POU1F1_Q6	0.077	0.008	10.271	0.032	2.427	0.008	9.672
POU3F2_02	2.051	0.371	5.532	1.010	2.030	0.486	4.219
POU6F1_01	0.073	0.007	10.168	0.034	2.166	0.013	5.653

PPAR_DR1_Q2	0.019	0.032	0.597	0.036	0.536	0.019	1.016
PPARA_01	0.241	0.257	0.941	0.300	0.806	0.225	1.072
PPARA_02	0.052	0.146	0.359	0.140	0.372	0.135	0.387
PPARG_01	0.012	0.022	0.547	0.019	0.656	0.020	0.613
PPARG_02	0.983	1.047	0.938	1.173	0.838	1.094	0.898
PR_01	0.001	0.001	0.500	0.001	0.554	0.001	0.641
R_01	0.000	0.002	0.011	0.001	0.019	0.002	0.012
RBPJK_Q4	0.002	0.016	0.119	0.013	0.141	0.031	0.060
RFX_Q6	0.086	0.292	0.293	0.246	0.347	0.419	0.204
RFX1_02	0.089	0.177	0.503	0.142	0.625	0.208	0.428
RORA1_01	0.011	0.007	1.598	0.011	1.028	0.004	2.839
RP58_01	0.000	0.000	1.190	0.000	0.713	0.000	NaN
RSRFC4_Q2	0.002	0.000	6.305	0.001	1.542	0.000	NaN
RUSH1A_02	0.118	0.061	1.950	0.087	1.360	0.103	1.147
S8_01	0.079	0.014	5.797	0.054	1.449	0.016	4.906
SF1_Q6	0.025	0.052	0.480	0.063	0.394	0.021	1.185
SMAD3_Q6	0.040	0.050	0.805	0.051	0.779	0.026	1.541
SOX9_B1	0.209	0.080	2.617	0.152	1.381	0.089	2.352
SP1_Q2_01	0.002	0.024	0.064	0.018	0.084	0.004	0.384
SP3_Q3	0.009	0.065	0.143	0.076	0.123	0.048	0.194
SPZ1_01	0.073	0.204	0.360	0.205	0.357	0.270	0.272
SREBP_Q3	0.055	0.085	0.649	0.135	0.408	0.104	0.530
SREBP1_01	0.038	0.044	0.860	0.173	0.221	0.069	0.554
SREBP1_Q6	0.049	0.170	0.287	0.180	0.271	0.107	0.457
SRF_C	0.006	0.005	1.222	0.008	0.709	0.006	0.919
SRF_Q4	0.003	0.002	1.846	0.005	0.646	0.000	NaN
SRF_Q6	0.006	0.005	1.326	0.010	0.621	0.006	1.078
STAF_02	0.003	0.012	0.230	0.011	0.252	0.005	0.539
STAT_Q6	0.120	0.194	0.619	0.144	0.831	0.063	1.905
STAT1_01	0.053	0.107	0.494	0.079	0.674	0.078	0.680
STRA13_01	0.000	0.000	0.198	0.000	0.081	0.003	0.004
SZF11_01	0.052	0.296	0.174	0.263	0.196	0.176	0.293
TAL1BETAE47_01	0.125	0.208	0.601	0.170	0.735	0.114	1.098
TAXCREB_01	0.004	0.043	0.096	0.028	0.149	0.114	0.037
TAXCREB_02	0.074	0.243	0.307	0.201	0.370	0.483	0.154
TBP_Q6	0.384	0.085	4.493	0.193	1.992	0.133	2.887
TBX5_01	0.125	0.186	0.672	0.231	0.542	0.173	0.722
TCF11_01	0.188	0.158	1.185	0.177	1.061	0.143	1.311
TEL2_Q6	0.036	0.114	0.318	0.079	0.456	0.075	0.483
TFE_Q6	0.041	0.021	1.969	0.049	0.830	0.023	1.780
TGIF_01	0.013	0.021	0.595	0.021	0.586	0.020	0.626
USF_Q6_01	0.011	0.038	0.296	0.043	0.259	0.080	0.139
VDR_Q3	0.045	0.161	0.281	0.180	0.251	0.104	0.435
VJUN_01	0.004	0.015	0.274	0.009	0.448	0.026	0.158
VMYB_02	0.122	0.507	0.240	0.230	0.530	1.717	0.071
WT1_Q6	0.007	0.037	0.198	0.040	0.182	0.006	1.211
YY1_Q6	0.076	0.118	0.643	0.116	0.652	0.049	1.545
YY1_Q6_02	0.070	0.127	0.555	0.092	0.765	0.048	1.464
ZBRK1_01	0.001	0.002	0.357	0.003	0.259	0.000	NaN
ZEC_01	0.000	0.000	0.297	0.000	1.222	0.001	0.073
ZF5_B	0.023	0.708	0.033	0.199	0.116	1.357	0.017
ZID_01	0.001	0.006	0.166	0.005	0.214	0.011	0.089
ZNF219_01	0.000	0.002	0.061	0.001	0.109	0.001	0.133

