Deciphering Transcriptional Regulation
Computational Approaches

By

Eivind Valen

A dissertation submitted to the University of Copenhagen in partial fulfillment of the requirements for the degree of Ph.D. at the Faculty of Science, University of Copenhagen.

December 2009

Supervisors:
Assoc. Prof. Albin Sandelin
Prof. Anders Krogh
Assoc. Prof. Ole Winther

DEPARTMENT OF BIOLOGY
UNIVERSITY OF COPENHAGEN
“I may not have gone where I intended to go, but I think I have ended up where I needed to be.”

-Douglas Adams
Abstract

The myriad of cells in the human body are all made from the same blueprint: the human genome. At the heart of this diversity lies the concept of gene regulation, the process in which it is decided which genes are used where and when. Genes do not function as on/off buttons, but more like a volume control spanning the range from completely muted to cranked up to maximum. The volume, in this case, is the production rate of proteins. This production is the result of a two step procedure: i) transcription, in which a small part of DNA from the genome (a gene) is transcribed into an RNA molecule (an mRNA); and ii) translation, in which the mRNA is translated into a protein. This thesis focuses on the first of these steps, transcription, and specifically the initiation of this.

Simplified, initiation is preceded by the binding of several proteins, known as transcription factors (TFs), to DNA. This takes place mostly near the start of the gene known as the promoter. This region contains patterns scattered in the DNA that the TFs can recognize and bind to. Such binding can prompt the assembly of the pre-initiation complex which ultimately leads to transcription of the gene. In order to achieve the regulation necessary to produce the multitude of tissues we observe, there exists a wide range of these TFs having different binding preferences and targeting different genes. By activating different TFs in a context dependent manner the organism can produce customized sets of proteins for each cell resulting in different cell types.

This thesis presents several methods for analysis and description of promoters. We focus particularly the binding sites of TFs and computational methods for locating these. We contribute to the field by compiling a database of binding preferences for TFs which can be used for site prediction and provide tools that help investigators use these. In addition, a de novo motif discovery tool was developed that locates these patterns in DNA sequences. This compared favorably to many contemporary methods.

A novel experimental method, cap-analysis of gene expression (CAGE), was recently published providing an unbiased overview of the transcription start site (TSS) usage in a tissue. We have paired this method with high-throughput sequencing technology to produce a library of unprecedented depth (DeepCAGE) for the mouse hippocampus. We investigated this in detail and focused particularly on what characterizes a hippocampus promoter. Pairing CAGE with TF binding site prediction we identified a likely key regulator of hippocampus.

Finally, we developed a method for CAGE exploration. While the DeepCAGE library characterized a full 1.4 million transcription initiation events it did not capture the complete TSS-ome of hippocampus. We fitted two statistical models to the CAGE data and extrapolated how deep sequencing needs to be to capture most of the events. We concluded that while most genes are discovered, tag clusters and TSSs are not fully explored.
Research Papers

First author positions are underlined, my name is listed in bold. Corresponding author is marked with a "*".

Primary Research Publications

I Valen, E.*, Winther, O., Sandelin, A. and Krogh, A.
Discovery of Regulatory Elements is Improved by a Discriminatory Approach
PLoS Computational Biology, 5(11) 2009

II Bryne, J.C., Valen, E., Tang, M.H.E., Marstrand, T., Winther, O., da Piedade, I., Krogh, A., Lenhard, B.* and Sandelin, A.*
JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update
Nucleic Acids Research, 36 (D102) 2008.

Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE
Genome Research, 19 2008. p255-265

IV Valen, E., Sandelin, A., Carninci, P., Krogh, A. and Winther, O.*
Estimating the coverage of tag-sequencing experiments at multi-level resolution
Submitted
Research Publications not Part of the Thesis

- Frith, M.C.*, Valen, E., Krogh, A., Hayashizaki, Y., Carninci, P. and Sandelin, A.*
  A code for transcription initiation in mammalian genomes
  Genome Research, 18 (1) 2008.

- Ahmed, S., Valen, E., Sandelin, A. and Matthews, J.*
  Dioxin increases the interaction between aryl hydrocarbon receptor and estrogen receptor alpha at human promoters
  Toxicological Sciences, 111 (2) 2009

  JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles
  Nucleic Acids Research 2009 [Epub ahead of print]

- The Fantom Consortium
  The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line
  Nature Genetics, 41 2008.

- Liu, Y., Gao, H., Marstrand, T.T., Strom, A., Valen, E., Sandelin, A., Gustafsson, J.A.* and Dahlman-Wright, K.*
  The genome landscape of ERα- and ERβ-binding DNA regions

  Asap: A Framework for Over-Representation Statistics for Transcription Factor Binding Sites

Book Chapters not Part of the Thesis

- Valen, E.* and Sandelin, A.*
  Future Challenges in CAGE Analysis
  Cap-Analysis Gene Expression (CAGE): the Science of Decoding Genes Transcription
  Chapter 15

- Sandelin, A.* and Valen, E.*
  Lessons Learned from Genomic CAGE
  Cap-Analysis Gene Expression (CAGE): the Science of Decoding Genes Transcription
  Chapter 14
Contents

Overview

Introduction ......................................................... 1
  Transcription Initiation by RNA PolIII and the Core Promoter .......... 4
  The Emerging Promoter in Light of CAGE .......................... 7
  Computational Sequence Analysis .................................. 12

Present Investigation ............................................. 21

Perspectives ......................................................... 31

Acknowledgement .................................................. 35

Research Papers

Paper I ................................................................. 43
  Discovery of Regulatory Elements is Improved by a Discriminatory
  Approach
  PLoS Computational Biology 5(11).

Paper II ................................................................. 53
  JASPAR, the open access database of transcription factor-binding pro-
  files: new content and tools in the 2008 update.
  Nucleic Acids Research 36:D102-D106.

Paper III ............................................................... 61
  Genome-wide detection and analysis of hippocampus core promoters
  using DeepCAGE
  Genome Research 19:255-265.

Paper IV ............................................................... 75
  Valen, E., Sandelin, A., Carninci, P., Krogh, A. Winther, O.
  Estimating the coverage of tag-sequencing experiments at multi-level
  resolution
  Submitted

Abbreviations ....................................................... 97
Introduction

Every organism on this planet consists of one or more small basic functional units. These building blocks of life are known as cells and come in two basic forms: prokaryotic and eukaryotic. The major difference between these is that the eukaryotic is larger and have several specialized sub-compartments called organelles. The prokaryotic cells are the choice of many unicellular organisms including two of the major branches of life: bacteria and archaea. Most multicellular organisms have more demanding organization and the vast majority of these are therefore eukaryotic.

Residing within these cells lies our basic hereditary material. As is common knowledge now this consists of deoxyribonucleic acid (DNA). DNA is a long molecule built from small basic building blocks: the nucleotides. These come in four different flavors: adenine, cytosine, guanine and thymine that fits together in pairs. Adenine can bond (or basepair) with thymine and cytosine can basepair with guanine. Assembled sequentially these form a long strand of pairs that is twisted around itself forming the famous double helix. Since the DNA in its native form is of substantial length it is usually packed into a compact structure known as chromatin. Besides the DNA, chromatin consists of a set of basic proteins called histones that the DNA is wound around which can attain condensed conformations.

A complete collection of DNA is known as a genome and for many organisms is often divided into smaller chunks known as chromosomes. In the eukaryotic cell these reside mainly in a large organelle: the nucleus, but a small part is present in many copies with each one placed in a comparably tiny organelle: the mitochondrion.

Spread across the DNA of each chromosome lies segments known as genes. While the exact definition of a gene is debatable and somewhat complex, a simplified definition would be: a segment of DNA forming a blueprint for a protein. This happens in two steps. The first, transcription, is facilitated by a protein using the gene DNA as a template for making a complementary ribonucleic acid (RNA) strand. The RNA is constructed by basepairing RNA to the DNA and then releasing the RNA product. This is possible because RNA is very similar to DNA just having a different backbone linking the nucleotides into strands. It also has a replacement for the nucleotide thymine known as uracil which can also basepair to adenine. Due to the rules of basepairing the new RNA molecule will now be the exact complement of the DNA template and is known as a transcript or messenger RNA (mRNA). Transcription starts at the transcription start site (TSS) residing at the very beginning of the gene in an area known as the promoter region and proceeds to the transcription termination site.

After a mRNA has been produced and undergone some processing it is exported out of the nucleus and into the cytosol which is the large space outside the organelles.

---

1 At this point technically a pre-mRNA
Here the second of the two steps, translation, take place. This is the process where a protein is produced based on the mRNA transcript. The main actor in this process is the large protein complex (a collection of many proteins and other molecules) known as the ribosome. Proteins are the basic building blocks of the cell and, similar to DNA, their primary structure is a long chain of smaller components. In this case they are called amino acids and there exists about 20 of them. Since the mRNA transcript consists of just 4 elements and a protein has 20 elements a certain language or code is needed to translate between the two “alphabets”. This translation is known as the genetic code and is with some exceptions universal to all life. The basic idea is that 3 nucleotides (a codon) uniquely identifies one amino acid. The translation is performed by small cloverleaf shaped molecules known as transfer RNAs (tRNAs) of which there are many types. Each of them can basepair with a codon at one end, using the region known as the anticodon (consisting of the complementary bases of the codon) and can have a certain amino acid corresponding to the anticodon attached to the other end. Using this system, the tRNAs can basepair with the next three nucleotides on the mRNA bringing with it the correct amino acid. This is then attached to the previous amino acids forming a chain. The ribosome makes sure that the mRNA is processed sequentially so that the amino acids are assembled in the correct order.

At the turn of the millennium a gargantuan task was nearing completion. Large headlines appeared in the papers that the human genome was finally sequenced promising to bring a cornucopia of benefits to medicine and basic understanding. Sequencing of DNA is the process of figuring out the order the 4 types of basepairs are arranged in on a DNA molecule. In the case of the human genome project the DNA was the 24 different chromosomes and 3.3 billion basepairs that the human genome consists of. It was the culmination of ten years of hard labor and 3 billion dollar in expenses, but sequencing technology has progressed a lot since then. For comparison the company Illumina recently announced that it could do full genome sequencing for $48,000 and it is expected that the $1000 genome will soon be within reach.

With the sequencing of an ever increasing number of genomes, one of the major challenges ahead is to understand how these blueprints result in the diversity of cells and behavior we see in every organism. In humans, liver cells are radically different from those of the cerebellum yet they both share the same genome. The difference lies in the interpretation of this recipe and deciding which genes and proteins are turned on and off during their life-time.

At the heart of the diversity lies the concept of gene regulation. This is a multifaceted, messy and hugely interconnected machinery featuring a plethora of cooperative and antagonistic agents that work on various steps and levels in the life cycle of the cell. Genes do not simply have on/off buttons, but function more akin to a volume control. They can be everything from more or less mute to turned up to maximum, cranking out RNAs at an alarming pace. While there are no actual buttons, the machinery takes cues from signals superimposed or encoded in the vicinity of the genes, binding agents that inhibit or stimulate the transcriptional expression.

Isolated, the in silico representation of the genome is only a long string of letters.

\[2 \text{Since } 4^3 = 64 \text{ multiple triplets code for the same amino acid}\]
(A,C,G and T representing each of the four nucleotides). For most organisms these strings are of a magnitude that makes human curation and interpretation impractical, hence the introduction of bioinformatics or computational biology. Building upon a long tradition in computer science and machine learning, it has rapidly gained popularity due to its necessity in the age of high-throughput experiments. Through the diligent application of both experimental and computational methods, patterns have started to emerge in the genome. These can make sense of how the cell is able to keep track of what is supposed to be expressed, how much and where.

The steps of transcription and translation provide at least two levels where regulation of this expression can act. This is naturally a simplistic view, almost deceivingly so, as we are continually observing RNAs that are not translated into proteins and are functional in and of themselves. These non-coding RNAs (ncRNA) are abundant and many of them seem to affect the production and life-cycle of other RNA molecules [1–5].
**Transcription Initiation by RNA PolII and the Core Promoter**

The general principles underlying the transcriptional machinery is the same for prokaryotes and eukaryotes, but is in the latter case much more complex. In bacteria and archaea transcription is carried out by a single enzyme: RNA polymerase (RNA Pol) while in eukaryotes there are three (I-III) used depending on the final product where RNA PolII is the one responsible for protein coding genes.

Transcription facilitated by RNA PolII can be roughly divided into three phases: i) initiation, where the factors assemble on the promoter and transcription starts; ii) elongation, where transcription moves out of the promoter and through the gene; and iii) termination, when RNA PolII reaches the end of the transcript and stops. Additional steps often included are “pre-initiation” being the formation of the so-called pre-initiation complex (PIC) and “promoter clearance” between initiation and elongation which, due to its complexity, merits its own stage. Several studies have targeted the complex behavior of the polymerase during this phase [6, 7]. This thesis however, will mainly focus on some of the more traditionally recognized steps of regulation: the involvement of transcription factors (TF) in transcription initiation and the concept of promoters particularly in light of new experimental techniques.

The DNA-dependent RNA PolII has the ability to unwind and rewind DNA as well as synthesize RNA [8], but it requires help in the recognition of promoter sequences shown by *in vitro* attempts to reproduce mRNA transcription [9,10]. RNA PolII therefore forms a larger system, the PIC, that assembles on the DNA. It comprises six proteins: RNA PolII and the five basal/general transcription factors: TFIIB, TFIID, TFIIE, TFIIF and TFIIH. In total over 30 polypeptides, but the exact subunits can vary according to the tissue and promoter type [11]. Various openings facilitate nucleotide access and exit for the RNA [12, 13]. Close to the exit lies the carboxyl terminal domain (CTD) of the largest subunit of RNA PolII. This contains a repeat sequence that is bound by several proteins involved in processing the RNA as it leaves the transcriptional complex.

Early promoter studies were mostly performed on highly expressed promoters, often tissue specific and with the property that they could easily be induced. Most of these contained an AT-region approximately 30bp upstream of the TSS termed the TATA-box after its consensus sequence. This element was discovered in 1979 by comparing 5’ flanking sequences of genes from Drosophila [14]. It was also revealed that the TATA-binding protein (TBP), a subunit of TFIID, was an essential part of the initiation complex that bound to this element. Thus the picture emerged that a promoter consisted of a TATA element surrounded by other more variable elements that produced differential expression.

The promoter core elements were initially believed to be invariant, but further studies revealed that not all promoters contained a TATA-box. In addition they could also contain other recurring patterns that were shown to be functional. Among these were the initiator element (INR), TFIIB recognition element (BRE) and the downstream promoter element (DPE) [15], where BRE is recognized by TFIIB and INR and DPE are recognized by TBP associated factors (TAFs). A given promoter could contain none,
one or more of these, but rarely many at the same time. Recently, other signals not part of the DNA, but instead modification made to the DNA packaging material: histones have been shown to facilitate binding. In particular a trimethylation of the amino acid lysine on the fourth residue of the histone 3 (H3K4) has been shown to promote binding by TFIID [16].

In addition to various novel core elements other findings also emerged to challenge the old dogma. Studies showed examples of promoters that had multiple initiation sites challenging the assumed specificity of TSS usage. Despite this, the TATA-box promoter with a single initiation site prevailed as the image of a typical promoter and is still the one most often presented in textbooks.

Transcription Factors and their Binding Sites

While the core initiation complex is required for transcription, it is certainly not the only agent involved. To achieve differential expression across genes, evolution has produced an abundance of trans-acting TFs that act on various cis-regulatory elements present in the vicinity of their target genes. By binding in proximity to their targets these act by repressing or activating their targets through chromatin modifications, recruitment of co-factors, or direct interaction and summoning of the PIC.

Unlike the TATA-box which is a general motif present in a multitude of genes these cis-elements can be limited to small groups of, and in some cases only, genes involved in very specific, rarely used pathways. Consequently, the TF or TFs\(^3\) that targets this element may only activate or repress a small collection of genes. Such targeted regulation is not uncommon and by having different cis-elements at different promoters and different transcription factors active in different contexts the cell can tightly regulate which genes are expressed at any given time during its lifetime.

To target a particular binding site it is necessary that TFs have certain preferences for the DNA that it binds to, binding strongly to some sequences and weakly to other. The actual binding site we will refer to as a transcription factor binding site (TFBS) while the model that attempts to represent and summarize the binding preferences of a TF will be referred to as a motif.

TFBSs can take many forms and it was shown early on that unlike restriction enzymes which have very specific targets, TFBSs can display considerable variance. An example of this can be seen in the \(\lambda\) operators which contain 12 half-sites where only two out of eight positions are completely conserved [17]. The reason for this lack of specificity is assumed to stem from the advantages this provides in regulating expression. As different instances of the motifs provide different binding efficacy, the promoter can be fine-tuned to provide a certain transcriptional output for that gene [18]. Since TFs also target multiple genes, a single factor can give rise to a wide range of output for its target genes using this system.

TFBSs are often fairly small, on the order of 6-20 bases. The longer ones are rarely conserved across the whole length, but often consists of two half-sites separated by a

\(^{3}\)Different TFs may target the same cis-regulatory elements
stretch of non-informative bases. A few are palindromic and many resemble each other and are closely targeted by classes of factors with similar binding profiles.

The quest to understand sequence specific binding started with the lac operon and the discovery that the regulation of expression was connected to a protein factor [18]. Concurrently with the experimental work, theoretical analyzes were being performed to understand the amount of information required for the functioning of regulatory systems [19, 20]. TFs are now understood to be highly dynamic, diffusing rapidly through the nucleus, directed to their sites partly by their transient binding to chromatin [21]. They may associate with DNA all over the genome, but having affinity for certain sequences ensures that the TFs spend more time at these positions thus increasing or preventing the chance of PIC assembly.

Unfortunately, the detection of these binding sites is complicated by their weak spatial restrictions. While most binding sites are assumed to bind in the vicinity of the TSS in the so-called proximal promoter region [22], many binding sites shown to be functional are clearly nowhere near this. Instead, these can reside in distal sites often tens of thousands of Kbp away (Fig. 1) and are then usually referred to as enhancers or silencers.

![Figure 1: Transcription factor binding sites (TFBS) can reside in the proximal promoter region or in more distal sites. Often sites cluster together into so-called cis-regulatory modules (CRM). Figure from [23].](image-url)
The Emerging Promoter in Light of CAGE

The advent of the cap-analysis of genome expression (CAGE) protocol brought many surprises that challenged the classical promoter view. CAGE is a sequencing dependent protocol developed in 2003 at RIKEN [24] that gives an unprecedented overview of the TSSs landscape. It does not capture all transcripts, only the subgroup that have undergone a modification known as capping, a universal step for RNA PolII transcripts.

Capping is a modification of the 5’ end of a transcript and takes place shortly after transcription initiation. It is the first of several modifications made to the mRNA and results in an inverted 7-methyl-guanosine being attached through a 5’-5’ triphosphate bridge to the initial nucleotide of the mRNA (Fig. 2) [25]. This structure is known as a cap or more specifically (as there are other types of caps): a m’Gppp cap or a Type 0 cap. The cap has several purposes: it can protect mRNA from degradation by exoribonucleases [26]; it is required for splicing; and it is involved in export from the nucleus.

The capping complex is associated with the CTD of RNA PolII and proceeds in three steps:

- RNA 5’-triphosphatase removes the terminal phosphate group from the mRNA, leaving 2 phosphate groups
- Guanylttransferase adds GTP to the phosphate groups, removing two phosphate groups from GTP in the process.
- Guanine-N7 methyltransferase methylates the 7-Nitrogen of the inverted guanosine.

The two former reactions are in mammals catalyzed by one protein: RNGTT/HCAP1, while the latter is overseen by RNMT.

The key element in CAGE is to sequence only the first 20bp of a transcript and map this back to the genome rather than the whole transcript. Since this greatly reduces the number of bases necessary to sequence from each transcript one can assess the start sites of a great number of these to the limit of the sequencing technology used. The downside is that you do not know the whole transcript and will therefore lose information about termination sites, length and splicing.

The protocol starts by using random or Oligo(dT) priming to synthesize cDNA depending one whether one wants to select for polyadenylated transcripts (Fig. 3). Using the Cap-trapping method [27] full length capped cDNAs are selected. The RNA is subsequently hydrolyzed and a double stranded linker is attached to the 5’ ends of the cDNA [28]. This linker contains the recognition sequence for the class II restriction enzyme MmeI close to its 3’ end. The second strand cDNA is synthesized and after that the sequences are subjected to MmeI. Upon recognizing the linker sites MmeI cuts 20bp downstream of these, effectively retaining the 5’ end of the cDNA. A second linker is then ligated to the new 3’ end and the whole is subsequently amplified by PCR using the linkers as primer sites. The resulting products are then concatenated and cloned into a vector for sequencing [29].
After sequencing, the 20bp long sequences called 'tags’ can be mapped to the genome using short read mappers such as bowtie [30] or bwa [31]. This effectively gives a genomic view over the transcription start site usage of the cell only limited by the sequencing depth. Also, since transcripts are sampled roughly proportionally to their abundance in the cell, CAGE correlates with the relative expression of each start site.

An initial surprise emerging from the first runs was that it appeared that most promoters did not conform to the TATA-box archetype. In fact the majority of promoters appeared to lack this particular element and could rather be characterized by their association with CpG rich regions termed CpG islands. Additionally, and perhaps the most surprising, a single initiation site for each gene seemed to be the exception rather than the rule. Instead most genes featured a distribution of TSSs where some were favored above others [33–35]. The initial reaction to this was of one disbelief and suspicion of experimental noise, but it was quickly realized that noise was not an adequate explanation for the observed data. In particular the following reasons were considered [36]:

- The classical single TSS, TATA-box-containing promoters were still observed albeit as a minority in the total population of promoters. These were positionally highly constrained across all tissues were the gene was expressed.

- In the broader promoters the relative distribution of tags were often preserved across many tissues. This is not compatible with random experimental noise. This was also observed across the species barrier as one discovered similar promoter distributions in human and mouse.
Figure 3: The CAGE protocol in 9 easy steps. Figure from [32].
After concluding that most genes were not too particular about their TSSs it quickly became apparent that it was beneficial to group these TSS into units reminiscent of the classical promoter. This was motivated by the fact that proximity made it likely that they shared regulatory input or were subject to the same constraints on nucleosome occupancy. To begin with, one grouped all tags that mapped to the same 5’ position into a CAGE-tag start site (CTSS) and used these to further cluster (based on overlap of tags) into larger units called tag clusters (TC) [34]. The TC including some of the surrounding sequence corresponds roughly to what in classical terms was known as the core promoter. As the overlapping of tags is a somewhat arbitrary criteria for clustering lacking a clear biological motivation more sophisticated clustering methods taking co-expression and strength (number of tags) into account were later devised [37, 38] to group TSS that have a high probability of being subject to the same regulatory machinery.