References

1. Hofmann O, Caballero OL, Stevenson BJ, Chen YT, Cohen T, Chua R, Maher CA, Panji S, Schaefer U, Kruger A et al.: **Genome-wide analysis of cancer/testis gene expression.** *Proc Natl Acad Sci U S A* 2008, **105**:20422-20427.

2. Sagar S, Kaur M, Dawe A, Seshadri SV, Christoffels A, Schaefer U, Radovanovic A, Bajic VB: **DDESC: Dragon database for exploration of sodium channels in human.** *BMC Genomics* 2008, **9**:622.

3. Kaur M, Radovanovic A, Essack M, Schaefer U, Maqungo M, Kibler T, Schmeier S, Christoffels A, Narasimhan K, Choolani M et al.: **Database for exploration of functional context of genes implicated in ovarian cancer.** *Nucleic Acids Res* 2009, **37**:D820-D823.

4. Essack M, Radovanovic A, Schaefer U, Schmeier S, Seshadri SV, Christoffels A, Kaur M, Bajic VB: **DDEC: Dragon database of genes implicated in esophageal cancer.** *BMC Cancer* 2009, **9**:219.

5. Schmeier S, MacPherson CR, Essack M, Kaur M, Schaefer U, Suzuki S, Hayashizaki Y, Bajic VB: **Deciphering the transcriptional circuitry of microRNA genes expressed during human monocytic differentiation.** *BMC Genomics* 2009.

6. Dawe A, Radovanovic A, Kaur M, Sagar S, Seshadri SV, Schaefer U, Christoffels A, Bajic VB: **DESTAF: A Database of Text-Mined Associations for Reproductive Toxins Potentially Affecting Human Fertility.** *Reproductive Toxicology* 2009, **under review**.

7. Wei CL, Ng P, Chiu KP, Wong CH, Ang CC, Lipovich L, Liu ET, Ruan Y: **5' Long serial analysis of gene expression (LongSAGE) and 3' LongSAGE for transcriptome characterization and genome annotation.** *Proc Natl Acad Sci U S A* 2004, **101**:11701-11706.

8. Hashimoto S, Suzuki Y, Kasai Y, Morohoshi K, Yamada T, Sese J, Morishita S, Sugano S, Matsushima K: **5'-end SAGE for the analysis of transcriptional start sites.** *Nat Biotechnol* 2004, **22**:1146-1149.

9. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: **Serial analysis of gene expression.** *Science* 1995, **270**:484-487.

10. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T et al.: **Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage.** *Proc Natl Acad Sci U S A* 2003, **100**:15776-15781.

11. Ng P, Wei CL, Sung WK, Chiu KP, Lipovich L, Ang CC, Gupta S, Shahab A, Ridwan A, Wong CH et al.: **Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation.** *Nat Methods* 2005, **2:**105-111.

12. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C et al.: **The transcriptional landscape of the mammalian genome.** *Science* 2005, **309:**1559-1563.

13. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC et al.: **Genome-wide analysis of mammalian promoter architecture and evolution.** *Nat Genet* 2006, **38:**626-635.

14. Bajic VB, Tan SL, Christoffels A, Schonbach C, Lipovich L, Yang L, Hofmann O, Kruger A, Hide W, Kai C et al.: **Mice and men: their promoter properties.** *PLoS Genet* 2006, **2:**e54.

15. Kapranov P, Drenkow J, Cheng J, Long J, Helt G, Dike S, Gingeras TR: **Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays.** *Genome Res* 2005, **15:**987-997.

16. Ovcharenko I, Loots GG, Nobrega MA, Hardison RC, Miller W, Stubbs L: **Evolution and functional classification of vertebrate gene deserts.** *Genome Res* 2005, **15:**137-145.

17. Rinn JL, Euskirchen G, Bertone P, Martone R, Luscombe NM, Hartman S, Harrison PM, Nelson FK, Miller P, Gerstein M et al.: **The transcriptional activity of human Chromosome 22.** *Genes Dev* 2003, **17:**529-540.

18. Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, Gingeras TR: **Large-scale transcriptional activity in chromosomes 21 and 22.** *Science* 2002, **296:**916-919.

19. Cohen N, Dagan T, Stone L, Graur D: **GC composition of the human genome: in search of isochores.** *Mol Biol Evol* 2005, **22:**1260-1272.

20. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G et al.: **Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution.** *Science* 2005, **308:**1149-1154.

21. Frohman MA, Dush MK, Martin GR: **Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer.** *Proc Natl Acad Sci U S A* 1988, **85:**8998-9002.

22. **Finishing the euchromatic sequence of the human genome.** *Nature* 2004, **431:**931-945.

23. Wakaguri H, Yamashita R, Suzuki Y, Sugano S, Nakai K: **DBTSS: database of transcription start sites, progress report 2008.** *Nucleic Acids Res* 2008, **36:**D97-101.

24. Bajic VB, Tan SL, Suzuki Y, Sugano S: **Promoter prediction analysis on the whole human genome.** *Nat Biotechnol* 2004, **22:**1467-1473.

25. Bajic VB, Brent MR, Brown RH, Frankish A, Harrow J, Ohler U, Solovyev VV, Tan SL: **Performance assessment of promoter predictions on ENCODE regions in the EGASP experiment.** *Genome Biol* 2006, **7 Suppl 1:**S3-13.

26. Knudsen S: **Promoter2.0: for the recognition of PolII promoter sequences.** *Bioinformatics* 1999, **15:**356-361.

27. Reese MG: **Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome.** *Comput Chem* 2001, **26:**51-56.

28. Davuluri RV, Grosse I, Zhang MQ: **Computational identification of promoters and first exons in the human genome.** *Nat Genet* 2001, **29:**412-417.

29. Down TA, Hubbard TJ: **Computational detection and location of transcription start sites in mammalian genomic DNA.** *Genome Res* 2002, **12:**458-461.

30. Solovyev VV, Shahmuradov IA: **PromH: Promoters identification using orthologous genomic sequences.** *Nucleic Acids Res* 2003, **31:**3540-3545.

31. Korf I, Flicek P, Duan D, Brent MR: **Integrating genomic homology into gene structure prediction.** *Bioinformatics* 2001, **17 Suppl 1:**S140-S148.

32. Ohler U, Stemmer G, Harbeck S, Niemann H: **Stochastic segment models of eukaryotic promoter regions.** *Pac Symp Biocomput* 2000,380-391.

33. Ponger L, Mouchiroud D: **CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences.** *Bioinformatics* 2002, **18:**631-633.

34. Bajic VB, Seah SH, Chong A, Zhang G, Koh JL, Brusic V: **Dragon Promoter Finder: recognition of vertebrate RNA polymerase II promoters.** *Bioinformatics* 2002, **18:**198-199.

35. Bajic VB, Chong A, Seah SH, Brusic V: **An intelligent system for vertebrate promoter recognition.** *Ieee Intelligent Systems* 2002, **17:**64-70.

36. Bajic VB, Seah SH: **Dragon gene start finder: an advanced system for finding approximate locations of the start of gene transcriptional units.** *Genome Res* 2003, **13:**1923-1929.

37. Bajic VB, Seah SH: **Dragon Gene Start Finder identifies approximate locations of the 5' ends of genes.** *Nucleic Acids Res* 2003, **31**:3560-3563.

38. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA et al.: **The sequence of the human genome.** *Science* 2001, **291**:1304-1351.