Given that one observed such different promoters as the TATA-box, single peak type and the broader, usually TATA-less and often CpG-associated types it was natural to classify the TCs into distinct categories. The first attempts at this divided the promoters into four categories (Fig. 4) [34], but this is often simplified into just a single peak (SP) category with 75% of the tags within 4nt of each other and a broad (BR) category which is a catch-all term for everything else.

Another surprise CAGE uncovered was the extensive use of alternative promoters. Some TCs targeting the same gene were clearly too far apart to be considered the same promoter. In fact the initial study revealed that 50% of known genes had two or more alternative promoters which is undoubtedly an underestimate. They are often clearly independently regulated and in those cases where you have multiple promoters each of them often show clear tissue preferences. Such a system makes it possible for a gene to have two different inputs which can help to organize differential regulation. Furthermore, it can also give rise to several different mRNAs and as such be a complement to alternative splicing.
Figure 4: The four types of promoters classified by Carninci et al. [34]. Figure from the same paper.
Computational Sequence Analysis

Identifying cis-regulatory elements and their shared motifs by experimental means can be an arduous task as experimental procedures have been either time-consuming (gene reporter/deletion assays) or lacking in the base level resolution that would be desirable (CHiP-chip [39]/CHiP-seq [40]). As a consequence much effort has been invested in producing computational methods as an alternative to experiments or as a complement to assays with lower resolution. Predictions from high quality methods can also be used to form hypotheses which could later be subjected to the more expensive, tedious and time consuming process of experimental verification. Furthermore, good models could shed light on the intrinsic features of the binding process and insight gleaned from the construction of these may increase our general knowledge of the nature of TF binding which could be exploited when new genomes are sequenced.

When discussing computational searches for motifs it is helpful to distinguish two different, but similar problems:

- In the first we know which motif we are searching for and are merely interested in its presence/absence or its position. I will refer to this as “motif finding”.

- In the second we suspect that some sites are present, but we do not know what the motif is nor the sites’ position. I will refer to this as “motif discovery”.

Note that these are not universal terms. They vary and are sometimes used interchangeably in the literature.

In both instances, it is common to start with a set of genes of interest derived from experiments or experience. Typically, one selects genes that are co-expressed and by extensions suspected to be under the control of the same regulatory machinery. Shared transcription factors often implies common motifs so this increases the likelihood of finding similar binding sites. Both “finding” and “discovery” are basically a matter of differentiating between signal and noise. The signals are the binding sites and everything that is usable in their detection and the noise is everything else. Besides the binding sites themselves there can be several other sources of information. Among the more popular is evolutionary conservation exploited in phylogenetic footprinting [41] and expression data which can be used to infer in which sequences the binding sites are most likely to be found [42].

Broadening our scope from a single factor, there are also many signals discernible in the interaction between factors. TFs often operate in larger groups known as cis-regulatory modules and we can infer information by looking at how the individual TF interacts at single promoters or across the genome.

De novo motif discovery is often aimed at either single motifs or the interaction of a few motifs since optimization of objective functions modelling larger contexts can be computationally hard. Interactions should in this case not be confused with simply searching for multiple motifs in a set of sequences which many motif discoverers do, but rather explicit modelling of co-occurrence. Motif finders on the other hand, starting

---

4 Other terms sometimes used in the literature are “motif scanning” and “motif searching”
<table>
<thead>
<tr>
<th>IUPAC Code</th>
<th>Meaning</th>
<th>Origin of Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>G Guanine</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>A Adenine</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>T Thymine</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>C Cytosine</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>G or A puRine</td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>T or C pyrimidine</td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>A or C aMino</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>G or T Ketone</td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>G or C Strong interaction</td>
<td></td>
</tr>
<tr>
<td>W</td>
<td>A or T Weak interaction</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>A or C or T not-G, H follows G in the alphabet</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>G or T or C not-A, B follows A in the alphabet</td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>G or C or A not-T (not-U), V follows U in the alphabet</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>G or A or T not-C, D follows C in the alphabet</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>G or A or T or C aNy</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: The IUPAC Nucleotide Code

with an already parameterized (set of) motif(s), are more free to focus on the higher levels of regulation as the motif optimization step is removed.

**Motif Definition**

Before embarking on the actual detection of a motif one needs a definition of what a motif is. In the case of DNA the simplest such definition is a consensus sequence using a string consisting of the four letters of the DNA alphabet: A,C,G and T. This, while sometimes used to describe binding preferences for certain TFs is unfortunately not a good general definition. Most TFs have lower specificity in their binding affinities at some of the spanned nucleotides perhaps preferring two out of the four bases or in some cases might even be totally indifferent. Indifference is usually modelled with the letter ‘N’ and in addition to this one can also include ambiguous letters representing one of two or even three nucleotides. This extended or “degenerate” alphabet (part of the IUPAC Nucleotide Code) comprises 15 letters (Table 1) and is used in several methods.

While the degenerate alphabet is more general than the base alphabet it still does not capture the full range of binding preferences for a TF. This has necessitated the exploration of more complex models of which the most used is the weight matrix (WM) first introduced for RNA translation initiation sites in *E. coli* [43]. Like the degenerate alphabet, a weight matrix allows ambiguous positions, but takes it one step further by letting each position vary freely. A weight matrix is commonly produced by aligning all known binding sites of a motif. Using 10 binding sites from the transcription factor ELK1 as an example (Fig. 5) we count the nucleotides at each position and enter them into a count matrix (Fig. 2(a)) also known as a position frequency matrix (PFM). This
is in turn can be normalized by the total number of counts in each column so that we end up with a probability matrix. As the name implies, this matrix gives the probability of having each letter at each position of the binding site (Fig. 2(b)).

\[ I(i) = 2 + \sum_{b=A}^{T} f_{b,i} \log_2 f_{b,i} \]

(1)

This is only meaningful if the distribution of bases is uniform in the genome, an assumption which is often violated as many organisms have biased genomes. *Saccharomyces cerevisiae* for instance has an A+T content of about 64%. Using equation 1 would result in seemingly information rich motifs when sampling and aligning random motifs from this genome [18]. To remedy this one can introduce the background distribution \( q \) which is simply the expected probability of each nucleotide given the genome. We then get the following equation known as “Kullbeck-Liebler divergence” or the “relative entropy”:

\[ I_{bg}(i) = \sum_{b=A}^{T} f_{b,i} \log_2 \frac{f_{b,i}}{q_b} \]

(2)

This is a measure of information in the site with respect to the background. There exists a nice graphical way of representing the information over a motif known as a logo. An example of this for the binding preferences of ELK1 factor can be viewed in figure 6. Here the height of the letter stack at each position corresponds to the information content at that site and the relative letter height is proportional to its frequency.

Motifs with higher scores imply lower variability and higher contrast to the background. As a side note, it is debatable whether the genome makes for a good background for matrices. Much of the genome is likely to be inaccessible and a good case
Table 2: Various matrices for the transcription factor ELK1 made from the above alignment (Fig. 5) and 18 other sites.

(a) Count matrix

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>7</td>
<td>10</td>
<td>9</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>27</td>
<td>21</td>
<td>13</td>
</tr>
<tr>
<td>C</td>
<td>7</td>
<td>6</td>
<td>4</td>
<td>19</td>
<td>24</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>10</td>
<td>6</td>
<td>10</td>
<td>1</td>
<td>24</td>
<td>27</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>T</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

(b) Probability matrix

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.25</td>
<td>0.36</td>
<td>0.32</td>
<td>0.18</td>
<td>0.07</td>
<td>0.00</td>
<td>0.04</td>
<td>0.96</td>
<td>0.75</td>
<td>0.46</td>
</tr>
<tr>
<td>C</td>
<td>0.25</td>
<td>0.21</td>
<td>0.14</td>
<td>0.68</td>
<td>0.86</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.18</td>
<td>0.00</td>
</tr>
<tr>
<td>G</td>
<td>0.36</td>
<td>0.21</td>
<td>0.36</td>
<td>0.04</td>
<td>0.04</td>
<td>0.86</td>
<td>0.96</td>
<td>0.04</td>
<td>0.00</td>
<td>0.50</td>
</tr>
<tr>
<td>T</td>
<td>0.14</td>
<td>0.21</td>
<td>0.18</td>
<td>0.11</td>
<td>0.04</td>
<td>0.14</td>
<td>0.00</td>
<td>0.00</td>
<td>0.07</td>
<td>0.04</td>
</tr>
</tbody>
</table>

(c) PSSM with a pseudocount of 0.1

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.00</td>
<td>0.51</td>
<td>0.36</td>
<td>-0.48</td>
<td>-1.76</td>
<td>-6.15</td>
<td>-2.69</td>
<td>1.93</td>
<td>1.57</td>
<td>0.88</td>
</tr>
<tr>
<td>C</td>
<td>0.00</td>
<td>-0.22</td>
<td>-0.79</td>
<td>1.43</td>
<td>1.76</td>
<td>-6.15</td>
<td>-6.15</td>
<td>-6.15</td>
<td>-0.48</td>
<td>-6.15</td>
</tr>
<tr>
<td>G</td>
<td>0.51</td>
<td>-0.22</td>
<td>0.51</td>
<td>-2.69</td>
<td>-2.69</td>
<td>1.76</td>
<td>1.93</td>
<td>-2.69</td>
<td>-6.15</td>
<td>0.99</td>
</tr>
<tr>
<td>T</td>
<td>-0.79</td>
<td>-0.22</td>
<td>-0.48</td>
<td>-1.20</td>
<td>-2.69</td>
<td>-0.79</td>
<td>-6.15</td>
<td>-6.15</td>
<td>-1.76</td>
<td>-2.69</td>
</tr>
</tbody>
</table>

could be made that one should only consider places where it is not impossible for the factor to bind.

All of the above mentioned models have in common that they consider each position in the motif to be independent of the others. While this is a nice property when doing calculations it is not simply something we can assume is true. Fortunately for us, while interdependencies do exist [45] in most cases assuming independence is a good approximation [46].

Despite this observation some attempts have experimented with higher-order models. This includes higher-order matrices and Markov models. A major roadblock for this has long been the lack of data necessary to fit the increased number of parameters, but with the introduction of high-throughput methods like CHiP-seq this should no longer be a hindrance.

**Motif finding**

Motif finding is the scenario where you have both motif model(s) and a set of sequences that you want to study. The models are usually compiled from experimental data (e.g. SELEX [47]) or a combination of this and computational methods (CHiP-seq [40] + motif discovery) which we will discuss in the next section. Recently k-mer arrays have been introduced [48] which have the potential to elucidate the binding preferences of
TFs in a comprehensive fashion. Several databases contain compilation of these motifs mostly in the form of WMs. JASPAR ([49, 50], paper II), TRANSFAC [51] and UniPROBE [52] are among the most well known.

When using a consensus sequence, motif finding is fairly straightforward. It is simply a matter of iteratively scanning through the sequences and seeing if any sites are present or not. However, knowing that sites are often variable one would normally allow some deviance from the consensus in the form of mismatches. This should be weighed up against the increasing number of false positives that we will identify if allowing too divergent sites. Using TATAAT as an example, given a random uniform sequence we can expect to see this every \(4^6 = 4096\) nucleotide. Allowing one mismatch this number is suddenly \(\frac{4^6}{2^6} = 227\) which given a typical range (e.g. -300 - +100 around the TSS) for searching a promoter would be expected to include at least one site by random.

For weight matrices the problem is somewhat more complicated. To assess whether a binding site for a probability matrix \(w\) of length \(|w|\) is present at a certain location \(i\) in sequence \(X\) one multiplies the probabilities from the weight matrix corresponding to the letter present in the sequences:

\[
P(X_i|w) = \prod_{j=1}^{\mid w \mid} w_{X_i+j,j}
\]

(3)

This gives the probability of the sequence given the weight matrix model. However, since long products of probabilities approach 0 quite rapidly this has the unfortunate consequence that computers (being poor at representing real numbers) experience problems accurately keeping track of the product\(^5\). Therefore, it is customary to use log transformed values and summing these instead:

\[
\log_2 P(X_i|w) = \sum_{j=1}^{\mid w \mid} \log_2 w_{X_i+j,j}
\]

(4)

\(^5\)This phenomenon is in computer science known as underflow.
The base two is a traditional computer science choice giving results in the unit “bits”. As before we would like to also consider the background of the genome to give us some indication of how surprising a certain result is. We can transform the probability matrix $w$ by dividing with the background probabilities $q$ and taking the logarithm (usually base 2) giving us what is known as a position specific scoring matrix (PSSM) (Fig. 2(c)):

$$PSSM_{b,i} = \log_2 \frac{w_{b,i}}{q_b}$$ (5)

This can be used to score sequences by summing over the cells corresponding to the sequence, where any score over 0 is more likely to be a product of the motif model rather than the background. It is also customary to add a small pseudocount to each cell in the matrix before transforming. The justification for this is that the data the matrix is built from is rarely complete and there is a big difference in having minuscule probability and a probability of 0. In addition log is not defined for 0 making the transformation problematic.

While a score larger than 0 indicates a higher probability of being generated by the motif rather than the background one rarely considers hits close to 0 as significant. Instead, a threshold is often used considering everything above as a real binding site. This can be calculated in various ways:

- Based on prior assumptions of how often the motif should occur [53].
- Based on how often it is deemed acceptable to occur in what is assumed to be non-regulatory sequences [54].
- Based on the scoring range possible using the PSSM. For instance 80% of the maximum achievable score.

No matter how one calculates the threshold, it is an ad hoc way of separating the wheat from the chaff based on the assumption that most PSSM-predicted binding sites are not involved in transcription factor binding. The assertion that most of these sites are not functional is known as the “futility theorem” [23]. Seen in the light of recent discoveries [21] it is likely that many of these are bound in vivo as well as in vitro, but that other conditions are not fulfilled for PIC assembly (e.g. other factors or histone modifications). Consequently, these events rarely lead to initiation.

**Motif discovery**

In motif discovery we also start out with a set of sequences that we are interested in, but unlike motif finding we do not know what pattern we are looking for. This is usually performed after first attempting motif finding since it is advantageous to exclude all known factors before attempting to discover new ones. The problem is analogous to the problem of finding optimal local multiple alignments and therefore, by reduction, belongs in the infamous class of intractable problems known as NP-complete [55].

Motif discovery is considered to be one of the classical problems in bioinformatics dating back at least 27 years [43]. As a consequence there are by now literally hundreds
of methods developed all varying in their choice of model or optimization. However, all of them rely on some common assumptions.

The primary reason computational prediction is at all possible is that the motif you are searching for is in some way over-represented compared to what is expected. This can be as simple as that a consensus sequence (e.g. TATAAT) occurs in every sequence, to the more realistic test that the number of times a PSSM scores above some specific threshold is greater than would be expected if the sequences were generated by a higher-order background model.

One can distinguish two approaches to the background model. In most cases this is modelled by a multinomial or higher order Markov chain that gives frequencies for the nucleotides, but lately more methods using a discriminative approach have been attempted [56, 57]. Discriminative in this context means that you have an additional set of sequences (a negative set) that are used to contrast with the set you suspect contain the motifs (the positive set). The advantage to this is that given a representative background it can perform better than any model approximation of this.

Additional constraints could also be added to get a more nuanced picture of the expected frequencies. For instance one could look at the position based over-representation noting that while TATAAT may not in itself be over-represented it could be the case that it is over-represented in the region -30 to -25 relative to the TSS. Sadly, such constraints are rarely as specific as in the case of the TATA-box. While most sites are in the vicinity of the TSS there is no guarantee they will have strict spatial constraints.

In addition to a motif model most methods feature an objective function describing what is a good configuration and how to interpret over-representation. This should capture whether whether motifs have to be present in each sequence (one occurrence per sequence (OOPS)), whether sequences without motifs are tolerated (zero or OOPS (ZOOPS)) or whether multiple motifs per sequence are allowed. It is also important in defining the nature of the motifs: are you trying to maximize information content, number of motifs, discriminative potential or a combination of these?

Finally, an algorithm for optimizing this function is needed. The choice of this is largely dependent on the motif model and objective function. Most methods fall into one of two categories: enumerative or statistical optimization methods, the latter can be further subdivided into deterministic and stochastic methods. Enumerative strategies proceed by in some way going through (“enumerating”) all possible motifs of a certain type. This is often used in conjunction with a sequence-based motif model. A good example is the Weeder algorithm [58] which explores sequences of length N with up to m mismatches. Statistical optimization on the other hand attempts to optimize an objective function through general or ad hoc optimization strategies. The Gibbs sampler [59] is one example using the Gibbs sampling method to optimize a weight matrix. Some other algorithms that have been attempted are expectation maximization [60], perceptron learning [57] and independent component analysis [61].

The accuracy of motif discovery tools has been fairly poor throughout the life time of the field. The main problem stems from the small size of motifs compared to the surrounding sequences. The positions with high information content are usually fairly short (6-8 bases) and can be hard to detect in the sea of noise surrounding it.
phenomenon of a too low signal to noise ratio is often known as “pattern drowning” and unfortunately affects many analyzes. To make things more difficult there are no hard rules for where motifs can be located only rough guidelines or conventions that have emerged. Typically, one ignores possible enhancers and silencers unless one has specific knowledge of their location. Instead it is common to search either the -300 to +100 region around the TSS or the -1000 to +200 region. The former choice is motivated by the fact that when plotting the mean evolutionary conservation of all promoters it tends to drop dramatically around the 300 mark. The latter choice is more motivated by the limits of pattern discovery tools. Both regions are asymmetric for the simple reason that searching far downstream quickly puts you in the middle of the coding region which has its own strongly biased sequences. However, recent investigations have shown that many sites are distributed symmetrically around the TSS [22]. Given that this is true this would indicate that many older results using these regions are possibly biased.
Present Investigation

“It’s sometimes called the final frontier. (Except that of course you can’t have a *final* frontier, because there’d be nothing for it to be a frontier *to*, but as frontiers go, it’s pretty penultimate...)”

-Terry Pratchett

The goal of promoter analysis is to elucidate the biological functions that produce the observed patterns. While the complete understanding of this lies beyond the ambition of this thesis, several steps have been made towards this goal. Some of the sub-goals include:

- Improvement of the historically poor performance of motif discovery tools.
- Extending the existing collections of motifs for transcription factors.
- Integration of these collections with tools to aid in the search for regulatory sites.
- Identification of key regulators in tissues by combining experimental data and computational predictions.
- Estimating the complexity of transcription initiation for tissues.
- Developing methods to deal with high-throughput tag data.

The four papers included in this thesis each addresses one or several of these points.

**Paper I** is a contribution to the motif discovery field featuring a novel discriminatory approach out-competing many contemporary methods.

**Paper II** extends the largest open-access collection of motifs and introduces several extensions and tools to help in this process.

**Paper III** presents the largest mapping of promoters in primary tissue to date. It uses a novel experimental technique paired with motif finding to identify one of the key regulators of this tissue.

**Paper IV** is the first attempt to estimate the size of the promoterome and TSS-ome based on CAGE data.
Paper I: Discovery of Regulatory Elements is Improved by a Discriminatory Approach

While published towards the end of my PhD, this paper marks the beginning of my venture into bioinformatics and addresses one of the fields’ classical problems: de novo pattern finding or “motif discovery”.

In particular, it is aimed at pattern finding in relation to transcription initiation. As discussed previously, transcription initiation is subject to control and regulation by several proteins termed transcription factors that can bind in the vicinity of the TSS. By bioinformatical standards this is an old field and by now there are literally hundreds of methods developed to attack this problem. Despite this, few are tested as extensively as we do in this article. This includes comparison to some of the more famous tools like MEME [60] and Weeder [58] where we compare favorably.

Using the classification from the introduction, Motif Annealer (MoAn) is a discrimi-
native statistical optimization method. The initial aim was to model cis-regulatory mod-
ules, but the nature of our model made the search space quite large making interactions hard to optimize. However, we realized that our objective function showed high corre-
lation with the correct solution on simulated data. This paired with a large negative set resulted in good performance on single motifs. The focus therefore shifted to whether we were able to out-compete contemporary methods at this task. Our method still has the capacity to model co-occurrence, but for this to be practical the optimization needs to be improved.

Comparing our method against the aforementioned tools as well as DEME [56] and NestedMICA [61] shows that our method is better at discriminating between the hetero-
geneous background of the mammalian promoter and transcription factor binding sites. This was concluded based on both synthetic sets (in line with the recommendations of a recent large scale evaluation of methods [62]) and on real data fetched from the PAZAR database [63]. We experience somewhat better results on the synthetic sets which we hypothesize may be due to the negative set being more representative of this background.

We also show that our method does not rely on the common pre-processing step of repeat masking [64]. This is a particularly good feature as some repeats have been shown to be functional [65, 66] with respect to transcription initiation. Indiscriminate use of masking, a step required by many methods, may therefore remove real signals.

Our method gains its power from two components: i) the large negative set and ii) the objective function. The complexity of scanning naively with a PSSM is linear with respect to the sequences (each positions is checked once) which would be intractable if the sequence set is sufficiently large. Instead we have implemented a data structure called an enhanced suffix array [67–69] which is a more efficient representation of a suffix tree [70]. This, combined with a threshold for PSSM searching, enables us to search all the sequences in an efficient manner.

As our objective function we use a conditional maximum likelihood for estimating the PSSM that best discriminates between the positive and the negative set. We let the matrix vary with counts independent of the sequences and therefore cannot necessarily
be derived from the sequences. The likelihood is formulated as the product over the sequences of the probability of the sequence label (positive/negative) given the sequence and the WMs.

Optimization is performed with a general optimization technique known as simulated annealing [71]. In short it performs random steps through the search space and then accepts or rejects these based on the difference in log-likelihood between the current and the proposed state. Acceptance is also influenced by a temperature parameter that gradually shrinks the space of valid moves, accepting fewer and fewer bad ones. The idea is that after exploring the search space the method gradually converges towards the global maxima before finally getting trapped there by the low temperature. At the time of writing the optimization is the weakest part of this method.
Paper II: JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update

The JASPAR database was motivated by the lack of an open source, free collection of binding motifs for transcription factors. Unlike its commercial competitor TRANSFAC [51], JASPAR focuses on quality before quantity. This is accomplished by using human curators who ensure that every matrix lives up to experimental standards and that every factor is represented by the best available matrix. The original database is highly cited and forms the backbone of many motif-finding applications.

This paper describes its third major release and is the culmination of a collaboration between the Computational Biology Unit in Bergen and The Bioinformatics Centre. It has three main authors focusing on different aspects. Jan Christian Bryne developed a web service Java library for interaction with JASPAR, Man-Hung Eric Tang focused on data curation and I developed and extended the functionality of the database through integration with tools for dynamic clustering and random profile generation.

The release features an extended core database with motifs that passed our quality standards as well as three new sub-databases:

**POLII** Core promoter motifs (e.g INR and BRE) [35, 72]

**CNE** Conserved non-coding elements from Xie et al. [73]. Many acting through long-range effects as enhancers.

**SPLICE** Splice sites as matching donor acceptor pairs [74]

In addition we provided several additional summary statistics showing among other things the number of spurious hits you would expect to see in random sets of promoters.

We also developed a WS-I compliant Web Service interface. This is coded in Java and is an API to simplify external utilization of the database. In particular to make it easier to use JASPAR in workflow managers like Triana [75] or Taverna [76].

Another important addition was the functionality to create custom familial binding profiles (FBP). Many TFs target similar motifs and by clustering them to a FBP you get models describing a set of matrices. This functionality is provided by the underlying STAMP [77, 78] tool. This also provides functionality to align matrices to the whole JASPAR database, searching for the best match. The latter is particularly useful if a motif discovery tool has found a motif you suspect is already known.

This release also provides a way of generating ’random’ matrices. This can be useful in many assessments for instance a scenario where you want to judge the significance of finding a certain number of sites matching a matrix. It is desirable that the matrices, though random, still share common properties with real binding sites. We therefore provide two methods to generate these:

- In the first we simply shuffle the columns within a matrix thus generating a new matrix with the same information content as the first.
• In the second method we sample random columns using a statistical model. A posterior distribution consisting of a multinomial using counts from real matrices and a Dirichlet mixture prior trained on the observed nucleotides in the JASPAR database.

Both methods assume independence of columns.

Since its inception the JASPAR database has become a mainstay of the motif-finding community and is used extensively in our work including paper III and [79–81].
Paper III: Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE

This work represents the first instance of the DeepCAGE protocol. DeepCAGE is the merging of the Cap Analysis of Gene Expression (CAGE) protocol with high-throughput sequencing. In this case the 454 Life Sciences (Roche) GS20 sequencer provided 2 million randomly primed tags from TSSs used in the mouse hippocampus. At the time of publication this was the most comprehensive landscape of PolII TSSs compiled for any tissue. Of the 2 million, we could unambiguously map 1.4 million to the mouse genome.

This is a technology-driven paper that explores what information one can get by sequencing to this depth. To this end, we compared our data with 7 other tissues: cerebellum, embryo, liver, lung, macrophages, somatosensory cortex and visual cortex compiled from a total of 39 libraries.

We first clustered all tags based on proximity which resulted in 18,948 TCs. Since the libraries spanned a wide range of sizes it was necessary to normalize the counts in each tissue to tags per million (TPMs) to have a basis for comparison. It should be noted that since we do not know the actual population size of the transcripts in the cells we can only compare relative preference within a cell. This is distinct from actual transcription strength, measured in initiation events per time unit. We discuss this further in [82].

After normalization we performed hierarchical clustering on the TCs versus the TPMs from each tissue for each of these. This showed, as expected, that the brain tissues are more highly correlated to each other than the rest of the tissues and, in particular, somatosensory and visual cortex are similar in terms of transcription initiation.

Seeking to identify the unique features of hippocampus, we sought to identify promoters that were primarily used in this tissue. Historically one has referred to tissue specificity, but with the advent of CAGE this term has lost much of its meaning since:

- Few TCs have tags exclusively from one tissue.
- Compared to all cell types and tissues, we have sequenced relatively few.
- Tissues can not be sequenced to infinite depth and in most cases we are far from detecting everything (as discussed in paper IV).

To address this, we derived the concept of a preferentially expressed promoter (PEP). The PEP definition is not particularly robust to generalization, but focuses on creating a subset that is strongly expressed and over-represented in one tissue for further analysis. In total we ended up with 6536 of these divided among the tissues with hippocampus having the highest number of biased promoters. In particular we identify many hippocampus TCs in intronic and intergenic spaces.

We further discovered that many genes harbor alternative promoter for different brain tissues. This is notable since it gives the cell the possibility of having individually customized control for each tissue. In addition, having promoters that captures different open reading frames alters the final protein product and can include or exclude important
protein domains. To investigate this possibility we used annotation of protein domains to search for hippocampus PEPs located downstream of these, but within the same gene. This resulted in about 50 genes where the protein products changed in an important manner compared to the full length alternative.

Following this we performed motif finding using the in-house developed ASAP [79] tool. Scanning the -1000 to +200 region of each PEP we identified a number of binding sites that were over-represented in each tissue. Since CAGE tags are also correlated with expression we used these to investigate which TFs were both strongly expressed and specific to hippocampus. Pairing this with the over-represented binding sites, we observed that the Arnt2 fulfilled all of these criteria, being highly expressed, specific to hippocampus and having over-represented binding sites in hippocampus PEPs. We compared this to in situ images from mouse brain, which confirmed distinct expression of Arnt2 in the C1 region of hippocampus. These converging lines of evidence implicated Arnt2 as an important factor in hippocampus.

All PEP analysis was performed on fairly strong TCs (>30TPMs). Since TCs with low number of tags are often met with distrust and dismissed as experimental noise we decided to explore some of these further. Using in situ we investigated the spacial expression of several promoters for a wide range if TPMs. The picture that emerged was that promoters with low expression is often a result of a small number of cells with strong expression. Since we are sampling from the whole hippocampus these get averaged out in the expression estimates. Sometimes these cells form well defined groups (e.g. Chek2) and probably play an important physiological role. In other cases (e.g. Nmbr) no clear groups are formed and functional roles are more uncertain.
Paper IV: Estimating the coverage of tag-sequencing experiments at multi-level resolution

This paper was inspired by a question that came up during the making of Paper III: How much do we need to sequence to hit every promoter and every TSS for a particular tissue? That is, how many CAGE tags do we need for detecting “everything”? While seemingly a simple question it conceals many complexities. One of these arise from an argument we put forth in [37] where we present the case that any nucleotide can be a potential TSS under the right conditions.

Under this paradigm, the conclusion is that we need enough CAGE tags to hit all positions in the genome once. This is not a satisfactory answer because obviously some nucleotides will have a minuscule chance of initiating a transcript. Nucleotides in constitutive heterochromatin or in the vicinity of (but not the target of) strong directional signals (e.g. TATA-boxes) are good examples.

Initially, we started out with very simple models trying to fit an exponential function to the data. These results are described in a book chapter recently published [82]. To capture the complexity of the promoterome it was necessary to employ more advanced models. We introduce two methods not previously applied to TSS analysis. One based on a generalized inverse Gaussian model (GIGP) that has been used in the context of SAGE. The other, a Pitman-Yor process makes no assumptions about the distribution (non-parametric) and has been previously applied to small EST libraries.

We use these to estimate transcription on three different levels: i) gene, ii) TC and iii) TSS. For both models there are two variables of particular importance: \( N \) being the size of the library (i.e. the number of CAGE tags) and \( k \) which is the number of entities (genes/TCs/TSSs) that are observed at size \( N \). By fitting our models to the observed data we can extrapolate to larger libraries and estimate how many new entities we expect to observe in larger libraries. This provides us with some indication of the total number of \( k \)'s for each class and tissue and can provide us with clues to how much depth we need for an adequate overview of the transcriptional initiation landscape.

The non-parametric model assumes an infinite population and as such will never have a probability of 0 of sampling a novel entity\(^6\). GIGP on the other hand operates with a finite population and can therefore estimate the total population size. Unfortunately, this is usually a considerable extrapolation. Small errors will therefore have a large impact on the total size and the huge estimates may therefore not be particularly useful. As an alternative in this paper we operate with the concept of “coverage”. This is estimated with the non-parametric model and is the one minus the probability of sampling a new unseen entity given our current data. Consequently this gives an indication of when further sampling is unlikely to result in more information.

On the biological side we can conclude that for the larger libraries few new genes will be hit with an increase in sequencing. Most genes that are expressed at a reasonable level are with large probability already hit by at least one tag. Of course, larger libraries will

\(^6\)Only for special parameter settings, not seen when estimated in our libraries, will the estimate be finite.
still be valuable as verification since one usually discounts single tags as unreliable.

For TSSs on the other hand we still have a long way to go. The largest library hippocampus, also obtains the largest coverage of 86%. This means that we have a 14% chance of sampling a novel TSS if we sampled a single new tag. This indicates that we still have much to gain from larger libraries. TCs naturally lie somewhere in the middle of genes and TSSs with a coverage of about 90% for the large libraries, showing that we also here have some room for new discoveries. The good news is that for many of the tissues a decent coverage of TCs and TSSs seem within reach of current technologies. Embryo, hippocampus and macrophages all achieve or come close to 90% coverage on TSS level and 95% on TC level with about 5 million tags.

We also assess the methods by sub-sampling half the size of each library (no replacement). Subsequently, the models are fitted to this reduced data set and are then made to predict the size of the original library. Ideally, the predicted $k$ should be close to the actual observed $k$. To increase the certainty of our assessment we do this by cross-validating ten samples.

In general the methods perform well with the non-parametric model having a slight tendency to over-estimate and the GIGP to underestimate. Both are still within a few percent margin of error.
Perspectives

“May you live in interesting times.”

-Chinese curse (reputedly)

While there are many remaining mysteries in gene regulation, the initiation of transcription has been particularly targeted in the last few years. Interesting aspects of both the dynamics of transcription factors and the initiation machinery [21] have recently been uncovered making this an exiting time to be in gene regulation.

Despite the strong efforts that have been put into this field, there are many fascinating problems that are still unsolved. Understanding the intricacies of polymerase behavior in initiation such as backtracking, pausing, recycling and the switch to the elongation phase is now a high priority. In parallel, many researchers are exploring the large number of small non-coding RNAs (ncRNA) that seem to originate from the vicinity of the TSS or the termination site and some evidence has emerged that they play a role in regulation [2, 3].

The studies presented in this thesis helps to shed light on some of the fundamental features of this process and provides tools for researcher interested in promoter analysis. We have advanced the predictive powers of motif discovery by showing that our tool compare favorably to many contemporary methods. In particular MoAn dispenses with the pre-processing step of repeat masking which at best is *ad hoc* and in the worst case can obfuscate real signals.

We have also furthered the motif finding field and during the writing of this thesis, JASPARs fourth release was also published, providing the largest motif set expansion to date [83]. Included in this release were also three new sub collections featuring data from recent high-throughput experiments. With the emergence of new methods this database is likely to increase exponentially in the following years.

The CAGE data demonstrates that we still have much to learn on the extent of transcription. We have observed that many tissue-specific, clearly differentially regulated promoters reside in intergenic space suggesting that many genes may still be undiscovered. We also have presented evidence that far from all promoters have been discovered yet. We should therefore be cautious in claims about how much of the genome is functional which seems to be fashionable these days as this promises to uncover many new transcripts. This is further strengthened by the recent discoveries that transcription is ubiquitous in that most nucleotides are at some point part of a transcript [22, 33, 84, 85].

In the intersection of machine learning and CAGE a particularly interesting problem would be to create a model for the genomic landscape of CAGE. Since CAGE is correlated with expression this should be possible with the right input. A first stab at this
was made by us in Frith et al. [37] by trying to predict CAGE distributions in isolated promoters using only sequence data. That study revealed that while it was possible to predict the relative expression of the TSSs (normalized over a TC), predicting the absolute number of tags was not. This hints that there are several layers of regulation where the local structure is the major determinant of which TSS to use while other non-local or epigenetic signals determine the rate of initiation.

In the classification of promoter types work is also progressing. While it is known that the general promoter classes SP and BR discovered by CAGE are of biological significance they have been mainly classified in an ad hoc manner. New methods are now emerging using unsupervised clustering showing that the old classes can be recreated in an unbiased manner (Zhao, X. et al. unpublished) and that further subdivision can be achieved correlating well with independent sequence features.

CAGE is also likely to scale with new sequencing technologies. As we demonstrated for some tissues we are almost there on the gene and TC level, but still far from saturation when considering TSSs. Recently we and others released the main paper of Functional Annotation of the Mouse Part 4 (FANTOM4) [38]. This pairs CAGE with the Illumina Solexa sequencer giving unprecedented depth across several time-points of THP-1. The Solexa produces orders of magnitude more reads than the 454, promising a new era of promoter discovery. The use of multiple time points monitors the dynamic nature of CAGE tags which was in turn used in combination with motif discovery to identify key regulators. It will be interesting to apply our method to this data set.

In the future, CAGE is likely to be used increasingly in functional and disease-related studies. For instance, while microarrays have long been showing that oncogenes may be misregulated in cancer only CAGE or similar protocols may uncover whether a particular promoter is to blame. This is particularly important now that we know that most genes are under control of multiple promoters. As the in situ images in paper III hinted at, rarely sampled tags might actually belong to highly expressed genes that are expressed in a minority of cells which will be important to map. This is particularly so if the cells are on the path to becoming cancerous. The problem is also attacked from another angle in that CAGE is now in the preliminary stages of being applied to a tiny number of cells. This is expected to culminate in single-cell CAGE (Piero Carninci personal communication). This will be of immense interest as sampling from multiple cells will tend to average out any dynamics the individual cell is experiencing.

Lately, much focus has been on non-coding transcripts (ncRNA) which appear to be abundant, but have been much more elusive than their protein-coding counterparts [86, 87]. Protein-coding genes, by their nature, are easier to detect both because of strong sequence biases, long open reading frames and the fact that they produce a protein product which can also be detected. They are also seemingly more sensitive to evolutionary change and are in many cases strongly conserved on the amino acid level which can be detected by synonymous/non-synonymous mutation rates. Non-coding genes on the other hand seem to be more free to diverge and some are only conserved on the structural level. CAGE can be a great help in detecting or confirming these and has been used in recent studies as additional evidence [87]. With the ever expanding pantheon of ncRNAs being discovered CAGE may provide a valuable tool in uncover-
ing new classes using the intergenic promoters as a guide.

All in all, high-throughput sequencing is launching transcriptomics to new heights providing us with a more unbiased view than ever before of the transcriptional complexity of the cell.
Acknowledgement

Thanks to the following people who have helped me and influenced me over the last three years:

Albin Sandelin for showing me that biology can be even more fun than computer science. For introducing me to CAGE and always including me in his latest cool project. You were never too busy for a discussion and I could not have asked for a better supervisor.

Anders Krogh for introducing me to bioinformatics, hiring me and going out of his way to help me in all my endeavours, even when planning to leave. Without you, I would not be were I am today.

Ole Winther for guiding me through the wilderness of machine learning and never laughing at my stupid questions.

Brian Parker for critically reading the thesis and correcting my horribly mangled English. Troels Marstrand for great discussions, both drunken and otherwise. Hanne Munkholm for her mastery of the ancient art of Linux-foo. Sanne Nygaard for introducing me to ever new ways of conquering the world. Stinus Lindgreen for infinite supplies of beer without divine additives. And to the rest of Binf, I will miss you all and I hope to return to you in a year.

Carsten Daub and Piero Carninci for welcoming me to RIKEN and Japan. You introduced me to working at a bigger lab and involved me in your projects. Also thanks to the other people I met at RIKEN, in particular Joost Boele, Sylvia Victor, Morana Vitezic, Marina Lizio (M&M) and the pink bunnies. Japan would have been much less fun without you.

To my family for support and Eirik, my brother, for taking care of Christmas presents this year (again). Finally, my girlfriend Pernille for her support, patience and companionship. Without you I would be lost.
Bibliography


Paper I
Discovery of Regulatory Elements is Improved by a Discriminatory Approach

Eivind Valen¹*, Albin Sandelin¹, Ole Winther¹,², Anders Krogh¹

¹The Bioinformatics Centre, Department of Biology and the Biotech Research and Innovation Centre (BRIC), University of Copenhagen, Copenhagen, Denmark, ²DTU Informatics, Technical University of Denmark, Lyngby, Denmark

Abstract

A major goal in post-genome biology is the complete mapping of the gene regulatory networks for every organism. Identification of regulatory elements is a prerequisite for realizing this ambitious goal. A common problem is finding regulatory patterns in promoters of a group of co-expressed genes, but contemporary methods are challenged by the size and diversity of regulatory regions in higher metazoans. Two key issues are the small amount of information contained in a pattern compared to the large promoter regions and the repetitive characteristics of genomic DNA, which both lead to “pattern drowning”. We present a new computational method for identifying transcription factor binding sites in promoters using a discriminatory approach with a large negative set encompassing a significant sample of the promoters from the relevant genome. The sequences are described by a probabilistic model and the most discriminatory motifs are identified by maximizing the probability of the sets given the motif model and prior probabilities of motif occurrences in both sets. Due to the large number of promoters in the negative set, an enhanced suffix array is used to improve speed and performance. Using our method, we demonstrate higher accuracy than the best of contemporary methods, high robustness when extending the length of the input sequences and a strong correlation between our objective function and the correct solution. Using a large background set of real promoters instead of a simplified model leads to higher discriminatory power and markedly reduces the need for repeat masking; a common pre-processing step for other pattern finders.

Introduction

The rapid emergence of experimental techniques that can probe for functional elements at whole-genome scales[1] necessitates computational methods to analyze data in these settings. In particular, methods that locate promoters or measure gene expression on genome-wide scales (e.g. [2,3]) must be complemented by algorithms that can find the active regulatory elements within the larger promoters. Ab initio computational search for transcription factor binding sites (TFBS) in DNA sequences is often termed “motif discovery”. “Motif” here refers to a general pattern describing what DNA sequences the transcription factor binds[4]. Motif discovery is one of the classical problems in computational sequence analysis and can be briefly stated as: Given a set of sequences containing one or several short overrepresented sites, locate these and produce a model describing them.

There are two main avenues used to attack this problem: i) enumerative algorithms based on word counting, such as [5,6], and ii) pattern-based approaches often using position specific weight matrices (WMs), which scores sites based on position specific weights [4]. Since the binding preferences of transcription factors (TFs) are not easily captured by a single word or consensus string, pattern-based approaches can give solutions closer to the biological reality and it has been argued that the matrix score is related to the binding energy [7,8]. However, such approaches correspond to the problem of finding local, optimal multiple alignments, which is NP-complete [9]. Therefore, almost all pattern-based motif finders use statistical optimization methods such as Gibbs sampling or expectation maximization [10,11].

A typical instance of motif discovery starts with a set of upstream promoter regions of co-expressed genes suspected to be co-regulated and by extension more likely to be under control by the same regulatory machinery. This set is called the “positive set” and most methods proceed from here by locating motifs that are in some way statistically overrepresented in this set. The most successful applications of motif discovery have been in organisms whose regulatory information is densely aggregated around transcription start sites, such as Saccharomyces cerevisiae (baker’s yeast). In mammalian genomes, regulatory information is spread out over wider regions, which makes “pattern drowning” a significant issue; in other words, the information in the regulatory sites is too small to stand out in the large genomic region of interest. In this context, the accuracy of contemporary pattern finders is not sufficient for many biologically important problems [12].

Most methods operate with some notion of a background model describing “generic DNA” against which the over-representation is measured. The model is often a multinomial or a Markov model. The choice of model is important for obtaining good results [13,14]. However, most such models have difficulty in capturing the complexity of the highly heterogeneous mammalian genome.
In the years following the sequencing of the human genome focus have shifted towards trying to understand how this blueprint results in the diversity of cells that we observe. Part of the answer lies in the regulation of transcription and how the proteins responsible for this recognize where they should attach to the DNA. This is a well studied problem, but most methods developed for this have a hard time dealing with the heterogeneity of the mammalian genomes. Here we present a method that greatly improves the efficiency of this search by contrasting the DNA with a large number of background DNA sequences. This enables us to handle repetitive segments of the genome that may be functional, but are usually considered intractable by most methods.

sequence, which has a multitude of different promoter architectures[15], numerous interspersed repeats, low complexity sequences, CpG islands, etc. [16]. Instead of simplifying the underlying DNA sequence by a general model, we take this to its extreme conclusion and use a very large set of promoters as the actual background instead of building a model describing the sequences in the promoters. For simplicity, we use the term “negative set” to describe the background set; this is strictly speaking not true as sites could occur in this set at a much lower frequency, since real promoters are sampled randomly. By contrasting the sets, it is possible to see what common features make the sequences in the positive set unique.

Discriminatory motif searching is not a new idea; several methods have been developed that take advantage of a negative set [17–24]. However, many of these use word-based models [19–21], which might not capture the diversity of binding sites. Others again use PWMs, but have binary hit models that do not distinguish between hits as long as they are over a threshold [22].

A discriminatory approach similar to ours has been combined with the use of expression data [18], but depending on the regions that are being investigated this might often not be available or even possible. We adopt an approach similar to DEME [23] to identify the most discriminative set of motifs by modeling the sequence labels (positive or negative) rather than using the conventional generative approach[10,11]. However, there are some important differences to DEME. Firstly, DEME uses a global string-based search followed by a local gradient refinement, which may miss patterns that are not well-represented by a consensus string, whereas we use a global optimization technique (simulated annealing) for optimizing the model, which does not have this limitation, although it may have others (see below). Secondly, our method (Motif Annealer - MoAn) uses and optimizes a threshold, and uses an enhanced suffix array (ESA) to speed up pattern searches. Thirdly, in MoAn the length of the motif is also optimized. DEME is also particularly targeted towards proteins while our approach is intended for use with DNA.

Specifically, we use conditional maximum likelihood to estimate the WM and their thresholds such that the probability of the positive and negative sets is maximized (see Methods). Thus, the resulting matrices cannot be derived from the frequency matrix for the sites found – it is rather the matrices that lead to the best discrimination. The probability of a sequence is calculated as a product of the probabilities given by the matrices matching above a threshold and a simple null model for non-matching regions. From this and prior probabilities for matches in the positive and negative sets, the probability of the set label (positive or negative) is calculated. In this probability the background model cancels. The total likelihood is a product of the class probabilities for all sequences (positive and negative).

This conditional likelihood leads to a non-trivial optimization problem which is handled using simulated annealing (see Methods), where we iteratively change the WMs and their thresholds, retaining changes that lead to higher discriminatory power using the Metropolis-Hastings algorithm [25,26].

Given sufficient iterations, the method guarantees convergence on the optimally discriminatory motifs. To cope with the vast size of the sets we utilize a highly efficient data structure, the ESA, for searching DNA for pattern instances[27]. With reasonable cutoffs, this reduces the computation by an order of magnitude[28].