39. Deininger PL, Batzer MA: **Alu repeats and human disease.** *Mol Genet Metab* 1999, **67**:183-193.

40. Kazazian HH, Jr.: **Mobile elements: drivers of genome evolution.** *Science* 2004, **303**:1626-1632.

41. Zhi D, Raphael BJ, Price AL, Tang H, Pevzner PA: **Identifying repeat domains in large genomes.** *Genome Biol* 2006, **7**:R7.

42. Engstrom PG, Fredman D, Lenhard B: **Ancora: a web resource for exploring highly conserved noncoding elements and their association with developmental regulatory genes.** *Genome Biol* 2008, **9**:R34.

43. Chen N: **Using RepeatMasker to identify repetitive elements in genomic sequences.** *Curr Protoc Bioinformatics* 2004, **Chapter 4**:Unit.

44. Kanhere A, Bansal M: **A novel method for prokaryotic promoter prediction based on DNA stability.** *BMC Bioinformatics* 2005, **6**:1.

45. Carninci P, Westover A, Nishiyama Y, Ohsumi T, Itoh M, Nagaoka S, Sasaki N, Okazaki Y, Muramatsu M, Schneider C et al.: **High efficiency selection of full-length cDNA by improved biotinylated cap trapper.** *DNA Res* 1997, **4**:61-66.

46. Carninci P, Hayashizaki Y: **High-efficiency full-length cDNA cloning.** *Methods Enzymol* 1999, **303**:19-44.

47. Maruyama K, Sugano S: **Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides.** *Gene* 1994, **138**:171-174.

48. Sugahara Y, Carninci P, Itoh M, Shibata K, Konno H, Endo T, Muramatsu M, Hayashizaki Y: **Comparative evaluation of 5'-end-sequence quality of clones in CAP trapper and other full-length-cDNA libraries.** *Gene* 2001, **263**:93-102.

49. Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pheasant M et al.: **The UCSC Genome Browser Database: update 2009.** *Nucleic Acids Res* 2009, **37**:D755-D761.

50. Frith MC, Valen E, Krogh A, Hayashizaki Y, Carninci P, Sandelin A: **A code for transcription initiation in mammalian genomes.** *Genome Res* 2008, **18**:1-12.

51. Kawaji H, Frith MC, Katayama S, Sandelin A, Kai C, Kawai J, Carninci P, Hayashizaki Y: **Dynamic usage of transcription start sites within core promoters.** *Genome Biol* 2006, **7**:R118.

52. Bajic VB, Seah SH, Chong A, Krishnan SP, Koh JL, Brusic V: **Computer model for recognition of functional transcription start sites in RNA polymerase II promoters of vertebrates.** *J Mol Graph Model* 2003, **21**:323-332.

53. Geoffrey J.McLachlan: *Discriminant Analysis and Statistical Pattern Recognition.* Wiley Interscience; 2005.

54. Strausberg RL, Levy S: **Promoting transcriptome diversity.** *Genome Res* 2007, **17**:965-968.

55. Modrek B, Lee CJ: **Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss.** *Nat Genet* 2003, **34**:177-180.

56. Brett D, Pospisil H, Valcarcel J, Reich J, Bork P: **Alternative splicing and genome complexity.** *Nat Genet* 2002, **30**:29-30.

57. Watson MA, Darrow C, Zimonjic DB, Popescu NC, Fleming TP: **Structure and transcriptional regulation of the human mammaglobin gene, a breast cancer associated member of the uteroglobin gene family localized to chromosome 11q13.** *Oncogene* 1998, **16**:817-824.

58. Venables JP: **Aberrant and alternative splicing in cancer.** *Cancer Res* 2004, **64**:7647-7654.

59. Wang Z, Lo HS, Yang H, Gere S, Hu Y, Buetow KH, Lee MP: **Computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer.** *Cancer Res* 2003, **63**:655-657.

60. Cohen BA, Mitra RD, Hughes JD, Church GM: **A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression.** *Nat Genet* 2000, **26**:183-186.

61. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D et al.: **A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome.** *Science* 2008, **321**:956-960.

62. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5**:621-628.

63. Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G et al.: **Stem cell transcriptome profiling via massive-scale mRNA sequencing.** *Nat Methods* 2008, **5**:613-619.