Results

We evaluated our method by comparing its accuracy to a set of widely used motif discovery methods (MEME[29], DEME[23], Weeder[3] and Nest-MICA[14] in several different ways. In all runs, we used the same background set, which consists of 1000 experimentally defined promoters randomly sampled from the mouse genome (Text S1). The evaluation statistics are the same as used in [12] (see Methods) and we also pooled the results from all motifs (grouped by length of the input sequence; see below) and calculated the compound statistics on this. To reduce the influence of the optimization method, we ran all non-deterministic methods five times on each set selecting the best run according to their own scoring function.

In line with the recommendations of [12] we used synthetic data sets for the inter-method comparison. These were constructed by taking experimentally defined promoter regions based on strong CAGE tag clusters [2] and planting binding sites from various TFs inside these (Text S1). To decrease possible biases for the methods towards certain specific motif types, we randomly selected one TF from each of the 11 JASPAR[30] families as well as an example of a zinc-finger factor (Table S1). For a given matrix, we randomly chose sites from experimentally validated binding sequences used for constructing the JASPAR matrix instead of generating sites using the matrix. Since the accuracy of motif discovery methods normally deteriorates when sequence length is increased (“pattern drowning”), we evaluated the various methods on sets with sequence lengths varying between 200 and 1200 nucleotides (Table S3). This gave a total of 84 sets (12 motifs ×7 lengths) with 100 sequences in each. Sequences had a site from a given motif planted with a probability of 0.5. For those methods that support it, a background/negative set was provided containing 1000 sequences sampled in the same way and with the same length as the positive sequences. We used default settings for all methods except where there were obvious reasons not to (Text S2). Since DEME requires motif length as input we decided to input the correct length of the matrix. This provides DEME with an informational advantage over the other methods.

Fig. 1 (and Figs. S4, S5, S6, S7, S8) shows a significant performance gain in using MoAn compared to the other methods as measured by Matthews correlation coefficient on nucleotide level (nCC) and average site performance (ASP) – an average over the positive predictive value and the sensitivity on binding site level (see Methods for details). With both measures, MoAn performs better than any other method on all sequence lengths. In particular, the performance is not as affected by increasing the input sequence length as the other methods; at certain sequence lengths(300, 1200) MoAn has more than twice as high ASP values as the second best method. We also evaluated MoAn with the applicable subset of the evaluation set proposed by [12] (Text S3 and Table S4), where the OligoDyad, AnnSpec and MoAn...
achieve the highest sASP values. We note that this set is challenging as none of the methods perform well overall, and the difference in performance between methods might not be significant due to this fact. In addition, this set does not evaluate how well the method can deal with increasing lengths of input sequences, which is highly relevant.

Correlation of score and solution

The relationship between our objective function and the correct solution was assessed by plotting the MoAn scores against the sensitivity obtained in all five runs on each of the 84 sets (not just the best from each run) (Fig. 2). There is a clear correlation (Pearson CC: 0.90) between these two measures. There is a similar correlation with other measures, such as the nCC (Fig. S1).

This finding is important, because it indicates that the raw score is an indication of quality independent of the motif analyzed. It also shows that choosing the best scoring run of several will often give the best result.

Repetitive sequences

Aside from the problem with decreasing sensitivity as the length of the input sequences increase, repetitive sequences represent a severe problem for motif discovery, as these will often seem to be over-represented, and therefore it is common to mask these repeats. However, masking is always arbitrary, and some repeats are functional [31,32], so indiscriminate repeat masking is not optimal. When using a large negative set, repeat masking is unnecessary since repeats, if commonly occurring, will feature in the negative set and therefore be avoided as potential hits in the positive. At the same time, we can avoid the reverse problem – if a type of repeat actually is over-represented in the positive set, it can still be found. To demonstrate the insensitivity to repeats on a practical level, we planted repetitive sequences in each of the positive sets with a slightly higher frequency than the real motifs and ran our predictor on these sets both with the normal background and with a background similarly spiked with repeats. Specifically, we planted 1 to 10 consecutive instances of CACTA with a probability of 60% in each sequence. Fig. 3 shows, as expected, that the results do not deviate much from the repeat-less run when repeats are planted in both the positive and negative sequences, while the method picks up the repeats instead when there are no repeats in the negative set. We also performed this test using decoy motifs instead of repeats with similar results (Text S4, Fig. S2).

Real data

Evaluation of methods on real data is difficult and often a poor indication of general performance due to lack of insight into the
correct solution [12]; on the other hand, it is necessary to show that the method can be applied to real problems.

MoAn and four other methods were run on a collection of real data sets consisting of the binding sites of four human and mouse factors from the PAZAR database[33] and their associated genomic sequence. The sets were split by organism into 7 sets and the regions adjacent on the genome were merged resulting in sets ranging in size from 14 to 118. The merging means that the base sequences can have a varying number of sites and may be of different lengths. The sets were then subsequently enlarged by adding an equal number of randomly selected promoters to increase the difficulty (Text S6 and Table S5) and also padded with their cognate upstream and downstream regions of varying lengths (200–1200, as in the synthetic evaluation) to estimate the impact of noise.

Fig. 4 shows the performance over the real sets. MoAn’s performance is clearly superior, but not as spectacular as in the more controlled environment with synthetic sequences. We speculate that the reason for this is that the background and foreground of the synthetic sets are essentially sampled from the same pool (RefSeq promoters), while we have made no effort to customize the background for the PAZAR sets. If the genomic environment of the factors differ from normal promoter sequences this could lead to a reduced performance. There are also fewer sets (7 versus 12) in this evaluation leading to a higher variability.

We report additional trials using ChIP-chip data in supplementary material (Text S7, Fig. S3 and Tables S6, S7). MoAn has also been used successfully to discriminate between binding regions of human ESR1 and its paralog ESR2; the results were comparable with matrix-scanning approaches with pre-defined motifs[34].

Co-occurrence of binding sites

An additional aspect of the motif finding problem is that TFs often work by forming complex interactions [35]. Examples include mutually exclusive and cooperative binding. Clusters of TFBSs are commonly termed cis-regulatory modules, and are often responsible for tissue-specific expression. We try to capture these interactions by incorporating co-occurrence of sites from different motifs into our model, with the goal of further increasing predictive power. To test whether our objective function is capable of capturing interactions between factors we constructed a set where co-occurrence of sites from different motifs occurs. We randomly chose 5 pairs of new motifs (Table S2) and planted their corresponding sites in a positive set of 100 promoters with a 40% chance of co-occurrence and 10% of single occurrence. We then spiked the background set with sites from each of the motifs (10% chance each for all sequences) to mimic a situation where it is the interactions of the two sites rather than single sites that are responsible for the regulation. MoAn was then run in co-occurrence mode and compared to two single-occurrence runs in a series. In the serial runs we masked out the predictions from the first iteration before running the second iteration. In Fig. 5 the ASP and nCC is plotted. In our experiment three of the pairs turned out to be composed of motifs with relatively low information, leading to poor performance. However, the two remaining ones show that modeling of co-occurrence can significantly improve performance. This extended model is unfortunately computationally taxing and requires more than twice the number of iterations compared to the single prediction.

Discussion

In this work we have shown the value of using a large negative set instead of a pre-defined background model in motif discovery. Using raw sequences more accurately portrays the background than any general model and therefore higher discriminatory power is achieved. This method is also much less sensitive to “pattern drowning” in larger sequences, which is a bottleneck in computational analysis of mammalian regulatory regions. However, while our method takes a significant step towards routine motif discovery on large sequences, the problem cannot be considered fully solved. In particular, MoAn accuracy may be further improved by
incorporating information on evolutionary constraints (phylogenetic footprinting) or DNA accessibility. MoAn seems to be better at balancing the sensitivity and specificity. On the other hand, DEME is also given an artificial advantage by having the correct motif length as input and it is uncertain how advantageous this is. Weeder performed surprisingly poorly given its stellar performance in a recent evaluation. This might be due to motif selection which we did according to the most redundant motif, but was in [12] done in a more complicated manner not part of the current Weeder package. This procedure led to no predictions on several of the harder sets which might give Weeder a statistical advantage (as discussed in [12]).

A concern that might be raised is that optimizing a cutoff might lead to a conservative estimate of binding sites at the expense of weaker sites. However, assessing this is hard since experiments have their own thresholds in the post-analysis and any evaluation of MoAn’s threshold will be dependant upon those. Investigations where we artificially forced the cutoff to remain low, lead to a reduction in performance (data not shown). We address this potential problem indirectly by providing a matrix that can be used to search sequences at a lower threshold.

Future improvements of MoAn will focus on the optimization algorithm, which currently is not robust enough to always produce reliable results. In our current implementation we avoid this problem by running the algorithm many times to see that the solution is stable.

Methods

Evaluation is done on both site and nucleotide levels. The statistics used are similar to those in the recent large scale evaluation. To get a compound statistic for all motifs at each length we used what is there described as the “combined” method for summarizing. This consists of treating all sets of a given length as one big set, summing up all the basic statistics below (nTP, nTN … nFP) before calculating the compound statistics. This removes the problem of undefined statistics in those cases where a method does not predict any sites.

Basic statistics

- nTP Number of nts of a site correctly predicted.
- nFP Number of background nts predicted to be part of a motif.
- nFN Number of nts of a site predicted as background.
- sTP Number of real sites that share over 50% of its nts with a predicted site.
- sFP Number of predicted sites that share less than 50% of its nts with a real site.
- sFN Number of real sites that share less than 50% of its nts with a predicted site.

Note that we are more conservative with respect to the site prediction than [12] in that we demand at least half of the nucleotides overlapped to get a single sTP.

Compound statistics

Derived from the basic statistics:

\[ s\text{Sn} = \frac{s\text{TP}}{s\text{TP} + s\text{FP}} \]

\[ s\text{PPV} = \frac{s\text{TP}}{s\text{TP} + s\text{FP}} \]

\[ s\text{ASP} = \frac{s\text{Sn} + s\text{PPV}}{2} \]

\[ n\text{CC} = \frac{n\text{TP} \times n\text{TN} - n\text{FP} \times n\text{FN}}{\sqrt{(n\text{TP} + n\text{FN})(n\text{TN} + n\text{FP})(n\text{TP} + n\text{FP})(n\text{TN} + n\text{FN})}} \]

Objective function

A sequence \( x \) is assumed to be described by a mixture model consisting of a background distribution \( q \) and a set of WMs \( W \) describing the binding affinities of the TFs. The WMs contain log-odds scores of the type:

\[ w_{i,b} = \log_2 \frac{P(b|i)}{q(b)} \] (1)

where \( i \) is the position in the WM, \( b \) is a letter in the DNA alphabet and \( P(b|i) \) is the probability of having letter \( b \) at position \( i \) in the motif described by \( w \). The score of a matrix \( w \) aligned at a position \( a \) in a sequence \( x \) is therefore:

\[ S(a,x,w) = \sum_{i=1}^{\left|\{a\}_{\leq a}\right|} w_{i,b+i-1} \] (2)

where \( x(i) \) is the DNA letter at position \( i \) in sequence \( x \).

The aim is to discriminate between two sets of sequences \( x^0 = \{ x_1^0, x_2^0, \ldots, x_N^0 \} \), where label \( y = 1 \) denotes the positive set and \( y = 0 \) the negative. The prior probability of binding site occurrence in a sequence contained in set \( y \) is called \( v^y \). We assume that there is a marked difference in the site occurrence between the two sets and want to construct a score that captures how well a set of WMs describe this difference. Using two WMs as an example, \( v^1 \) and \( v^0 \), there are four possible ways for a sequence \( x \) to be generated. With prior probability \( v^0 \) it contains no sites and is only generated by the background model \( q \). Or, with prior probability \( v^1 \), it contains a single site (one of the two) corresponding to one WM \( w^k \) positioned at nucleotide number \( a_k \) (\( k \) is equal to 1 or 2 corresponding to the two different matrices). This is written \( q(x) \gamma^k(\{a_k,x,a_k^k\}) \), where \( S(a_k,x,a_k^k) \) is the score of the matrix aligned to the nucleotides at position \( a_k \) (eq. 2) and \( 2 \) is the base of the log scores contained in the WM. Note that the log scores in a WM are divided by the background model, so the background \( (q) \) cancels out in sites where the motif occurs. The final case, with prior probability \( v^2 \), is the co-occurrence of two sites in a sequence, which is \( q(x) \gamma^2(\{a_k,x,a_k^k\}) \gamma^2(\{a_k,x,a_k^k\}) \). However, this is only correct when the sites are not overlapping since otherwise the overlapping nucleotides would be included in the product twice. Therefore we disallow overlaps.

For efficiency reasons, we do not calculate the score in its entirety. We assume that it is the strong sites that contribute the most to the equation and introduce a cutoff for each WM on the minimum score of a site. This enables an efficient search in the ESA. This is not without biological merit since WM scores and binding energies for known TFs are correlated, and at some point the binding energies of a TF and a poor binding sequence must be too small to
Discriminatory Discovery of Regulatory Elements

matter [4]. It is also a standard method to use when scanning with known matrices [38]. So we only consider sites that score above a threshold, which is called \( c_\text{th} \) for matrix \( X \). Then the probability of a sequence \( x \) from the set \( y \) being generated by the WM's is

\[
P(x|y,W,c) = q(x) \left[ v_{0,x}^y + v_{1,x}^y r(x,w^1,c_1) + v_{2,x}^y r(x,w^2,c_2) + v_{3,x}^y r(x,w^3,c_3) \right]
\]

where \( r(x,w,c) \) is the expectation over \( a \) of \( \theta(S(a,x,w) - c)2^a \) over all predicted sites:

\[
r(x,w,c) = \frac{1}{N} \sum_a \theta(S(a,x,w) - c)2^a
\]

with \( \theta(\cdot) \) being the step function (1 above 0 and zero otherwise). The co-occurrence expectation \( r(x,w^1,w^2,c_1,c_2) \) is defined in a similar way with overlaps disallowed. The effective weight of no sites

\[
v_{0,x}^y = v_0 + \frac{N-N_i}{N} v_1 + \frac{N-N_i}{N} v_2 + \frac{N^2-N_iN_2}{N^2} v_{12}
\]

accounts for extra weight given to no sites due to alignments not meeting the threshold. With this definition, \( q(x)r(x,w^1,c_1)/N_i \) is the probability or generative model of the sequence conditioned on the WM and threshold, \( p(x|w^1,c_1) \).

To find the WM's that best explain the difference in occurrence between the sets we use a discriminative objective function based on the probability of the labels \( Y \) given the sequences \( X \) and WM's, formally:

\[
\mathcal{L}(W,c;X,Y) = P(Y|x,W,c).
\]

This is the logistic likelihood function for binary classification, see e.g. [39]. The discriminative model can thus be viewed as logistic regression with an adaptive set of basis functions. For multiple sequences assumed to be independent, the joint probability is the product of the single sequence probabilities over all sequences in both the positive and negative set:

\[
\log_2 \mathcal{L}(W,c;X,Y) = \sum_{(x,y)\in(T,Y)} \log_2 P(y|W,x)
\]

We refer to this function as the (log likelihood) score, \( S_c \).

Based on the sequence density \( P(y|x,W,c) \) we can use Bayes theorem to calculate the probability of the label \( y \) given the WM's \( W \), the thresholds \( c \), and the sequence \( x \):

\[
P(y|x,W,c) = \frac{P(x,y|W,c)}{P(x|W,c)} = \frac{P(x,y|W,c)P(y)}{\sum_y P(x,y|W,c)P(y)}.
\]

We observe that the prior probability of \( y \) is proportional to the number of sequences in the set divided by the total number of sequences \( P(y) = \frac{|X|}{|x|} \).

A very high threshold will give no matches, and the probability will then be a constant given by the priors and the size of the two sets. Matches that score above the threshold in the negative set will lower the score and matches above the threshold in the positive set will increase the score, so the game is to obtain as many high-scoring matches in the positive set as possible without introducing too many matches in the negative set.

The prior is conservative i.e., in our runs in that we are strict about promoting hits in the positive set, but only moderately strict about disallowing negative hits. For a single matrix the prior on \( v_k \) is \( 0.01; v_1: 0.99; v_{12}: 0.80; v_2: 0.20. \) For two matrices: \( v_1: 0; v_{12}: 0.1; v_2: 0.9; v_{12}: 0.80; v_1: 0.15; \) and \( v_2: 0.05. \) These priors can be set by the user if prior knowledge is available about the set (i.e. a high confidence negative set or an uncertain positive set).

In the evaluation we deliberately chose a probability of having a site (0.5) in a sequence very different from the model prior (0.99) to avoid giving our own method a big advantage. It shows that the method is not very sensitive to the choice of prior.

Optimization

The objective function outlined above is optimized using simulated annealing [40]. Informally, it proceeds by iteratively proposing a candidate solution and then accepting or rejecting it depending on how good it is compared to the current solution. It sometimes accepts changes for the worse and therefore possesses the power to escape local maxima. The hope is that it will converge on a solution that is close to optimal. Formally, this translates to a walk over the search space \( E \) where in the current state \( E_i \) the next state \( E_{i+1} \) is either the same or the candidate solution \( \xi \) depending on their relative scores and a temperature parameter \( t_i \).

\[
E_{i+1} = \begin{cases} \xi & \text{with probability } p = \min \left( 1, \exp \left( \frac{S_c(E_i) - S_c(\xi)}{t_i} \right) \right) \\ E_i & \text{with probability } 1 - p \end{cases}
\]

The temperature parameter is lowered for each iteration using as default an exponential cooling scheme (for details see Text S5), thus incrementally constraining the neighborhood of accepted changes.

Candidate solutions are proposed by applying one of several steps outlined in the list below. In the case of multiple matrices, only one is changed at a time. We perform all steps on a integer "count" matrix which is then translated into a log-odds WM prior to searching the ESA, but notice that the "count" matrix does not represent actual letter frequencies in the selected sites. The steps are:

- Alter the contents of the WM columns by moving counts from one random cell to another within a column. The number of counts moved is selected uniformly from 1 to the current count number for the cell.
- Extend the WM in either direction. A uniformly sampled number of columns (1 to 5) is added and counts of these are decided by consulting the sequence locations of hits scoring above \( c \). The counts are proportional to the counts in the columns from the extended hits, but normalized so that all columns have the same counts.
- Decrease the length of the WM by deleting columns. Similarly to adding columns a uniformly selected number between 1 and 5 columns are deleted.
- Slide the WM across the sequences. Columns are deleted on one site and extended on the other according to the two steps above.
- Alter the cutoff \( c_k \) of the matrix \( k \). The cutoff is expressed in bits per column and a new candidate \( c_k \) is proposed by sampling uniformly from 0.6 to 2 bits.

- Extend the WM in either direction. A uniformly sampled number of columns (1 to 5) is added and counts of these are decided by consulting the sequence locations of hits scoring above \( c \). The counts are proportional to the counts in the columns from the extended hits, but normalized so that all columns have the same counts.
- Decrease the length of the WM by deleting columns. Similarly to adding columns a uniformly selected number between 1 and 5 columns are deleted.
- Slide the WM across the sequences. Columns are deleted on one site and extended on the other according to the two steps above.
- Alter the cutoff \( c_k \) of the matrix \( k \). The cutoff is expressed in bits per column and a new candidate \( c_k \) is proposed by sampling uniformly from 0.6 to 2 bits.
Note that for the extend and decrease step there is a minimum and maximum number of columns for a motif. The default for these are 5 and 15 respectively.

The matrix is initialized with random counts and the cutoff is also selected uniformly according to the last step in the list above. Termination of the optimization is only based on the number of iterations which is by default set to a rather conservative value of 30 million iterations. Time requirements for a single run is variable depending on the set size, but was for our runs comparable to NestedMICA (single threaded) and considerably faster than Weeder’s “large” run and DEME.

Availability
Source code as well as data sets is freely available at the author’s web site: http://moan.binf.ku.dk

Supporting Information

Figure S1 Correlation of MoAn’s objective function (Sc) and nucleotide correlation coefficient (nCC)
Found at: doi:10.1371/journal.pcbi.1000562.s001 (0.01 MB EPS)

Figure S2 Evaluation with decoy motifs. Average site performance (lines) and the nucleotide correlation coefficient (bars) of MoAn with decoy motifs planted in the two sets.
Found at: doi:10.1371/journal.pcbi.1000562.s002 (0.01 MB EPS)

Figure S3 Discriminatory power of matrices. ROC curve showing discriminatory power of matrices produced by MoAn and NestedMICA on the ESR1 data set. The line extends from the highest cutoff possible for that matrix (bottom right) to a cutoff of 0 (top left).
Found at: doi:10.1371/journal.pcbi.1000562.s003 (0.03 MB EPS)

Figure S4 Performance on individual sets for MoAn. The average site performance (lines) and the nucleotide correlation coefficient (bars) on the sets.
Found at: doi:10.1371/journal.pcbi.1000562.s004 (0.02 MB EPS)

Figure S5 Performance on individual sets for DEME. The average site performance (lines) and the nucleotide correlation coefficient (bars) on the sets.
Found at: doi:10.1371/journal.pcbi.1000562.s005 (0.02 MB EPS)

Figure S6 Performance on individual sets for MEME. The average site performance (lines) and the nucleotide correlation coefficient (bars) on the sets.
Found at: doi:10.1371/journal.pcbi.1000562.s006 (0.02 MB EPS)

Figure S7 Performance on individual sets for Weeder. The average site performance (lines) and the nucleotide correlation coefficient (bars) on the sets.
Found at: doi:10.1371/journal.pcbi.1000562.s007 (0.02 MB EPS)

Figure S8 Performance on individual sets for NestedMICA. The average site performance (lines) and the nucleotide correlation coefficient (bars) on the sets.
Found at: doi:10.1371/journal.pcbi.1000562.s008 (0.02 MB EPS)

Text S1 Data set construction
Found at: doi:10.1371/journal.pcbi.1000562.s009 (0.03 MB PDF)

Text S2 Running parameters
Found at: doi:10.1371/journal.pcbi.1000562.s010 (0.03 MB PDF)

Text S3 Tompa assessment
Found at: doi:10.1371/journal.pcbi.1000562.s011 (0.03 MB PDF)

Text S4 Sequences spiked with decoy motifs
Found at: doi:10.1371/journal.pcbi.1000562.s012 (0.02 MB PDF)

Text S5 Annealing schedule
Found at: doi:10.1371/journal.pcbi.1000562.s013 (0.03 MB PDF)

Text S6 PAZAR data sets
Found at: doi:10.1371/journal.pcbi.1000562.s014 (0.03 MB PDF)

Text S7 ChIP-chip data sets
Found at: doi:10.1371/journal.pcbi.1000562.s015 (0.04 MB PDF)

Table S1 Length of upstream and downstream extensions
Found at: doi:10.1371/journal.pcbi.1000562.s016 (0.01 MB PDF)

Table S2 Motifs planted in single occurrence sets
Found at: doi:10.1371/journal.pcbi.1000562.s017 (0.04 MB PDF)

Table S3 Motifs planted in co-occurrence sets
Found at: doi:10.1371/journal.pcbi.1000562.s018 (0.03 MB PDF)

Table S4 Results on the mammalian subset of the Tompa assessment
Found at: doi:10.1371/journal.pcbi.1000562.s019 (0.01 MB PDF)

Table S5 Sizes of PAZAR data sets
Found at: doi:10.1371/journal.pcbi.1000562.s020 (0.01 MB PDF)

Table S6 Sizes of ENCODE data sets
Found at: doi:10.1371/journal.pcbi.1000562.s021 (0.01 MB PDF)

Table S7 Performance on ENCODE data sets
Found at: doi:10.1371/journal.pcbi.1000562.s022 (0.07 MB PDF)

Acknowledgments
Thanks to Brian Parker for help with the manuscript.

Author Contributions
Conceived and designed the experiments: EV AS OW AK. Performed the experiments: EV. Analyzed the data: EV. Wrote the paper: EV AS OW AK.

References

Paper II
JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update

Jan Christian Bryne¹, Eivind Valen², Man-Hung Eric Tang², Troels Marstrand², Ole Winther²,³, Isabelle da Piedade⁴, Anders Krogh², Boris Lenhard¹,⁵,* and Albin Sandelin²,*

¹Computational Biology Unit, Bergen Center for Computational Science, University of Bergen, Thormøhlensgate 55, N-5008 Bergen, Norway, ²The Bioinformatics Centre, Department of Molecular Biology & Biotech Research and Innovation Centre, University of Copenhagen, Ole Maaløes Vej 5, DK-2100 København Ø, ³Informatics and Mathematical Modeling, Building 321, Technical University of Denmark, DK-2800 Kgs. Lyngby, ⁴Center for Comparative Genomics, Institute of Biology, University of Copenhagen, Universitetsparken 15, DK-2100 Copenhagen Ø, Denmark and ⁵Sars Centre for Marine Molecular Biology, University of Bergen, Thormøhlensgate 55, N-5008 Bergen, Norway

Received September 14, 2007; Revised October 15, 2007; Accepted October 16, 2007

ABSTRACT

JASPAR is a popular open-access database for matrix models describing DNA-binding preferences for transcription factors and other DNA patterns. With its third major release, JASPAR has been expanded and equipped with additional functions aimed at both casual and power users. The heart of the JASPAR database—the JASPAR CORE sub-database—has increased by 12% in size, and three new specialized sub-databases have been added. New functions include clustering of matrix models by similarity, generation of random matrices by sampling from selected sets of existing models and a language-independent Web Service applications programming interface for matrix retrieval. JASPAR is available at http://jaspar.genereg.net.

INTRODUCTION

Computational analysis of regulatory properties of DNA is most often based on the use of matrix models describing binding preferences of transcription factors, or other DNA patterns. Such matrices are based on sets of known or inferred sites for a DNA-binding protein, and can be scanned over genomic sequences to predict novel binding sites (1,2). JASPAR is the most comprehensive open-access database holding such models. The heart of JASPAR is the JASPAR CORE sub-database, holding curated, non-redundant matrix models from multi-cellular eukaryotes. The methodology for JASPAR CORE curation has been described previously (3). JASPAR CORE is now a standard resource in gene regulation bioinformatics and is used as a matrix set in a wide variety of other services [for instance (4–9)], and large-scale projects (10,11). Besides JASPAR CORE, the database contains several sub-databases (JASPAR Collections) holding matrix models produced by different methods and for different purposes (Table 1).

Here we present the recent JASPAR expansion, which includes a significant increase of the JASPAR CORE content and an addition of three new sub-databases focusing on core promoter patterns, splice sites and motifs detected in vertebrate highly conserved non-coding elements, respectively. In addition, we present several unique functional features in the web interface aimed at both casual and power users, including statistics on expected number of predictions each matrix will yield at several different thresholds in random sequences generated by three commonly encountered sequence background models, dynamic clustering of matrices by similarity and generation of random matrices using a selected set of matrices as background model.

*To whom correspondence should be addressed. Boris Lenhard: Tel: +47 55584362; Fax: +47 55584295; Email: boris.lenhard@bccs.uib.no
Albin Sandelin: Tel: +45 52321285; Fax: +45 35325669; Email: albin@bias.ku.dk

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

© 2007 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
RESULTS

Here we briefly describe the new data and functional features; more detailed descriptions are available at the documentation at the web site.

Expansion of JASPAR CORE

The JASPAR CORE database holds a curated set of transcription factor-binding profiles from multi-cellular eukaryotes: this is a unique feature with respect to databases of similar scope. We have extended JASPAR CORE with 15 new, high-quality profiles from recent experimental literature, increasing the total number of JASPAR CORE models to 138 (Table 1). In addition, annotation for all models in the database has been updated [e.g. to standard gene symbols from Entrez Gene (12)] and expanded. Prompted by user feedback, several existing matrices have been updated or corrected.

New sub-databases

Existing and new sub-databases within JASPAR and their specific features are described in Table 1. Since the last update, we have added three new sub-databases, which are briefly described below (see the web documentation for details):

**JASPAR POLII.** The large body of novel data pertaining transcription start sites (13,14) has triggered a new interest in computational studies of core promoters. The JASPAR POLII sub-database holds 13 known DNA patterns linked to RNA polymerase II core promoters, such as the Inr and BRE elements, each based on experimental evidence: each model must be constructed using five or more experimentally verified sites. An important difference to the transcription factor profiles in JASPAR CORE is that patterns here do not necessarily have a specified protein that binds them [See Ref. (15) for a review on core promoter patterns]. When possible, profiles were extended by 2 nt more than the core motif. We consistently report positions relative to the TSS as the position of 5' and 3' edge of the matrix.

**JASPAR CNE.** Highly conserved non-coding elements (CNEs) are a distinctive feature of metazoan genomes. Many of them can be shown to act as long-range enhancers that drive expression of genes that are themselves regulators of core aspects of metazoan development and differentiation. Since they act as regulatory inputs, attempts at deciphering the regulatory content of these elements have started (16–18). JASPAR CNE is a collection of 233 matrix profiles derived by Xie et al. (19) by clustering of overrepresented motifs from human conserved non-coding elements. While the biochemical and biological role of most of these patterns is still unknown, Xie et al. have shown that the most abundant ones correspond to known DNA-binding proteins, among them the insulator-binding protein CTCF. These matrix profiles will be useful for further characterization of regulatory inputs in long-range developmental gene regulation in vertebrates.

**JASPAR SPLICE.** This small collection contains matrix profiles of human canonical and non-canonical splice sites, as matching donor:acceptor pairs. It currently contains only six highly reliable profiles (two canonical and four non-canonical) obtained from human genome (20). In the future, we shall include additional eukaryotic species, as well as new models for exonic splicing enhancers (ESE) and inhibitors (ESI).

Extended functionality

In addition to data extension, we have implemented a number of functional improvements in the web interface of the JASPAR database. These range from static statistics, such as expected number of hits on typical DNA sequence for any factor, to dynamic tools for similarity-based profile clustering and for generating random profiles based on a subset of known profiles.

**Web service interface.** The JASPAR database can now be reached remotely through a new Web Service interface. Current functionality includes retrieval of profiles by name, by identifier and by searching profile annotations. The purpose of providing an external application programming interface (API) is to simplify the utilization of JASPAR in distributed applications and in scientific workflows created in workflow editors like Triana (21), BPEL (http://www.bpelsource.com/) or Taverna (22). Other benefits include platform- and language-independent access, as well as constant up-to-date access to

---

**Table 1. JASPAR databases**

<table>
<thead>
<tr>
<th>Database</th>
<th>Number of models</th>
<th>Scope</th>
<th>Species coverage</th>
<th>When to use</th>
</tr>
</thead>
<tbody>
<tr>
<td>JASPAR CORE</td>
<td>138</td>
<td>Curated, non-redundant matrix models</td>
<td>Multi-cellular eukaryotes</td>
<td>'Standard' promoter analysis</td>
</tr>
<tr>
<td>JASPAR FAM</td>
<td>11</td>
<td>Familial 'consensus' patterns for major structural families of transcription factors</td>
<td>Multi-cellular eukaryotes</td>
<td>Matrix-to-matrix comparison and classification, or as prior knowledge for pattern finders</td>
</tr>
<tr>
<td>JASPAR PHYLOFACTS</td>
<td>174</td>
<td>Evolutionary conserved patterns in 5' promoter regions</td>
<td>Multi-cellular eukaryotes</td>
<td>Core promoter analysis for large-scale studies</td>
</tr>
<tr>
<td>JASPAR POLII</td>
<td>13</td>
<td>Core promoter element models</td>
<td>Multi-cellular eukaryotes</td>
<td>Core promoter analysis</td>
</tr>
<tr>
<td>JASPAR CNE</td>
<td>233</td>
<td>Motifs overrepresented in vertebrate highly conserved non-coding elements</td>
<td>Human</td>
<td>Analysis of regulatory content of long-range enhancers</td>
</tr>
<tr>
<td>JASPAR SPLICE</td>
<td>6a</td>
<td>Splice sites</td>
<td>Humana</td>
<td>Splice site analysis</td>
</tr>
</tbody>
</table>

*aExpansion under way.

---
the database over time. The API is implemented as a
WS-I compliant Web Service, identical to the technology
used for the services made available through the
EMBRACE Network of Excellence (www.embracegrid.
info), and the Web Service technology chosen by the
European Bioinformatics Institute (EBI) (23). Its
basic usage is described in tutorials at the JASPAR web
site. The WSDL describing this service can be found at:
http://api.bioinfo.no/wSDL/JasparDB.wsdl. Further information
about the Web Service, including example clients in Java and Python, is available on the Jaspar
web site and in the WSDL file.

Expected predictions/base-pair statistics for all
models. An important problem with genome-wide scan-
ning with matrix models is the limited information content
in a typical matrix, resulting in numerous spurious hits
just due to sequence background (1,2). The number of
false positives varies considerably between factors and
also depends on what type of sequences that models are
applied to, user-defined cutoffs and to a more limited
extent on the type of scoring scheme used. For a first-
glance assessment of the rate of spurious predictions of a
given model, we apply the model to three distinct sequence
sets: known promoters from the EPD database (24), CpG
islands and randomly selected genomic DNA, respec-
tively. For different score thresholds, we plot the mean
number of hits per 1000 nt for each sequence set. The
resulting bar plots are available for each JASPAR matrix
(Figure 1).

Dynamic clustering by similarity and creation of familial
binding profiles from a given profile subset. Many
transcription factors bind similar targets and it is often
helpful to cluster similar binding profiles to generate
familial binding profiles—models describing a set of
matrices (25). Part of this problem is matrix profile
comparison and alignments, explored by several research-
ers (25–30). Recently, Mahony et al. (27,28) made a
comprehensive study on alignments of matrices and
construction of familial binding profiles, resulting in the
STAMP tool, which is now used within JASPAR to
cluster matrix models. Hierarchical clustering is per-
formed on a selected set of matrices using the UPGMA
algorithm with a Pearson Correlation Coefficient distance
metric. Then the optimal number of clusters is selected
using a log variant of the Calinski and Harabasz statistic
[See Ref. (27) for details]. Finally, the clusters are
partitioned and a familial binding profile is created for
each cluster using iterative refinement (a multiple align-
ment method). An example is shown in Figure 1.

Dynamic random profile generation. In many computa-
tional studies, it is helpful to have a set of ‘random’
matrices. This is particularly true for assessment of
distances between putative sites and reference points as
transcription start sites, and also for matrix-to-matrix
comparisons. In these cases, it is desired that the
randomized matrices should share properties with the
true matrix set—for instance having the same nucleotide
content and/or the same general information content.

Within any JASPAR sub-database, users can select a
subset of matrices, which will then be used to generate
random matrices using one of two methods:
(i) Permutations: Columns of the selected matrices are
shuffled: either constrained to shuffling of columns
within each matrix or between all selected matrices.
(ii) Probabilistic sampling: This enables the users to
generate random Position Frequency Matrices from
selected profiles. In our model, each random column is
sampled from a posterior distribution—a 4D Dirichlet
mixture distribution. The posterior distribution has
two contributions: a multinomial with counts of columns
selected as in (i), and a Dirichlet mixture prior trained
from all observed nucleotides in the JASPAR database.
We assume that column positions are independent.

DISCUSSION

We have presented a significant update to the JASPAR
database, including an expansion of the core database,
three new sub-databases and many new utilities. The new
web service interface enables easy interaction with
scientific workflows and an increasing number of pro-
gramming languages that support this technology.
We project that the new features, together with the
open-access policy, will further consolidate the JASPAR
database as a standard resource in the field of gene
regulation bioinformatics.

Towards a comprehensive set of models for
most known transcription factors

The lack of models for the binding specificity of most
transcription factors is a significant bottleneck for com-
prehensive computational analysis of genomes. Only a
fraction of transcription factors have been characterized in
enough detail to allow the construction of adequate
models of their binding specificity. This problem is being
solved in two principally different ways. First, tiling
array approaches for measuring binding preferences
en masse are being developed (31); these technologies
show great promise and are expected to make their mark
on the field in the near future. Second, a wealth of cis-
regulatory elements, characterized in painstaking detail, is
hidden in experimental literature; many of these sites are
not included in any database. There is a growing awareness
of this problem in the field, resulting in online open-access
databases such as ORegAnno (32) and PAZAR (33), where
one of the goals is to house expert-curated binding sites.
We are currently developing services to enable cross-talk
with these databases to enable matrix models built on
curated sites that exceed a certain quality threshold.
JASPAR, ORegAnno and PAZAR face the same chal-
lenge: to build models or sites, it is necessary to mine the
literature, which inevitably means that the curators will
miss many important studies. The only long-term solution
would be a requirement by scientific journals for research-
ers to deposit protein–DNA interactions in public data-
bases prior to publication, much in the same way as
mRNAs must be submitted to Genbank (34). Part of such
a system will be to establish a minimal standard for
reporting these interactions, much like the MIAME standard (35) for microarray data. As before, JASPAR team is always prepared to incorporate new matrices and matrix sets provided by external contributors.

Data availability

All the data in JASPAR are available without any restrictions, either from the web interface, as flat files or through the Web service interface.

ACKNOWLEDGEMENTS

Thanks to Katsuya Shigesada for pointing out errors in matrix MA0002, Shaun Mahony and Panayiotis V. Benos for generously sharing the STAMP code and general helpfulness and Vladimir B. Bajic for kindly providing the frequency matrices for JASPAR SPLICE. E.V., M.-H.E.T., T.M., O.W., A.K. and A.S. were supported by a grant from the Novo Nordisk foundation to the Bioinformatics Center. I.P. was supported by a grant from Carlsberg Foundation (21-00-0680). J.C.B. was supported by EMBRACE—an EU Sixth Framework Network of Excellence. B.L. was supported by the Functional Genomics Programme (FUGE) of the Research Council of Norway, and a core grant from the Sars Centre. Funding to pay the Open Access publication charges for this article was provided by a grant from the Novo Nordisk Foundation and the Functional Genomics Programme of the Research Council of Norway.
Conflict of interest statement. None declared.

REFERENCES

Paper III
Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE

Eivind Valen, Giovanni Pascarella, Alistair Chalk, et al.

Genome Res. 2009 19: 255-265 originally published online December 11, 2008
Access the most recent version at doi:10.1101/gr.084541.108

Supplemental Material
http://genome.cshlp.org/content/suppl/2009/01/14/gr.084541.108.DC1.html

References
This article cites 41 articles, 14 of which can be accessed free at:
http://genome.cshlp.org/content/19/2/255.full.html#ref-list-1

Article cited in:
http://genome.cshlp.org/content/19/2/255.full.html#related-urls

Open Access
Freely available online through the Genome Research Open Access option.

Email alerting service
Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here

To subscribe to Genome Research go to:
http://genome.cshlp.org/subscriptions

© 2009 American Meteorological Society
Methods

Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE

Eivind Valen,1 Giovanni Pascarella,2 Alistair Chalk,3 Norihiro Maeda,4 Miki Kojima,4 Chika Kawazu,4 Mitsuyoshi Murata,4 Hiromi Nishiyori,4 Dejan Lazarevic,2,8 Dario Motti,2 Troels Torben Marstrand,1 Man-Hung Eric Tang,1 Xiaobei Zhao,1 Anders Krogh,1 Ole Winther,1 Takahiro Arakawa,4 Jun Kawai,4 Christine Wells,3 Carsten Daub,5 Matthias Harbers,7 Yoshihide Hayashizaki,4 Stefano Gustincich,2 Albin Sandelin,1,9 and Piero Carninci4,6,9

1The Bioinformatics Centre, Department of Biology and Biotech Research and Innovation Centre, University of Copenhagen, Ole Maaløes vej 5, DK-2200, Denmark; 2The Giovanni Armenise-Harvard Foundation Laboratory, Sector of Neurobiology, International School for Advanced Studies (SISSA), Basovizza, 34012 Trieste, Italy; 3The National Centre for Adult Stem Cell Research, The Eskitis Institute for Cell and Molecular Therapies, Griffith University, Nathan QLD 4111, Australia; 4LSA Technology Development Group, Omics Science Center, RIKEN Yokohama Institute, Yokohama, Kanagawa 230-0045, Japan; 5LSA Bioinformatics Team, Omics Science Center, RIKEN Yokohama Institute, Yokohama, Kanagawa 230-0045, Japan; 6Functional Genomics Technology Team, Omics Science Center, RIKEN Yokohama Institute, Yokohama, Kanagawa 230-0045, Japan; 7DNAFORM Inc., Yokohama, Kanagawa 230-0046, Japan; 8CBM Scrl–Consorzio per il Centro di Biomedicina Molecolare, Basovizza, 34012 Trieste, Italy

Finding and characterizing mRNAs, their transcription start sites (TSS), and their associated promoters is a major focus in post-genome biology. Mammalian cells have at least 5–10 magnitudes more TSS than previously believed, and deeper sequencing is necessary to detect all active promoters in a given tissue. Here, we present a new method for high-throughput sequencing of 5' cDNA tags—DeepCAGE: merging the Cap Analysis of Gene Expression method with ultra-high-throughput sequence technology. We apply DeepCAGE to characterize 1.4 million sequenced TSS from mouse hippocampus and reveal a wealth of novel core promoters that are preferentially used in hippocampus: This is the most comprehensive promoter data set for any tissue to date. Using these data, we present evidence indicating a key role for the Arc/m2 transcription factor in hippocampus gene regulation. DeepCAGE can also detect promoters used only in a small subset of cells within the complex tissue.

[Supplemental material is available online at www.genome.org. CAGE tag sequences have been submitted to the DNA Data Bank of Japan (http://www.ddbj.nig.ac.jp/) under accession nos. AGAAA0000001–AGAAA0552486. Processed CAGE data sets are freely available at http://people.binf.ku.dk/albin/supplementary_data/hcmap/]

Transcription initiation is one of the most fundamental cellular processes. The identification of transcription start sites (TSS) leads to the detection of the associated core promoters. Historically, precise definition of TSS has been laborious and addressed one gene at a time. Therefore, few genes have had their start sites mapped in detail. We, and others, have presented techniques that can identify TSS on a genome-wide scale—typically, by generating full-length cDNAs and then sequencing short tags at their 5' ends (Ng et al. 2005; Kodzius et al. 2006). The largest study to date using such methods (Carninci et al. 2006) was carried out by taking advantage of the cap analysis of gene expression (CAGE) technology (Kodzius et al. 2006). In the FANTOM3 project, tag libraries were sequenced from 22 tissues with an average 48,500 tags per library in mouse. This study gave new insights into how transcription initiation works, as reviewed (Muller et al. 2007; Sandelin et al. 2007), and suggested that mammalian cells have many more core promoters than previously appreciated. In a previous study (Gustincich et al. 2006), we have also shown that mouse brain regions have a higher number of active TSS than other tissues, presumably leading to a higher diversity of distinct transcripts in these tissues. Since the average mammalian cell is estimated to express at least 350,000 mRNAs (Jackson et al. 2000) and the brain is a highly complex tissue with many distinct regions, which in themselves have high cellular heterogeneity, it is evident that the promoters sampled so far are just skimming the surface of the brain transcriptional complexity.

Among the different regions of the brain, the hippocampal formation (hippocampus, dentate gyrus, and subiculum) has been the subject of intense studies due to its essential role in the formation of new episodic and long-term memories (Bird and Burgess 2008). In particular, the hippocampus has been a major experimental system to unveil the role of synaptic plasticity. The hippocampus has also been implicated in a number of neurological and psychiatric disorders including epilepsy and Alzheimer’s disease. Additionally, the subgranular zone of the dentate gyrus is one of the sites of adult neurogenesis (Zhao et al. 2008). The study of the hippocampal formation has relied on its distinctive and readily identifiable structure at both gross and histological levels (Paxinos 2004). The hippocampus is divided

9Corresponding authors.
E-mail albin@binf.ku.dk; fax 45-3532-1281.
E-mail carninci@riken.jp; fax 81-48-4624868.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.084541.108.
into fields CA1–CA3 comprising the pyramidal cell layer, where the pyramidal cells are present, and a heterogeneous group of diverse GABAergic interneurons (Parra et al. 1998; Maccaferri and Lacaille 2003). The dentate gyrus comprises the molecular, principal, and polymorphic layers where the granule cells are the principal cells and the pyramidal basket cells are the most prominent class of interneurons. In a few cases, the use of cell-type-specific knockout mice for ligands or receptors of neuroactive molecules has made it possible to integrate physiological, anatomical, and molecular data from synapses to understand animal behavior (Tsien et al. 1996; Nakazawa et al. 2004). However, our understanding of the physiological organization of the hippocampus has been hampered by difficulties in describing the complete repertoire of neuronal cell types and their properties (Parra et al. 1998).

Thus, the identification of specific promoters that initiate transcription and drive gene expression in hippocampus, or even in specific subsets of cells within the tissue, will be important for developing new lines of mice with cell- or tissue-specific gene isoforms labeled and/or knocked out, or by the use of neuronal cell ablation (Watanabe et al. 1998; Tonegawa et al. 2003). The identification of novel promoters, particularly those that drive expression in a small number of neurons, may also lead to the description of rare neuronal types that have not been previously characterized.