76. Wingender E, Dietze P, Karas H, Knuppel R: **TRANSFAC: a database on transcription factors and their DNA binding sites.** *Nucleic Acids Res* 1996, **24:**238-241.

77. Pheasant M, Mattick JS: **Raising the estimate of functional human sequences.** *Genome Res* 2007, **17:**1245-1253.

78. Amaral PP, Dinger ME, Mercer TR, Mattick JS: **The eukaryotic genome as an RNA machine.** *Science* 2008, **319:**1787-1789.

79. Zoubak S, Clay O, Bernardi G: **The gene distribution of the human genome.** *Gene* 1996, **174:**95-102.

80. Charlesworth B, Sniegowski P, Stephan W: **The evolutionary dynamics of repetitive DNA in eukaryotes.** *Nature* 1994, **371:**215-220.

81. Elder JF, Jr., Turner BJ: **Concerted evolution of repetitive DNA sequences in eukaryotes.** *Q Rev Biol* 1995, **70:**297-320.

82. Finnegan DJ: **Eukaryotic transposable elements and genome evolution.** *Trends Genet* 1989, **5:**103-107.

83. Kidwell MG, Lisch DR: **Transposable elements and host genome evolution.** *Trends Ecol Evol* 2000, **15:**95-99.

84. Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T et al.: **The regulated retrotransposon transcriptome of mammalian cells.** *Nat Genet* 2009, **41:**563-571.

85. Brookes AJ: **The essence of SNPs.** *Gene* 1999, **234:**177-186.

86. De Gobbi M, Viprakasit V, Hughes JR, Fisher C, Buckle VJ, Ayyub H, Gibbons RJ, Vernimmen D, Yoshinaga Y, de Jong P et al.: **A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter.** *Science* 2006, **312:**1215-1217.

87. Knight JC, Udalova I, Hill AV, Greenwood BM, Peshu N, Marsh K, Kwiatkowski D: **A polymorphism that affects OCT-1 binding to the TNF promoter region is associated with severe malaria.** *Nat Genet* 1999, **22:**145-150.

88. Ward JF: **Radiation mutagenesis: the initial DNA lesions responsible.** *Radiat Res* 1995, **142:**362-368.

89. Hrabe de Angelis MH, Flaswinkel H, Fuchs H, Rathkolb B, Soewarto D, Marschall S, Heffner S, Pargent W, Wuensch K, Jung M et al.: **Genome-wide, large-scale production of mutant mice by ENU mutagenesis.** *Nat Genet* 2000, **25:**444-447.

90. Brookes AJ: **The essence of SNPs.** *Gene* 1999, **234:**177-186.

102. He X, Treacy MN, Simmons DM, Ingraham HA, Swanson LW, Rosenfeld MG: **Expression of a large family of POU-domain regulatory genes in mammalian brain development.** *Nature* 1989, **340**:35-41.

103. Stopkova P, Saito T, Papolos DF, Vevera J, Paclt I, Zukov I, Bersson YB, Margolis BA, Strous RD, Lachman HM: **Identification of PIK3C3 promoter variant associated with bipolar disorder and schizophrenia.** *Biol Psychiatry* 2004, **55**:981-988.

104. Tang R, Zhao X, Fang C, Tang W, Huang K, Wang L, Li H, Feng G, Zhu S, Liu H et al.: **Investigation of variants in the promoter region of PIK3C3 in schizophrenia.** *Neurosci Lett* 2008, **437**:42-44.

105. West AG, Fraser P: **Remote control of gene transcription.** *Hum Mol Genet* 2005, **14 Spec No 1**:R101-R111.

106. Tenen DG: **Disruption of differentiation in human cancer: AML shows the way.** *Nat Rev Cancer* 2003, **3**:89-101.

107. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D: **Ultraconserved elements in the human genome.** *Science* 2004, **304**:1321-1325.

108. Baira E, Greshock J, Coukos G, Zhang L: **Ultraconserved elements: Genomics, function and disease.** *RNA Biol* 2008, **5**.

109. Kuntz SG, Schwarz EM, DeModena JA, De Buysscher T, Trout D, Shizuya H, Sternberg PW, Wold BJ: **Multigenome DNA sequence conservation identifies Hox cis-regulatory elements.** *Genome Res* 2008, **18**:1955-1968.