To this end, we present DeepCAGE, a method combining CAGE technology with a high-throughput tag sequencer, the GS20 sequencer (454 Life Sciences [Roche]). We use DeepCAGE to create a comprehensive resource of hippocampal TSS and core promoters. Since with this method we can reach an unprecedented sampling depth (2 million tags, of which 1.4 million can be unambiguously mapped to the genome), we explored this data set to find promoters preferentially used in the hippocampus, their effects on the proteome, correlation with spatial expression data, and to understand the transcriptional regulatory program in the hippocampus.

**Results**

DeepCAGE sequencing

We adapted the CAGE method (Shiraki et al. 2003; Kodzius et al. 2006) to the 454 Life Sciences (Roche) GS20 sequencer as described in Methods (Fig. 1). Briefly, total RNA from CS7B6f/pooled mice hippocampi was purified and used as a template in the first-strand cDNA reaction primed by random primers to capture both the poly(A)⁺ and poly(A)⁻ RNA species. To extend cDNA synthesis through GC-rich regions in the 5’ UTR, we carried out the reverse transcription reaction at high temperature in the presence of trehalose and sorbitol (Carninci et al. 2002). CDNAs reaching the cap site were then selected by cap-trapping. They were then ligated to a linker having a recognition site for the class-IIs restriction endonuclease Mmel just next to the start of the cDNAs corresponding to the 5’-end of the original RNAs. This linker was used to prime second-strand cDNA synthesis. Subsequently, Mmel digestion cleaved 20 ~ 21 bp within the double-stranded cDNA, releasing CAGE tags. After ligation of a second linker to the 3’ end opened by Mmel digestion, CAGE tags were PCR amplified, purified, and further amplified before restriction and concatenation for direct sequencing (see Methods). The DeepCAGE technology does not require cloning in bacteria.

The key step for direct sequencing on the 454 device is the introduction of specific primer sites at the ends of the concatamers by mixing the CAGE tags at a ratio of 20:1 with a mixture of the 454 linkers “A” and “B.” Since these linkers (Margulies et al. 2005) can ligate DNA only on one side, they terminate the concatenation reaction and provide ends suitable for sequencing when A and B appear on the opposite sides of the concatamer, regardless of the orientation of the CAGE tags and insert size, which was optimized to be ~500 bp.

After a first test run, we produced two large-scale reactions, achieving in total ~2 × 10⁹ CAGE tags. After sequencing, tags were mapped to the mouse genome (mm8 assembly) using an algorithm based on our previous studies (see Methods); in total, 1.4 × 10⁹ tags map with high stringency (see Supplemental Table S1). The same mapping protocol was applied to all other CAGE tag libraries that are part of the data set used in this study. All CAGE tags have been submitted to the DNA Data Bank of Japan (DDBJ) under accession numbers AGAA0000001–AGAA0552486.

**Data sets and resource preparation**

Similar to most other high-throughput genomic technologies (The ENCODE Consortium 2007), CAGE data are most useful together with other data sets. In this study, we compare the hippocampus CAGE data to seven other CAGE data sets—each corresponding to a different tissue—with varying sequencing depth, including three brain tissues: visual cortex, somatosensory cortex, and cerebellum (see Supplemental Table S1). We also use the FANTOM3 cDNA set for associating promoters with genes and gene annotation (Carninci et al. 2005). As a resource for the community, the CAGE data sets (including “tissue-specific” promoter sets discussed below) prepared here are freely available as data tracks and sequence files at http://people.binf.ku.dk/albin/supplementary_data/hcamp/, from where they can be directly uploaded to the UCSC browser for visualization, or downloaded for analysis by power users. Additional statistics such as the fraction of tags mapping to known 5′-ends are shown in Supplemental Figure S1.

**Exploration of tissue preferences for core promoters**

As shown previously, for analysis of core promoters, it is helpful to group CAGE tags that map close to each other on the genome. Using the method of Carninci et al. (2006), tags from any tissue were grouped into a tag cluster if their genome mapping coordinates overlap on the same strand.

To explore the overall tissue preference of the CAGE clusters, we selected all CAGE tag clusters that have more than 30 tags per million (TPM), when counting all tissues. For clarity, TPM normalization is commonly used in tag-based studies and can be described as normalizing all tag counts so that the total count of mapped tags within a library equals 1⁰⁶ tags. The reason for this conservative cutoff is that a certain number of tags are needed to assess tissue distributions. For simplicity, we refer to these clusters as core promoters in this study, in the same way as in Carninci et al. (2006) and Sandelin et al. (2007).

This analysis identified 18,948 core promoters. To explore both what fraction of promoters are expressed primarily in one or a subset of tissues and what tissues have similar promoter usage, we hierarchically clustered these core promoters in terms of their expression and visualized the results as a heatmap (Fig. 2A; Eisen et al. 1998). We observe that:

1. The brain tissues cluster together in terms of promoter usage; in particular, visual and somatosensory have the most similar
usage, with few CAGE tag clusters being preferentially used in only one of these tissues.

2. The cortex tissues have substantial "smearing." Many promoters are used, but they are also shared between at least two tissues. This is a property seen also for macrophage and lung tissues, which might be due to the large number of macrophage cells present in lung tissue.

3. Conversely, there is a large set of promoters that are used mostly in the hippocampus. This promoter set has very little smearing, a feature shared with less complex tissues such as liver. The cerebellum is somewhere in between the hippocampus and the somatosensory cortex have the fewest number promoters only detected in hippocampus PEPs. (Supplemental Fig. S2).

As noted previously (Gustincich et al. 2006), promoters used preferentially in brain generally have multiple TSS and higher CpG content compared to promoters used preferentially in other tissues, which often have a single peak distribution of TSS, governed by a TATA-box (Sandelin et al. 2007). We found that hippocampus PEPs share the properties of the other brain PEPs in this regard—broader, CpG-rich promoters with fewer TATA patterns (Supplemental Figs. S3–S5). We also note that for promoters that are used strongly in many brain tissues, the hippocampus tag usage at the nucleotide level generally correlates well with tag distributions from the other brain tissues (Supplemental Figs. S6 and S7). There are exceptions to this; promoters where the tag distribution shape differs between tissues have been explored previously by Kawaji et al. (2006).

We then assessed where the PEPs from the various tissues were located in terms of overlap with known genes (see Methods). Figure 2B shows that the hippocampus, apart from having the largest amount of PEPs overall, also has the greatest number of PEPs located in intronic and intergenic space. This indicates that there are many strong promoters preferentially expressed in the hippocampus that have no known corresponding gene. This observation is not simply a sample size effect, as the cerebellum has almost as many PEPs with just ~18% of the sequencing depth compared to the hippocampus.

RACE validates distal upstream promoters

While PEPs falling within genes can be considered candidate alternative promoters for the same gene, PEPs in intergenic space are more likely to be the promoters of novel transcripts. We selected 10 intergenic hippocampus PEPs for RACE validation. Importantly, the selection was not in any way based on additional information that would be a validation in itself, such as EST...
sequences. Out of the 10 cases, eight had a PCR product, and the sequenced product validated a hippocampus PEPs in six of these cases. Of the failed cases, one had supporting evidence from other sources (overlap of 5'-ends of spliced ESTs) (data not shown). The outcome is comparable to that of RACE validation of intergenic transcribed regions from tiling array data in the ENCODE project (50%–70% success rate) (The ENCODE Project Consortium 2007); however, this should not be viewed as a true sensitivity measure of
DeepCAGE, for three principal reasons: (1) Even in perfect circumstances, RACE does not have perfect sensitivity; (2) as we are focusing on novel core promoters, we do not know the exon structure of the downstream product, which makes primer design nontrivial; and (3) many of the promoters have high GC content, which makes amplification challenging. Nevertheless, these results show that promoters inferred by DeepCAGE can be detected by other methods, as already shown previously in extensive validation experiments of the original FANTOM3 CAGE study (see Supplemental Material of Carninci et al. 2006).

In the annotation process, we noticed a considerable number of cases in which intergenic hippocampus PEPs were located relatively close to the 5'-end of a known gene. These cases are likely novel alternative upstream promoters. An example is shown in Figure 2C: CAGE and RACE data show a novel hippocampus PEP upstream of the mouse Bai3 (brain-specific angiogenesis inhibitor 3) gene. An extreme case is shown in Figure 2D: CAGE identifies a hippocampus PEP that is upstream of the Arpc5 (actin-related protein 2/3 complex, subunit 5) gene, but on the other strand, forming a bidirectional promoter. RACE validation as well as human orthologous transcripts and EST evidence show that the novel promoter is likely a distal upstream alternative promoter of the Rgl1 (ruminat guanine nucleotide dissociation stimulator-like 1) gene, whose RefSeq-annotated start site (which is also a hippocampus PEP) (data not shown) is a remarkable ~141 kb downstream from the novel promoter. Therefore, while we focused on intergenic promoters to find novel transcripts, we often identified novel promoters that provide new ways to regulate the transcription of known genes.

Brain tissues use different alternative promoters within the same gene

As shown in Figure 2B, the majority of hippocampus PEPs are located inside genes, overlapping annotated exons. It is likely that many of these are alternative promoters for the same gene, since many of them are supported by full-length cDNAs (see examples in Figs. 3 and 4). Alternative promoters are interesting for three reasons: (1) They allow a gene to have multiple, distinct, regulatory inputs; (2) alternative promoter locations can affect the protein content of the gene product similarly to alternative splicing; and (3) it is important for molecular approaches in neurobiology to selectively knock down gene isoforms that are preferentially used in a given tissue.

We first identified all genes containing one or more PEPs from hippocampus, somatosensory cortex, visual cortex, and cerebellum inside exons. Then, we counted the number of genes with multiple distinct PEPs from the different tissues (Fig. 3A). The Dlgap1 gene (guanylate kinase-associated protein [GKAP] or SAPAP [synapse-associated protein 90-postsynaptic density-95-associated protein]) (Fig. 3B) is exceptional since it has four core promoters that are preferentially used in hippocampus, somatosensory cortex, visual cortex, and cerebellum, respectively. All of these PEPs overlap corresponding 5'-ends from full-length cDNAs (Fig. 3B); in this case, the CAGE verifies these 5'-ends and assigns tissue expression constraints. Dlgap1 is a scaffolding postsynaptic density protein at excitatory synapses that contains 14-amino-acid repeats at the N terminus involved in protein–protein interactions and that are affected by different promoter usage (Kim et al. 1997; Romorini et al. 2004); the CAGE data indicate that all the PEPs are upstream of these repeats except for the cerebellum, indicating that cerebellum transcripts do not include the repeats. Thus, the selection of alternative promoters has in this case a clear functional consequence.

We then sought to systematically identify potential changes in protein domain composition caused by usage of hippocampus PEPs. Using cDNA data, we predicted protein domains to genomic positions and determined in how many cases a hippocampus PEP falls within a gene but downstream from a protein domain within the same gene, which then would give a protein product that is lacking the domain in question. Using conservative criteria (see Methods), we found 50 such genes (see Supplemental Material). Three examples (Pchq, Bai1, and Myo10), showing dramatic protein domain content diversity, are shown in Figure 4.

Transcription factor binding sites analysis on specific core promoters

An advantage with the CAGE is that tags give high-resolution mappings of active TSS, which can be used to pinpoint core promoters for computational sequence analyses (Wasserman and Sandelin 2004). We first analyzed the ~1000 to +200 region surrounding PEPs from the tissues in Supplemental Table S1 for significantly overrepresented motif matches from the JASPAR database (Vlieghe et al. 2006). Our results are largely consistent with previous studies of promoters used primarily in single tissues—for instance, homeobox motifs are overrepresented in embryonic PEPs, ETS motifs in macrophage PEPs, and so on (Supplemental Tables S3 and S4). Since CAGE data may also be interpreted as promoter usage (the number of tags mapped to a loci), we investigated what transcription factor genes are strongly expressed in hippocampus, and whether their predicted transcription factor binding sites have a clear preference to the promoters that are preferentially used in the same tissue.

Figure 5A plots the fraction of hippocampus tags in transcription factor genes compared to other tissues versus the overall hippocampus expression of the same genes (CAGE TPMs) (see Methods). Only a handful of transcription factor genes stand out as very highly expressed in hippocampus, including Amt2, Sp3, and Aes, and only some of these have a clear preference for

**Figure 2.** Exploration and validation of identified core promoters. (A) Exploration of tissue usage in all core promoters having more than 30 tags per million using hierarchical clustering, with CAGE tag expression data from the actual core promoters. Preferential usage for a certain tissue (fraction of tags belonging to the tissue in question) is color-coded as shown in the legend. Rows represent individual promoters (the row dendrogram is omitted because of the large number of rows), while columns are the different tissues. (B) Number of core promoters used preferentially in just one tissue (PEPs, as defined in Methods), we found 50 such genes (see Supplemental Material). Three examples (Pchq, Bai1, and Myo10), showing dramatic protein domain content diversity, are shown in Figure 4.
We compared these highly expressed transcription factors with mouse in situ hybridization experiments from the Allen Brain Atlas (Lein et al. 2007, see image 5B-I). Overall, there is a high correspondence between both preferential expression in hippocampus and overall strength of expression between CAGE and in situ data (data not shown). Interestingly, its paralog Arnt2 is primarily expressed in brain. According to the CAGE data, Arnt2 is highly, and preferentially, expressed in hippocampus (Fig. 5A). Furthermore, in situ images of Arnt2 confirm a distinctive expression of Arnt2 in the C1 region of the hippocampus (Fig. 5B). This leads to the hypothesis that the Arnt predicted sites are, in fact, sites for Arnt2, which would make Arnt2 a major factor in hippocampus transcription regulation.

Promoters used in restricted cell types in hippocampus

We have focused above on relatively strong promoters having more than 30 TPMs in order to study the distribution of tags from different tissues in a statistically valid way. The significance of transcripts that are present in a tissue with low frequency has been met with suspicion, and the observations were often labeled either as methodological or transcriptional noise. Although both of these are still a possibility, we have explored the expression properties of known genes having substantially less than 30 TPM by analyzing their spatial expression patterns using in situ hybridization data. In Figure 6, we compare the number of tags hitting a known gene and the corresponding in situ images. Note that in almost all these cases, the tags hit the annotated 5′-end of the gene (Supplemental Fig. S8).

For most of the highly expressed transcription factors, we have no corresponding computational model for how they bind DNA. However, since many transcription factors from the same structural class bind similar target sequences (Sandelin and Wasserman 2004), observed overrepresentation of hits with a given model might be due to binding sites from a different factor with similar binding preferences. As an example, the predicted sites for the well-studied bHLH-PAS Arnt gene are overrepresented in hippocampus PEPs (Supplemental Table S3), but the Arnt gene is lowly expressed in the whole brain as measured by both CAGE and in situ data (data not shown).
Discussion

We are on the verge of a new era, in which sequencing technology can be used to infer biological function on a comprehensive scale, and the power of sequencing centers will be available to normal laboratories. Here, we have modified the CAGE protocol for the 454 Life Science instrument and demonstrate the usefulness of deep sequencing to discover new promoters in complex tissues.

We have identified a large number of core promoters that are preferentially used within hippocampus. Our results indicate that of the tissues we tested, the hippocampus has the largest number of such promoters, closely followed by cerebellum. These results may be due to two different factors: cell type diversity of the tissue and sequencing depth. The cerebellum is one of the least complex among brain tissues, while the hippocampus and the cortex tissues are very complex, with a plethora of different cell types. This is also shown in Figure 6, where it is evident that small, distinct cell populations within the hippocampus express a given gene, while most other cells do not. The methods we use to measure transcription cannot quantitatively measure the diversity of cells within a tissue and consequently neither the transcription dynamics within single cells. This strongly motivates further developments to assess the expression rates of genes within smaller cell populations. Such approaches in combination with the in situ data

Figure 4. Examples of changes of domain content for genes by use of hippocampus PEPs. Hippocampus preferentially expressed promoter (PEP) locations are shown as red triangles. Locations of predicted protein domains are shown as colored blocks (note that domains spanning more than exons are extended over the intron(s)). In all of these cases, at least one domain is upstream of the PEP, which means that this domain is not included in the isoform expressed in hippocampus. Known cDNA locations are shown below: transcription is right-to-left. BAI1 is a membrane protein whose N-terminal domain is extracellular, with a transmembrane region just downstream from the GPS domain (data not shown). The extracellular part can be cleaved off at the GPS domain, releasing a tumor-suppressing peptide; however, the hippocampus PEP is just downstream from the GPS domain, presumably giving a BAI1 variant that is attached to the membrane but without the extracellular domains, which lack the tumor-suppression capability (Kaur et al. 2005). Similarly, the PEPs in Myo10 confirm a previous study showing the neuronal expression of an isoform lacking the Myosin head domain (Sousa et al. 2006). In Pclo, the zf-piccolo domain cannot be included when using the hippocampus PEP.
from the Allen Brain Atlas may result in a new molecular taxonomy of the different types of hippocampal cells providing a framework for a complete description of the components of hippocampal cell's network (Gray et al. 2004; Ma 2006; Sugino et al. 2006).

The subset of novel hippocampus promoters that are not overlapping known transcripts could indicate noncoding RNAs that have as yet not been sampled in mouse; Mercer et al. (2008) showed the existence of other long noncoding RNAs that are expressed primarily in brain. On the other hand, we also find that many of the “novel” promoters, in fact, are new upstream promoters for known genes.

However, most of these novel promoters fall within known genes, and we find that many genes have different core promoters that are used preferentially by different brain tissues, which may give partially different RNAs and protein products. In extension, identification of cell-type-specific alternative promoters for genes encoding for proteins responsible for neuronal and synaptic activity (channels, receptors, etc.) may provide increased specificity for drug treatments for epilepsy and other hippocampal-related neuropsychiatric disorders. Although this approach may seem far from our current technology, the use of antigene RNAs (agRNA) or peptide nucleic acids (agPNA) as well as of Locked Nucleic Acids (LNA) that target specific promoters has been proposed and demonstrated in vitro (Janowski et al. 2005a,b). This is particularly relevant since new promising strategies for delivery of nucleic-acid-based modifiers of gene expression into the brain have been...
recently proven (Kumar et al. 2007). To this end, the data set we present here, enabled by DeepCAGE, is to date the most comprehensive brain-centric promoter-exploration resource.

### Methods

#### Preparation and sequencing of CAGE libraries

The preparation of the CAGE library is adapted from Shiraki et al. (2003) and Kodzius et al. (2006), to work with the 454 Life Sciences sequencer. The schema is represented in Figure 1. A detailed protocol of the CAGE library preparation, starting from trizol-extracted RNAs, is available in the Supplemental Material.

Once the CAGE library is prepared, we test various ratios of beads to CAGE library ratios, using usually an excess of DNA over beads (1:4 to 1:16 ratio beads:DNA) in the 454 GS20 protocol. During the calibration of the instrument, small-scale runs (1/8 of small kit runs) are used to calibrate the best DNA/beads ratio, followed by one or more runs of 454 large-scale sequencing kits (further details at http://www.454.com/).

#### In silico mapping of CAGE tags

Sequenced tags CAGE tags were mapped to mouse chromosomes and the mitochondrial genome (Genome build: mm8) using the BLAST/Vmatch alignment programs, and the longest full-matched (meaning no mismatches in the middle) positions were selected. These tags were referred to as “single-mapped” tags. Tags that map to multiple locations on the genome (with the same length) were called “multi-mapped,” and tags that did not map (mapped <18 bp long) were called “unmapped.” These multi-mapped and unmapped tags were passed to the rescue stage to increase the number of “single-mapped” tags (see Supplemental Material), since many promoters share identical subsequences (Faulkner et al. 2008). Rescued tags were incorporated into the single-mapped tag collection, and other tags were discarded. In the rest of the analysis, we use only the single-mapped tags; note that the same mapping procedure was applied to all CAGE libraries in the study.

#### Mouse hippocampus RNA preparation and 5'-RACE PCR validation of target intergenic core promoters

Adult C57/Bl wild-type mice (n = 5) were sacrificed by CO2 inhalation, and hippocampal regions were rapidly dissected and snap-frozen in liquid nitrogen. Total RNA was extracted with TRizol reagent (Invitrogen) following the manufacturer’s protocol; the RNA sample was treated with DNase (Ambion), aliquoted in RNase free LoBind tubes (Eppendorf), and stored at −80°C. 

RACE-ready cDNA was obtained with the Generacer kit (Invitrogen) following the manufacturer’s protocol with no modifications starting from 5 μg of hippocampus total RNA. 5'-RACE was carried out using Platinum Taq DNA polymerase High Fidelity (Invitrogen); each PCR product was cloned in a TOPO TA vector (Invitrogen) and transformed in OneShot Top10 chemically competent Escherichia coli cells. Five colonies from each plate were selected for growth, DNA extraction (DNA Mini Kit; QIAGEN), and sequencing.

---

**Figure 6.** CAGE identifies promoter activity from small subpopulations of hippocampal cells. Examples of correspondence between CAGE tags and signal detected by in situ hybridization, ordered from relatively high expression (from the top left quadrant), expected to correspond to RNAs that are expressed only in a specific subset of cells.
Oligonucleotide primers for the validation of intergenic core promoters were hand-designed according to guidelines from the Generacer kit manual and checked with PerlPrimer for possible primer-dimer formation. Primers used to validate the core promoters are shown in Supplemental Table S1.

Generation of tag clusters
A tag cluster (TC) was defined as the maximum set of tags where all 5'-ends are <20 bp from the closest neighbor, and on the same strand. We chose 20 bp because it is approximately the length of a CAGE tag, and thus we know with certainty that the transcript starting at the tag's 5'-end at least spans this region. This is the same definition as used in Carninci et al. (2006).

Exploration of tissue preferences of tag clusters
We first normalized the expression of each TC to TPMs for each tissue:

\[ n_{c,i} = r_{c,i} \times 10^6 / r_i \]

\[ n_c = \sum n_{c,i} \]

where \( r_{c,i} \) is the number of non-normalized tags in cluster \( c \) for tissue \( i \), \( n_i \) is the total number of non-normalized tags for tissue \( i \), and \( n_c \) is the total TPM for the cluster. Only TCs with \( n_c \geq 30 \) TPM were considered. We then normalized the total expression of each such tag cluster to sum to 1:

\[ t_{c,i} = n_{c,i} / n_c \]

where \( t_{c,i} \) is the normalized contribution of tissue \( i \) to cluster \( c \). We then hierarchically clustered the set of promoters in terms of expression in each tissue (the \( t_{c,i} \) values) using Euclidian distance measure and complete linkage as the clustering method (the defaults of the dist() and hclust() functions in R, respectively). The reordering of columns (tissues) and rows (tag clusters) was visualized using the heatmap.2() function in the gplots R package.

Generation of preferentially expressed promoters (PEPs)
To call a TC \( c \) preferentially expressed in a tissue \( i \), we considered:

1. \( t_{c,i} \) values as defined above. We required one such value to be >0.5, since with this cutoff, a TC will be preferentially expressed only in one tissue.
2. The assessment whether this over- or underrepresentation was significant, that is, unlikely to have arisen from random sampling from the underlying tags. This can be expressed as a binomial overrepresentation test. We required that the TC in question presented a \( P \)-value <0.05 in a one-tailed binomial overrepresentation test (R function: binom.test).
3. We only assessed core promoters \( n_i > 30 \) TPMs. The tag number constraint is not strictly necessary; we introduced the additional constraint to reduce the number of statistical tests (as tests with few tags will always be insignificant) and to focus on strong promoters.

Mapping PEPs to genes and introns
PEPs were considered to belong to a gene if they had at least one tag on the same strand within the boundary of its transcript (using the RIKEN cDNA database) including a 50-bp slack at the 5'-end of the gene. If a PEP had no such overlap, it was considered intergenic. PEPs belonging to genes were further divided into exonic if the PEP overlaps with an exon, or otherwise as intronic.

Domain annotation and PEPs
Domains were annotated using RIKEN cDNA annotation (corresponding to Interpro domain locations). To determine whether transcription initiation at the hippocampus PEPs changed the domain product, we used the gene mappings from above. Then we checked whether any domain in a gene containing a hippocampus PEP was upstream of this PEP and downstream from the annotated transcription start site. Usage of this PEP would result in the domain being lost and consequently in a different protein product.

TFBS overrepresentation analysis
We searched all sequence sets with the JASPAR matrices (Vlieghe et al. 2006) using the AASAP tool (Marstrand et al. 2008) with the following setting: uniform background model, a pseudo-count of 1, and threshold value of 0.7 relative to the matrix-specific scoring range. For all matrices, we calculated a \( P \)-value representing the overrepresentation using the binomial test as described in van Helden et al. (1998). For the tables, the \( P \)-value threshold is <0.01.

In situ comparison
For comparing expression and tissue preference of transcription factors between CAGE and in situ experiments available from the Allen Brain Atlas (Lein et al. 2007), we calculated the hippocampus strength versus tissue preference \( T \) for each transcription factor gene (using the RIKEN TF database [Kanamori et al. 2004]):

\[ T = \sum \frac{n_{c,\text{hippocampus}}}{\sum n_{c,i}} \]

where we sum over all TCs \( c \) that are within the boundary of the gene of interest, and all the brain tissues \( i \). \( n_{c,i} \) is the TPM count for respective TC and tissue. This was then visually compared with corresponding in situ images, downloaded from http://www.brain-map.org/.

Acknowledgments
P.C. and Y.H. are supported by the National Project on Protein Structural and Functional Analysis from MEXT and the National Project on Genome Network Analysis and the RIKEN Genome Exploration Research Project from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government. P.C. is also supported by a grant by the EU 6th Framework Program (NFG project). Authors affiliated with the Bioinformatics Centre are supported by a grant from the Novo Nordisk Foundation. The European Research Council has provided financial support to A.S. under the EU 7th Framework Programme (FP7/2007-2013)/ERC grant agreement 204135. S.G. is supported by a career development grant from “The Giovanni Armenise-Harvard Foundation.” We thank Susan Sunkin for help with in situ images and scientific discussion, Kazuho Ikeo and Toshitsugi Okayama for developing and sharing the CAGE mapping procedures before publication, and Akira Hasegawa for technical support.

References


Paper IV
Estimating the coverage of tag-sequencing experiments at multi-level resolution

Eivind Valen\textsuperscript{1}, Albin Sandelin\textsuperscript{1}, Piero Carninci\textsuperscript{2}, Anders Krogh\textsuperscript{1}, Ole Winther\textsuperscript{1,3,*}

\textsuperscript{1} The Bioinformatics Centre, Dept. of Biology and the Biotech Research and Innovation Centre (BRIC), University of Copenhagen, Ole Maaloes Vej 5, 2200 Copenhagen, Denmark
\textsuperscript{2} LSA Technology Development Group and Functional Genomics Technology Team, Omics Science Center, RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045 Japan
\textsuperscript{3} DTU Informatics, Technical University of Denmark, Building 321, DK-2800 Lyngby, Denmark

Email: Ole Winther\textsuperscript{*} - winther@binf.ku.dk;
\textsuperscript{*Corresponding author

Abstract

Background: Tag-based methods have become increasingly popular to measure expression, but also to discover transcription start sites (TSSs) and core promoters. These methods produce a number of reads sampled from a population of transcripts or sequences, which means that rare entities might not be observed if the sample size is not sufficiently high. This raises the questions: how many transcripts have yet to be discovered, what fractions of the expression are these representing (their coverage) and what is a reasonable sequencing depth to obtain a suitable coverage of the underlying transcript population?

Methodology/Principal Findings: Here, we implement and compare two state-of-the-art methods, a parametric and a non-parametric (or distribution free), with the goal to predict the coverage of observed transcripts, and, by extrapolation also predict the amount of unique transcripts to be attain at given sequencing depth. Both of these methods are implemented as R code which can easily be used by experimentalists. As an applied example, we use these methods on a large set of TSSs from various tissues to investigate the difference in transcriptional diversity in different samples and on different abstraction levels: individual transcription start sites, core promoters and genes.

Conclusions: We observe in cross-validation tests that both methods give quite precise predictions, comparable in performance, but tend to have different types of biases under and over-estimate the number of unique TSSs,
respectively. This suggests that it is worthwhile considering both, using the consensus as a better predictor. From a biological standpoint we observe that the present CAGE libraries are not close to 100% coverage (saturation) if assessing unique transcription start sites (TSSs), but the larger libraries are close to saturation if assessing larger entities such as core promoters. We also show substantial differences in the amount of distinct TSSs, core promoters and genes, and thereby transcriptional repertoires, that different tissues use. The libraries are not as close to saturation in terms of unique “atoms” (TSSs, core promoters or genes) showing that many rare atoms are present in the libraries at all resolutions. This means that both the expected cellular diversity of samples as well as the desired biological annotation level must be considered deciding on the number of tags to sequence in an experiment.

**Availability:** An implementation of the methods as well as the data sets used are available at [1].
Background
High-throughput sequencing is now the method of choice for gene expression, protein-DNA interaction and small RNA discovery. It can also be used to discover and characterize transcription start sites (TSSs) and promoters on unprecedented scales. All of these methods are sampling-based: the experiments yield a large number of double-stranded DNA sequences corresponding either to transcripts or bound DNA regions, where only a fraction will be sequenced. Therefore, a tag library should be viewed as a sub-sampling of a much larger population, meaning that some rare entities will only be detected consistently with a large sequencing depth. It follows that the number of tags sequenced is important - this will be a trade-off between sequencing cost on one hand and the need for a comprehensive coverage. However, in most cases the choice of sequencing depth is made ad hoc, and is not subject to a stringent analysis. Thus, estimating how much of the active targeted entities that are detected in a given experiment is important. While this problem is equally important in, for instance, RNA-seq and ChIP-seq, we have chosen to use 5’ cDNA tag data, aimed to find active transcription start sites, as our example data sets. These methods, oligo-capping [2] or Cap Analysis of Gene expression (CAGE) [3], are based on capturing capped mRNAs, and only extract a short sequence tag starting from the 5’ end of such mRNAs. These tags are then mapped to the corresponding genome. An advantage with these methods is that they are not biased by annotation, have a resolution on nucleotide level, and at the same time can measure the initiation site usage of any nucleotide, which makes it possible to dissect promoters in detail. This means that we can both detect functional entities like core promoters and TSSs, but also measure expression on promoter or gene level (simply the number of tags overlapping a gene). The data sets are also typically a magnitude larger than EST or SAGE data used for previous digital gene expression profiling. We will use CAGE data as the application example throughout, since it has larger data sets from primary tissues than oligo-capping. Despite the data depth, from initial studies of CAGE tag data, it became evident that even with the comparatively large sequencing depth (between 50,000 to 1,200,000 mapped tags from a single tissue or cell line - more if several libraries from the same tissue are pooled), some transcripts that could be detected with “subtracted” cDNA libraries (see [4] for a review of cDNA subtraction) and/or PCR, were not hit by any tags. Thus, in this work we will focus on assessing how many of the unique TSSs/core promoters/genes from a given cell sample that are found at a given sequencing depth. However, the same framework could be applied to both RNA-seq and ChIP-seq. In this work, we assess this by three complementary statistics

- The total number of unique TSSs is a difficult quantity to estimate because it in principle requires
extrapolating to infinite sample size. One therefore often opt for extrapolating to sample sizes of the same order as the current library to get the second statistic:

- The number of new unique TSSs we can expect to observe in a new sample. This growth curve statistic thus gives equal weight (one) to each unique TSS independently of whether it is a common or more rare TSS. This motivates the third statistic:

- The coverage. The underlying concentrations (or expression levels) of all unique TSSs are unknown but can be estimated from the frequencies of the observed unique TSSs in the library. We will both estimate the expression of each of the observed unique TSSs and the total amount of expression for the unobserved TSSs. The total estimated expression of the observed unique TSSs is called the coverage. This statistic thus weighs tag position by estimated expression giving a small weight to rare TSSs. A coverage of, say 90%, means that there is 10% chance that a newly sequenced tag is a new unique tag and 90% it is not.

Ultimately, we would like to achieve saturation, meaning qualitatively that further sampling will not result in any substantial gain of information in terms of increased coverage. This is a more realistic goal than aiming to capture all unique TSSs. One will very rarely have 100% coverage so what is regarded substantial will depend upon technical, biological and economical factors. The coverage and number of unique TSSs are not only affected by sequencing depth and methodology used (such as the sensitivity and noise level of a given technique): it also reflects a number of biological parameters, such as i) how diverse the collection of cells that is giving the RNAs that are ultimately sequenced, as even a single tissue have many sub-types of cells, ii) how diverse the actual repertoire of mRNAs in a cell type is. Therefore, assessing the amount of achieved saturation is not only motivated by economical constraints, but will also give biological insights. Motivated by this need, in this work we implement and compare two state-of-the-art parametric and non-parametric (or distribution free) methods of using the two above statistics to get insight about the coverage and total size of the entities targeted by sequencing experiments - the transcriptome if using RNA-seq, the “promoter-ome” if using CAGE [5]. To show that application of the methods can lead to new biological insights, we compare eight data sets from different tissues on different abstraction levels, and conclude that there is a large difference in the range of TSSs, core promoters and genes used by different tissues, and that the coverage for current experiments is insufficient if the goal is discovery of individual TSSs, but acceptable for larger entities such as core promoters. The context will thus depend upon the biological phenomena we are probing (here promoters), the tissue and the also the resolution.
Results
Data sets and data processing
The CAGE data protocol is described in [3]. Here, we use the same mouse CAGE data as presented in [6], and focus on tags from the tissues in Supplementary Table 1. Tags were mapped to the genome as described in [6]. We only consider the actual 5’ end position of the mapping. A nucleotide on a given strand that is indicated by at least one CAGE tag to be a 5’ end is, as in [7], defined as a CAGE-defined TSS (CTSS). The number of tags for a particular CTSS is assumed to be proportional to the actual usage rate of that nucleotide as a TSS. Additional information and material are available at [1].

Statistical approaches to estimating the size of the transcriptome
In this work, we are interested in estimating how large a fraction of the total population of genomic atoms that we have sequenced in a particular experiment. What is is considered a genomic atom depends upon the chosen resolution. In this paper we use CAGE data as an example and consider three levels of abstraction: transcription starts sites (TSSs), tag clusters and genes (described below). In statistical terms, we want to estimate the probabilities (expression levels) associated with a possibly infinite number of discrete outcomes. The number of possible outcomes (bins) is not known to us a priori. We will use the term species for the bins because this problem is equivalent to classical problem in ecological studies of estimating the number of unobserved species from a finite collection of observed species in a certain biotope [8, 9]. Clearly, to estimate the expression levels of unobserved species and the number of unobserved species require smoothness assumptions about the distribution of expression levels. Fitting this smooth distribution to the observed data thus allow us to make inferences about the number of unobserved species, what fraction of the expression they represent and how many new species we will observe in a new experiment of a given size. In our setting, this corresponds to asking, for a given tissue, how many new TSSs that will be discovered in a new CAGE sequencing experiment of a given size (= number of tags sequenced and mapped to the genome). Knowing this, we can estimate what sampling size that is necessary to obtain the desired coverage (= summed expression levels of observed species). We investigate and compare two distinct statistical approaches described in some detail below and in Methods.

The first approach, the so-called stick breaking process, is a way to generate a distribution of expression levels for potentially infinite number of species. Here we will focus on a specific computationally simple but powerful variant, the Pitman-Yor or two-parameter Poisson-Dirichlet process. We will also sketch how generalizations may be realized. We will refer to this method as non-parametric, meaning that no
assumptions are made that the data are drawn from a given probability distributions. The motivation for using a non-parametric approach is that the data is complex (spanning the range of many unique TSSs with low frequency to few unique TSSs with high frequency) that we need a flexible model to fit the data. In the second approach, rather than working directly on the species level, a statistical model is introduced for the frequency of frequency statistic which summarizes how many species have been observed one, two, etc. times. In this representation, the number of counts for a specific species is modeled by a binomial or (taking the limit of large library size) a Poisson distribution. To obtain a sufficient flexible model that can fit large complex data sets, the parameter of binomial (or Poisson) is assumed to be drawn from a flexible distribution so that the resulting model is an infinite mixture of binomials (or Poissons). We will refer to this methods as generalized inverse Gaussian-Poisson (GIGP) because the the model is Poisson with parameters drawn from a generalized inverse Gaussian [10].

### Stick-breaking processes

The stick-breaking process can most easily be understood from a species sampling formula. Here we consider the Pitman-Yor sampling formula (or two-parameter Poisson-Dirichlet process) [11–13]. Below we will outline how to go from the stick breaking process to the sampling formula. Recently, Lijoi, Mena, Prünster and co-workers [14–17] have introduced a general Bayesian nonparametric methodology for species discovery and applied it to EST data sets. Here we give an implementation and presentation of the Poisson-Dirichlet process suitable for large data sets since we are considering CAGE data sets two magnitudes larger than the previously considered EST libraries (order $10^6$ tags versus $10^4$).

The Poisson-Dirichlet process can be understood as a generative model for simulating TSS libraries. We have observed a library of $n$ tags specified as a list (sequence) of genomic coordinates $(c_1, c_2, \ldots, c_n)$. This library contains $k$ unique TSSs with counts $n = (n_1, \ldots, n_k)$, $n = \sum_{j=1}^{k} n_j$. Without loss of generality we will label the tags in order of their arrival such that $c_i \in \{1, \ldots, k\}$. For a given setting of the two model parameters $\sigma$ and $\theta$, the sampling formula specifies the probability distribution for the $n + 1$th sample to belong to one of the $k + 1$ possible categories, $c_{n+1} \in \{1, \ldots, k + 1\}$, the $k$ previously seen TSSs or a new one. In the sampling formula the probability of observing a previously unseen TSS, i.e. $c_{n+1} = k + 1$ and $n_{k+1} \leftarrow 1$, has probability:

$$P(c_{n+1} = k + 1|n, \sigma, \theta) = \frac{\theta + k\sigma}{n + \theta}. \quad (1)$$

6
The probability of re-observing unique TSS $j$ for $\in \{1, \ldots, k\}$, i.e. $n_j \rightarrow n_j + 1$, is

$$P(c_{n+1} = j|n, \sigma, \theta) = \frac{n_j - \sigma}{n + \theta}.$$  \hspace{1cm} (2)

The parameters of the sampling formula can be given the following interpretation: $\sigma \leq 1$ determines the tendency to observe species with few counts and $\theta$, which we take to be non-negative, acts as a pseudo-count for the new species probability. When $\sigma < 1$ there will be a decreasing tendency to draw new TSSs when the sample grows.

We can make a simulated experiment using the sampling formula: Starting from our observed library we draw a new tag with the probabilities specified by the formula. This modifies the sampling probabilities for the next tag. We then update the probabilities, draw a new tag and iterate this process. This recursion can be repeated to give predictions for the number of unique TSSs $k$ as a function of $n$. Simulating this process a number of times gives predictions with confidence intervals [14–16]. However, with the size of libraries considered here the relative uncertainty due to sampling of tags is negligible compared to the bias of the model (see below). So we omit the sampling step and use the exact form for the expected number of species of a sample of size $n + n'$, i.e. $n'$ additional tags), that has recently been derived in closed form [17]:

$$\text{Expected}[k](n + n') = k + \left(k + \frac{\theta}{\sigma}\right) \left(\frac{(\theta + n + \sigma)n'}{(\theta + n)n'} - 1\right).$$

Another quantity of perhaps even greater interest is the coverage which is defined as the sum of the expression estimates of TSSs we have observed so far. Eq. (2) gives the estimate of expression level of species $j$ so we have to sum this quantity over all $j \in \{1, \ldots, k\}$ to get the coverage:

$$\text{Coverage} = \sum_{j=1}^{k} \frac{n_j - \sigma}{n + \theta} = 1 - \frac{\theta + k\sigma}{n + \theta}. \hspace{1cm} (3)$$

An exact formula for the extrapolated coverage is also available [17]:

$$\text{Expected}[\text{Coverage}](n + n') = 1 - \frac{\theta + k\sigma (\theta + n + \sigma)n'}{\theta + n (\theta + n + 1)n'}.$$

By definition the coverage is equal to one minus the probability of observing a new species, which is also called the discovery probability. Note that the expression estimate of the process model eq. (2) is a modified version of the maximum likelihood estimate for a finite dimensional multinomial $n_j/n$. The two model parameters is of course crucial for the predictions we get and should be learned from the TSS library data. We discuss learning these parameters from the data by using maximum likelihood and in the Bayesian setting in Methods. The applicability of this model to genomic data has been validated on small data sets by [14, 15] and here we extend the analysis to library sizes relevant for high-throughput sequencing.
One should be aware that the non-parametric model assumes that the population is infinite, i.e. when \( \sigma > 0 \) there will always be non-zero probability to observe a new species. In practice this probability can be vanishingly small, but we can only talk about strict saturation when the probability for observing a new species goes exactly to zero. This happens when the numerator \( \theta + \sigma k \) in the the sampling formula for a new species is equal to zero. Since we have chosen \( \theta \geq 0 \) this requires \( \sigma < 0 \). In that case we get an upper bound for the number species \( k_{\text{max}} = -\theta/\sigma \).

In the stick breaking construction, species probabilities (expression levels) \( p_j \) are generating from the following stochastic process: draw weights \( v_j \in [0,1] \) \( j = 1, \ldots \) independently from a distribution \( \text{Dist}_j \) on the unit interval and set \( p_1 = v_1 \) and \( p_j = (1 - v_1) \ldots (1 - v_{j-1})v_j \) for \( j \geq 2 \). This stick-breaking analogy is understood from taking a stick of unit length and sequentially dividing into pieces according to \( p_1, \ldots \). The most well-known choice of unit interval distribution is the beta-distribution with parameters \( \alpha_j \) and \( \beta_j \): \( \text{Beta}(\alpha_j, \beta_j) \). It can be shown that the Pitman-Yor process corresponds to drawing \( v_j \) from \( \text{Beta}(1 - \sigma, \theta + j\sigma) \), see e.g. [11, 18]. The advantage of working with the sampling formula is that the explicit \( p_j \) dependence has been averaged (marginalized) out so that one doesn't have to work with a potentially very large number of probabilities. However, one may work with an explicit representation of the \( p_j \)s [18], allowing for more general stick breaking constructions such as the recently proposed probit based model [19]. In this paper we stick to the Pitman-Yor process because it is elegant, quite flexible and relatively accurate on our data sets, noting that given the size and the prediction accuracy on our data sets it is definitely an interesting methodological challenge to construct better more flexible processes.

**Frequency of frequency models**

The frequency of frequency model has been developed in a series of papers [5, 10, 20]. We can model the probability that species \( i \) is observed \( n_i = 0, \ldots, n \) times by a binomial distribution. To make the model more flexible we let each species have its own binomial parameter: \( \text{Bin}(n_i|q_i, n) \). The parameters \( q_1, \ldots, q_k \) should be tied together in some way for example by setting \( q_i = q \) or in a hierarchical fashion such that each \( q_i \) are drawn independently from the same distribution of parameters \( p(q_i|\gamma) \), where \( \gamma \) is short hand for shared hyper-parameters. In general the binomial with \( q_i = q \) is a too restrictive distribution to fit this type of data [5, 20]. Instead we will use an (infinite) mixture model

\[
p(n_i|\gamma, n) = \int p(n_i|q, n)p(q|\gamma) \, dq .
\]
Refs. [10, 20] investigated different alternatives and the empirical findings of [10] suggest that the four parameter generalized inverse Gaussian distribution is a better choice than the beta distribution. We will also use the generalized inverse Gaussian distribution. The binomial may be well approximated by a Poisson $\text{Poisson}(\lambda_i)$ in this case because $n$ is large and the mean of the process $\lambda_i = q_in$ is order one for most of the species. The Poisson distribution will be used in the implementation.

Once we have fitted the model, for example by maximum likelihood (ml) (see Methods), we can make predictions about different quantities of interest. Specifically, $p(0|\gamma_{ml}, n)$ is the probability that a species is not observed in a sample of size $n$. The expected fraction of observed species to the total number is therefore $1 - p(0|\gamma_{ml}, n)$. This means that we can estimate the expected total number of species as $k_{\text{total}} = k/ [1 - p(0|\gamma_{ml}, n)]$ and the expected number of species as function of the sample size $n'$ as $k'(n + n') \approx k [1 - p(0|\gamma_{ml}, n')]/ [1 - p(0|\gamma_{ml}, n)]$. This species growth curve is also estimated in the stick-breaking process model allowing for direct comparison between the two approaches. This approach does not estimate the coverage whereas the stick breaking approach does not estimate the total number of species. In these respects the two approaches are complementary.

**Implementation**

We implemented both these methods in R [21]. In the rest of the text, we will refer to the methods as non-parametric (=stick breaking process) and GIGP (=frequency of frequency model). R code is downloadable at [1]. The R code is split into four files: two containing methods specific to each of the methods (nonparam.r and GIGP.r) and one with the shared functions dependent on the two former (shared.r). The latter includes methods for generating and plotting cross validation data, extrapolations and the coverage tables. The final file contains code to read in data and generate all plots featured in this paper.

**Assessing performance by cross-validation**

Ideally, for assessing the performance of the methods, we should obtain additional larger CAGE libraries, one for each tissue, and see how well the models trained on the original data can predict the new libraries. In the absence of this we can instead simulate this process by a cross-validation scheme where we for each tissue, sub-sample half the tags of a library without replacement. We proceed to fit both models on the sub-sampled library and thereby gain an estimate of their parameters. These parameters can then be used to extrapolate to the full sized library and predict the number of species we expect to observe there. For a
good model this should ideally be close to the true number of species observed. To gain a better estimate of the models performance we repeat the sub-sampling scheme ten times for each tissue.

**Evaluating the methods on TSS, promoter and gene level: multi-resolution comparison** We also need to consider the abstraction level of the species we are counting - in other words the resolution of the data. As discussed above, the finest resolution of the CAGE data are individual nucleotides being hit with the 5’ end of at least one CAGE tag. These are called CTSSs. Nearby CTSSs can be clustered into “tag clusters” (TCs), which could be interpreted as the result of a core promoter’s activity. The last abstraction level is to use the CAGE tags as expression data, by simply counting the expression of a given gene as the number of tags hitting that gene. In previous studies, it is clear that the level of resolution/abstraction is important for biological interpretation, as we sometimes are interested in actual TSSs and in other situations wish to study promoter or gene activity. The three levels are not necessarily highly correlated, as we can have many close individual CTSSs making up one larger tag cluster (a so-called broad promoter), or many tag clusters mapping to the same gene (alternative promoter usage). Therefore, we apply both methods on these three abstraction levels - CTSS, TCs and genes.

Figure 1 shows examples of cross-validation and validation predictions for all libraries at CTSS level, for both methods. It is clear that both methods do well: the relative deviation of the cross-validated predictions when we extrapolate from half the library size to the full library is in almost all cases 2-3 percent for the non-parametric and 0-1 percent for GIGP (shown as box-plots in Figure 2). For the non-parametric method, the infinite species assumption of the model results in a consistent upward bias in the predicted number of species. The cross validation extrapolation shows that the bias is present but small. Conversely, the GIGP method has a not obviously explainable bias towards underestimating the number of species.

If we instead of TSSs analyze TCs or genes, we see similar performance. For TCs, the relative error for both methods vary between 0-2 percent. Interestingly, both methods agree more on this level, but also tend to under-estimate the number of tag clusters. Part of this explanation is likely that the definition of a TC is arbitrary - TCs can have different widths and also vary between a single tags and thousands. If these results are not due to technical reasons within the library preparations, they indicate that the tag cluster level might be harder to model. In the gene case, we observe the same bias as in the CTSS (over-estimation by the non-parametric method and under-estimation by GIGP), but also higher errors - we speculate this is due to the potentially higher dynamic range in signals, since a gene is a relatively large
Cross validation on TSS level

Figure 1: The plots show the number of species (unique transcription start sites) captured with a given sequencing depth. Red lines show subsamplings of the actual library, while green and blue lines are predictions based on half the data for the non-parametric and GIGP methods, respectively. The right sub-panel for each tissue show a zoom-in of the predictions made when the sequencing depth is approaching that of the actual library size. Similar plots where the species are tag clusters or genes can be found in Figure S1-2. Complementary boxplots showing the deviations of the final predictions are shown in Figure 2.
Figure 2: The plots summarize the final predictions as shown in the zoom-in panels of Figure1 (and Figure S1-S2) as boxplots. The Y axis shows the deviation between actual counts of entities (TSSs, Tag clusters and genes, respectively) and the predicted value, where 0 corresponds to no deviation, positive deviations as over-predictions and negative as under-predictions.
interval which only need to be hit by a single tag to be counted as active. Regardless, these results indicate that the model predictions can serve as a useful tool for experimental planning: that is, extrapolating the number of unique species at different resolution levels (CTSS, TC and gene) for a larger sequencing depth.

**Estimating TSS coverage and promoter-ome complexity**

One of the advantages of the non-parametric method is that coverage has an easy interpretation - it is the probability of sampling one of the already observed species. This also means that we can predict both the number of tags needed to achieve a given coverage $n'$, and the number of TSSs (or core promoters or genes) $k$ hit at least once at that coverage.

We predicted $n'$ and $k$ for TSSs, TCs and genes for the coverage values 0.9, 0.95 and 0.99, respectively (Table 1). Firstly, it is interesting to note that on CTSS level, we do not reach 0.9 coverage for any of the tissues (Table 1) - in particular, the brain tissues except hippocampus have the lowest coverage. For these tissues, a large amount of tags are predicted to be needed to reach the desired coverage on this resolution level (between 46 and 74 million). On the other hand, for some tissues, less than 5 million mapped tags are needed to reach this saturation - which is realistic with current sequencing technologies. However, if we repeat the analysis using TCs instead of individual TSSs, we see that liver, hippocampus and embryo libraries already have a coverage $>0.9$ with the present sequencing depth. Similarly, on gene level, most tissues have a high coverage. Altogether, this strongly indicates that there is little added benefit of sequencing more than 5 million tags if the goal is to detect most active promoters or genes, unless the tissue is complex. The additional information given by sequencing more are on the TSS level, not the promoter level. This likely reflects that most promoters have a complex distribution of TSSs [22, 23]. These results also indicate that brain tissues in general use many more unique TSSs than other tissues.

To see if these results are consistent regardless of the number of tags sequenced, we plotted the predicted number of TSSs to as a function of the number of sequenced tags (Figure 3), using both methods. The steepness of the curves in these graphs will tell us something about the complexity of the underlying TSS/promoter selection regime in the various tissues. As expected, the tissues with the highest growth of TSSs per sequenced tag are the three brain tissues cerebellum, somatosensory and visual cortex, closely followed by hippocampus, while embryo has the lowest growth. This does not change dramatically if we assess tag cluster level - the only real difference is the lung tissue which has a higher growth rate compared to the other tissues on tag cluster level than on TSS level. This might be due to that brain tissues tends to
Table 1: Summary of predictions for all libraries. \( n \) is the number of mapped tags in the library, \( k \) is the number of unique species (TSSs, TCs or genes) “Non-param \( n+n' \)” is the estimated size of the library at coverage 0.99 for genes, 0.95 for TCs and 0.9 for TSSs. “Non-param \( k \)” is the estimated number of unique species at that coverage.

<table>
<thead>
<tr>
<th>GENE</th>
<th>( n )</th>
<th>( k )</th>
<th>Coverage</th>
<th>Non-param ( n+n' )</th>
<th>Non-param ( k )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cerebellum</td>
<td>203481</td>
<td>15079</td>
<td>0.98</td>
<td>431645</td>
<td>18232</td>
</tr>
<tr>
<td>Embryo</td>
<td>510662</td>
<td>16869</td>
<td>0.99</td>
<td>555920</td>
<td>17335</td>
</tr>
<tr>
<td>Hippocampus</td>
<td>983532</td>
<td>18665</td>
<td>0.99</td>
<td>439613</td>
<td>14804</td>
</tr>
<tr>
<td>Liver</td>
<td>1015553</td>
<td>18711</td>
<td>0.99</td>
<td>444202</td>
<td>14675</td>
</tr>
<tr>
<td>Lung</td>
<td>1015700</td>
<td>23121</td>
<td>0.99</td>
<td>514792</td>
<td>19445</td>
</tr>
<tr>
<td>Macrophages</td>
<td>141236</td>
<td>10643</td>
<td>0.97</td>
<td>490917</td>
<td>15946</td>
</tr>
<tr>
<td>Som. Cortex</td>
<td>170823</td>
<td>13813</td>
<td>0.98</td>
<td>454209</td>
<td>17997</td>
</tr>
<tr>
<td>Vis. Cortex</td>
<td>189515</td>
<td>14080</td>
<td>0.98</td>
<td>421419</td>
<td>17323</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TC</th>
<th>( n )</th>
<th>( k )</th>
<th>Coverage</th>
<th>Non-param ( n+n' )</th>
<th>Non-param ( k )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cerebellum</td>
<td>253234</td>
<td>60523</td>
<td>0.84</td>
<td>9931255</td>
<td>734295</td>
</tr>
<tr>
<td>Embryo</td>
<td>632197</td>
<td>66236</td>
<td>0.93</td>
<td>2239080</td>
<td>160288</td>
</tr>
<tr>
<td>Hippocampus</td>
<td>1388237</td>
<td>147973</td>
<td>0.93</td>
<td>4182405</td>
<td>310915</td>
</tr>
<tr>
<td>Liver</td>
<td>1218713</td>
<td>132383</td>
<td>0.92</td>
<td>5882224</td>
<td>409844</td>
</tr>
<tr>
<td>Lung</td>
<td>1298243</td>
<td>205394</td>
<td>0.88</td>
<td>41652301</td>
<td>2775431</td>
</tr>
<tr>
<td>Macrophages</td>
<td>169919</td>
<td>29956</td>
<td>0.89</td>
<td>1561807</td>
<td>123250</td>
</tr>
<tr>
<td>Som. Cortex</td>
<td>208943</td>
<td>52519</td>
<td>0.82</td>
<td>12912795</td>
<td>930603</td>
</tr>
<tr>
<td>Vis. Cortex</td>
<td>232690</td>
<td>55093</td>
<td>0.83</td>
<td>11565639</td>
<td>834630</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TSS</th>
<th>( n )</th>
<th>( k )</th>
<th>Coverage</th>
<th>Non-param ( n+n' )</th>
<th>Non-param ( k )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cerebellum</td>
<td>253234</td>
<td>133923</td>
<td>0.59</td>
<td>46475190</td>
<td>6368559</td>
</tr>
<tr>
<td>Embryo</td>
<td>632197</td>
<td>138190</td>
<td>0.84</td>
<td>4572923</td>
<td>610046</td>
</tr>
<tr>
<td>Hippocampus</td>
<td>1388237</td>
<td>290038</td>
<td>0.86</td>
<td>3623567</td>
<td>558917</td>
</tr>
<tr>
<td>Liver</td>
<td>1218713</td>
<td>282555</td>
<td>0.83</td>
<td>7999818</td>
<td>1105824</td>
</tr>
<tr>
<td>Lung</td>
<td>1298243</td>
<td>393792</td>
<td>0.78</td>
<td>24840031</td>
<td>3410933</td>
</tr>
<tr>
<td>Macrophages</td>
<td>169919</td>
<td>66257</td>
<td>0.72</td>
<td>5537268</td>
<td>788337</td>
</tr>
<tr>
<td>Som. Cortex</td>
<td>208943</td>
<td>115480</td>
<td>0.56</td>
<td>74388846</td>
<td>9933489</td>
</tr>
<tr>
<td>Vis. Cortex</td>
<td>232690</td>
<td>123847</td>
<td>0.59</td>
<td>52702455</td>
<td>7144462</td>
</tr>
</tbody>
</table>

use TCs that are composed by many individual CTSSs, which would give a higher growth on TSS than on tag cluster levels for these tissues.

**Discussion**

When planning a sequence-based experiment it beneficial to have a reasonable estimate of the number of possible outcomes to decide on the necessary sample size. Ideally, this should be large enough to properly asses the true underlying distribution of targeted entities (in our case TSSs), but small enough to avoid unnecessary expenditure and wasted man-hours. We have here implemented and tested two state-of-the art methods and found that they both can be used as frameworks for estimating how much the contemporary
Figure 3: The plot shows the predicted growth of species ($k$) as a function of sequenced and mapped tags ($n + n'$), on different biological levels and using different detection cutoff. This is essentially the same type of analysis as in the cross-validation test (Figure 1), but results for all tissues are shown in each image, and we extrapolate the growth curve to see the relative difference between tissues. See main text for interpretation methods must be scaled to capture a reasonably complete repertoire of the active species in a tissue, and can be viewed as part of a road map for planning future experiments.

Comparing the methods:

Several authors have, mainly because it is problematic to distinguish genuine low expression from noise, questioned whether it is at all possible to use SAGE or EST data to estimate the total number of tags (i.e. the size of the transcriptome) [5, 20]. Our findings indicate that the both methods used here fits the observed data well and give realistic (cross-validated) predictions. At TSS and gene level we see some trends: the GIGP method tends to under-estimate the actual number of TSSs while the non-parametric methods tends to overestimate them. However, the differences are small - amounting to a few percent deviations compared to the true result. The systematic biases are likely due to the finite versus infinite population size assumption of the models and suggest that even more flexible models are needed if higher accuracy is required.

Different tissues require different sequencing depth

An important biological finding is the large variation in TSS, promoter and gene numbers per sequenced tag between tissues, showing that the necessary number of sequenced tags is largely tissue (or sample)-dependent and that what is right for one tissue might be insufficient to another. Unsurprisingly,
the number of necessary tags follows to a large extent the number of cell sub-types encompassed by the various tissues. Brain cortex tissues, for instance, have a vast array of specialized cell sub-types and would be expected to show greater complexity, which our model clearly shows. More importantly, our model also reveals that cellular diversity is not only the factor that affects the total number of TSSs; for instance, cerebellum has a higher growth of TSSs per sequenced tag than the more complex hippocampus.

Differences in coverage at nucleotide, core promoter and gene level

Using the two methods on different levels of resolution (TSSs, tag clusters and genes), show that many tissues are predicted to be close to saturation level in terms of core promoters or genes with the current sequencing depths, in contrast to the coverage on individual TSS level. This indicates that most TCs have a complex distribution of tags that require high amounts of tags to ascertain.

While knowing the tag distribution on nucleotide level for most promoters is important if the goal is to dissect individual promoters and understand the mechanics of the transcriptional process, one can argue that most experimentalists are primarily interested in roughly what regions that are initiating transcription. If this is the case, then a relatively modest increase in sequencing depth is predicted to be sufficient for reaching an acceptable level of coverage.

Conclusion

With the rapid development of sequencing technologies there is little doubt that we can reach near-saturation in terms of identifying essentially all unique positions in a genome with a non-vanishing expression level. It is evident that the both the choice of tissue type, or really the population of cells being sequenced as well as the goal of the downstream analysis will decide whether the current sequencing technologies are adequate or whether one must wait for the next generation to approach complete coverage. Our work gives an estimate based on smaller samples that can help an investigator to decide and plan his next experiment and judge whether it is possible to achieve good enough coverage for the relevant cell type. We believe that an estimation of the coverage in a newly sequenced library should be part of the standard post-analysis of tag sequencing data. As such our work presents an important piece on the way towards establishing sequencing technologies that can achieve detection of most relevant promoters in any cell type. While we have scoped this work for TSS analysis, it can also be use for any sequencing-based method aimed to detect biological features, including ChIP-seq [24] and RNA-seq [25].
Methods

This section gives the technical details about the models used. In Results we have discussed how to make inferences about coverage and discovered species in a new sample for fixed model parameters. In the following we describe how to estimate the model parameters.

Availability. An implementation of the methods as well as the data sets used are available at [1]

Stick-breaking process We let $n, k$ and $\mathbf{n} = (n_1, \ldots, n_k), \sum_{i=1}^{k} n_i = n$ denote the quantities that defines the observed data set. The likelihood function for the model parameters $\sigma, \theta$ is the probability that the model assigns to the observed data sequence $\mathbf{c} = (c_1, \ldots, c_n)$. An important requirement for the construction of sampling scheme for independently sampled data is exchangeability which means the probability of observing a data sequence $\mathbf{c} = (c_1, \ldots, c_n)$ should not depend upon the order of the observations. It is easy to show that the Pitman-Yor sampling formula is the most general sampling scheme to fulfil this [12, 13].

We can illustrate this property with an example: Imagine we record the following sequence of species $\mathbf{c} = (E, E, M, T, T)$ (e.g. E=elephant, T=tapir and M=monkey). We can translate this to the sampling formula by indexing according to occurrence such $\mathbf{c} = (1, 1, 2, 3, 3)$, with count vectors $\mathbf{n}^{(1)} = (1), \mathbf{n}^{(2)} = (2), \mathbf{n}^{(3)} = (2, 1)$ and so on. Iterating the sampling formula and using the chain rule $p(\mathbf{c}) = p(c_5|\mathbf{n}^{(4)}) \cdots p(c_2|\mathbf{n}^{(1)}) p(c_1)$ we can write down probability of the sequence of observations $\mathbf{c}$ as

$$
\frac{\theta \theta 1 - \sigma \theta + \sigma \theta + 2 \sigma 1 - \sigma}{\theta 1 + \theta 2 + \theta 3 + \theta 4 + \theta}.
$$

On the other hand if we observe $M, E, T, T, E$, i.e. $\mathbf{c}' = (1, 2, 3, 3, 2)$ and $\mathbf{n}' = (1, 2, 2)$ we get

$$
\frac{\theta \theta \theta + \sigma \theta + 2 \sigma 1 - \sigma}{\theta 1 + \theta 2 + \theta 3 + \theta 4 + \theta}.
$$

Comparing the two probabilities we see that we get the required exchangeability: $P(\mathbf{c}|\sigma, \theta) = P(\mathbf{c}'|\sigma, \theta)$.

This is the likelihood function for $\sigma, \theta$ for that specific example. The general likelihood function follows by exactly the same argument, i.e. by using the sampling formula and the chain rule of probability:

$$
P(\mathbf{c}|\sigma, \theta) = \frac{1}{(1 + \theta)^{n-1}} \prod_{i=1}^{k-1} (\theta + i \sigma) \prod_{j=1}^{k} (1 - \sigma)_{n_j-1},
$$

where

$$(a)_n = a(a+1) \cdots (a+n-1) = \frac{\Gamma(a+1)}{\Gamma(a)} \frac{\Gamma(a+2)}{\Gamma(a+1)} \cdots \frac{\Gamma(a+n)}{\Gamma(a+n-1)} = \frac{\Gamma(a+n)}{\Gamma(a)}$$

is the ascending factorial and $\Gamma()$ is the Gamma-function. The first term in the likelihood comes from the denominator in the $n-1$ conditionals (the first observation contributes a one), the second term is the
contribution from the first sample in species 2 to \( k \) and the last term gives the contributions from observation 2 to \( n_j \) for species \( j = 1, \ldots, k \).

The simplest way to estimate the parameters is by maximum likelihood:

\[
(\sigma_{\text{ML}}, \theta_{\text{ML}}) = \arg\max_{(\sigma, \theta)} P(c|\sigma, \theta) .
\]  

(5)

Maximum likelihood in the context of this model was also investigated in [16]. Just using this set of parameters in the species sampling step omits the contribution of variability that comes from parameter uncertainty. Alternatively, we may take a fully Bayesian approach and specify a prior over the parameters \( P(\sigma, \theta) \) and make predictions averaging over the posterior of the model parameters. We have investigated this but has found that the large library sizes makes the posterior highly peaked at the maximum likelihood so that the variability from parameter uncertainty is much smaller than the bias we observe. We therefore for clarity stick to maximum likelihood estimation.

**Frequency of frequency model** The likelihood function is probability of the observed data given the parameters. Since the binomial distribution is also assigning probability to zero observations, that is the unobserved species, we need to zero-truncate and re-normalize the probabilities in the likelihood function:

\[
p_{\text{like}}(\cdot|\gamma, n) = p(\cdot|\gamma, n)/(1 - p(0|\gamma, n)):
\]

\[
p_{\text{like}}(n|\gamma, n) = \prod_{i=1}^{k} p_{\text{like}}(n_i|\gamma, n)
\]

We can also write this in a slightly more compact form first computing the frequency of frequency summary statistic, that is counting how many species occur \( i \in 1, \ldots, n \) times: \( m_i = \sum_{j=1}^{k} \delta_i(n_j) \), where \( \delta_i(l) \) is the Kronecker \( \delta \)-function which is one if \( i = l \) and zero otherwise. The number of samples and number of species can be calculated from frequency of frequency vector \( m = (m_1, \ldots, m_n) \) as: \( n = \sum_i m_i \) and \( k = \sum_i m_i \). With this expression the likelihood becomes \( p_{\text{like}}(n|\gamma, n) = \prod_{i=1}^{n} p_{\text{like}}^m(i|\gamma, n) \).

**Mapping of CAGE tags to genes** 61815 representative transcripts were downloaded from the [7]. These were then sorted by length (from smallest to largest). Any TC overlapping its start-end region (with a 50nt slack upstream) was assigned to that transcript if it has not already been assigned to another transcript. This means that small transcripts take precedence, a choice made due to the possibility of genes contained within introns.
Authors contributions

OW adapted the methods used. EV implemented the methods and made all related tests. AS initiated and coordinated the project, prepared data and interpreted outcomes. All authors wrote the article.

Acknowledgements

OW would like to thank Igor Prünster for many useful discussions. Authors affiliated with the Bioinformatics Centre are supported by a grant from the Novo Nordisk Foundation. The European Research Council has provided financial support to AS under the EU 7th Framework Programme (FP7/2007-2013)/ERC grant agreement 204135. PC is supported by the National Project on Protein Structural and Functional Analysis from MEXT and the National Project on Genome Network Analysis and the RIKEN Genome Exploration Research Project from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government. PC is also supported by a grant by the EU 6th Framework Program (NFG project).

References

1. Supplementary data [http://people.binf.ku.dk/albin/supplementary_data/tss_saturation/].


Abbreviations

CAGE  Cap Analysis of Gene Expression. An high-throughput experimental protocol for mapping TSSs.

CTD  Carboxyl Terminal Domain. The end of RNA PolII.

CTSS  CAGE-Tag Start Site. A collection of all CAGE tags mapping to the same location.

DNA  DeoxyriboNucleic Acid.

GIGP  Generalized inverse Gaussian model. One of two models used in paper IV.

MoAn  Motif Annealer. The name of our *de novo* motif discovery method.

mRNA  Messenger RNA. A RNA molecule that codes for a protein.

ncRNA  Non-coding RNA. RNA not coding for a protein.

PFM  Position Frequency Matrix. A matrix containing counts.

PIC  Pre-Initiation Complex. A complex formed by RNA PolII and and other factors before initiating transcription.

PEP  Preferentially Expressed Promoter. A weaker alternative to tissue specificity.

PSSM  Position Specific Scoring Matrix. A matrix consisting of log-odds scores.

RNA  RiboNucleic Acid. Similar to DNA, but has a different backbone and the uracil base instead of thymine.

TC  Tag Cluster. A collection of CTSSs that are clustered together. Roughly comparable to a core promoter.

TSS  Transcription Start Site. The genomic position of the first transcribed nucleotide.

tRNA  Transfer RNA. Used in translation of RNA into protein.

WM  Weight Matrix.